w241

Gene Ahn, Mike Gruzynski, Matt Stevenson

November 2017

1 Introduction

Measuring systematic racial discrimination has been a popular research area for the past several decades. While racial equality has been a cornerstone of America's modern social progress, the existence of innate racial bias of individuals and societies has been subject of many social experiments within the U.S.

Several past studies have sought to quantify the residual and systematic discrimination that remains in post-civil rights American society. Despite the rise of political constructs such as "color-blindness," many studies have inferred that racial biases may be more deeply rooted in our cognitive makeup. For example, Devah Pager and Bruce Western published several field experiments on the effects of race on employment. In their seminal work, *Identifying Discrimination at Work: The Use of Field Experiments*, Pager and Western sent "matched pairs of individuals (called testers) to apply for real job openings in order to see whether employers respond differently to applicants on the basis of selected characteristics." The testers were identically trained with similar physical attributes (i.e. height, weight, etc.) and work experience. Randomized intervention was on race, and the testers included white, black, and Hispanic applicants. Pager and Western reported a prevalence in racial discrimination with "Whites receiving positive responses at roughly twice the rate of equally qualified Black applicants."

More recently, social media provides a compelling forum for executing a field experiment to measure racial bias. More so, much of today's political discourse among citizens as well as from the current U.S. President occurs on Twitter. In contrast to interventions such as in-person interactions through confederates (as used by Pager and Western), Twitter and other social media platforms allow us to implement double-blind experiments across multiple topics of discussions at internet-scale. In addition, social media allows for rapid and transactional engagements with easily quantifiable results.

With many trending political debates occurring on Twitter, we want to understand the impact of race on social media response rate. Are Tweets coming from certain races more likely to be seen and/or read than others? If so, this opens up further questions regarding race-based social and political equality. Specifically, our research question is as follows:

What is the impact of the race of the commenter on the responses they receive on Twitter? The responses include impressions, engagement rate, retweets, replies, and likes.

2 Research Design

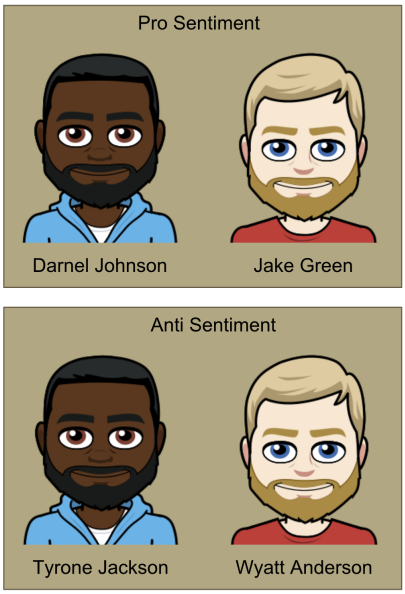
To answer our research question, we targeted a broad set of existing tweets with specific hashtags regarding trending current events (i.e. sexual harrassment or NFL players kneeling for the national anthem) and replied to them with pre-scripted tweets.

Figure 1: High-Level Experimental Design

R	$X_{1,1}$	O
R	$X_{1,2}$	O
R	$X_{2,1}$	O
R	$X_{2,2}$	O

We used a 2x2 factorial design to conduct our experiment and capture heterogeneous treatment effects. The two factors we manipulated were (1) race of our Twitter avatar (i.e. white vs. black avatars and white vs. black-sounding names) - see figure below - and (2) point of view of our reply (i.e. pro vs. anti the majority argument at hand) for a total of four different experimental conditions as shown above. We chose the names of each avatar by randomly choosing first names and last names from top 20 lists from ABC News for the most common white and black names in the United States (http://abcnews.go.com/2020/top-20-whitest-blackest-names/story?id=2470131). Our goal was to measure racial bias across both sides of a Twitter debate. Additionally, our experiment was a double-blind audit study. The Twitter profiles and homepage designs were identical for all four experimental conditions with the exception of race and point of view, and with all of our Tweets pre-scripted and assignment automated, those administrating the experiment had no say in whom we replied to or how we replied.

Figure 2: Twitter Avatars

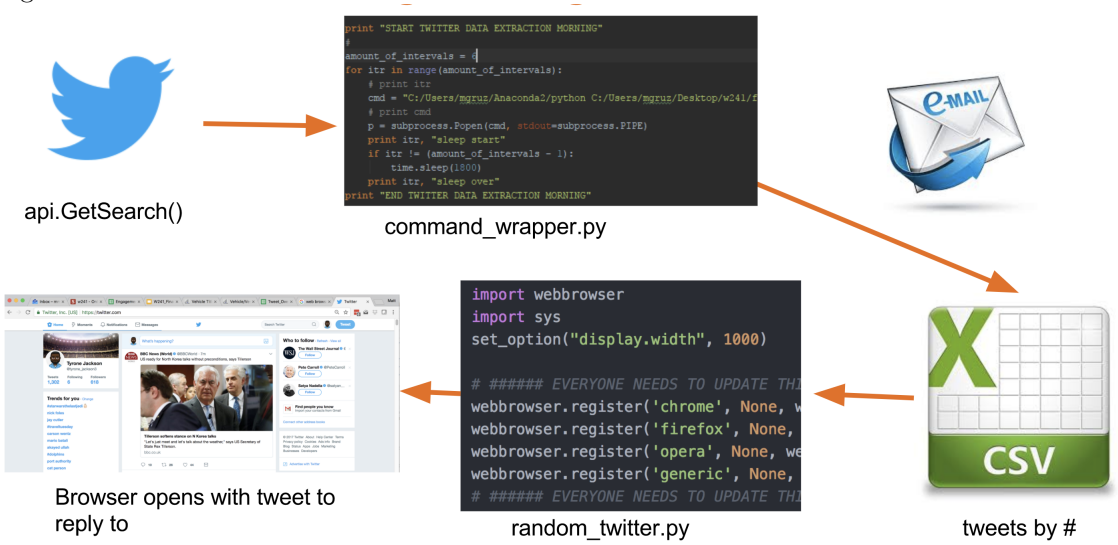


After randomly assigning to the four experimental conditions, we applied the treatment and measured the intent to treat effect (ITT), captured by an outcome variable constituted of the total number of likes, retweets, and replies that we received.

3 Randomization engineering

The randomization performed on the final experiment was performed using Python. We used a customized Python Github package to access the Twitter search API called "Twitter". We would pull up to 100 tweets per pull request using the Twitter package in Python. A pull was performed every hour for 6 hours and then compiled, filtered to make sure there were unique users ids and tweet ids. A maximum of 100 tweets was the limit for the Python package we used, and could not be inflated, so instead of worrying about pulling more tweets per pull, we ran more pull requests. A filter for unique ids was used in order to reduce spillage among the Twitter subjects. The dataframe then was run through a random sample generator that assigned each Twitter avatar to a group of subjects. The potential subjects would take the total amount of tweets per session and then divide that number by 2 (50 %) and then split up the remaining data evenly to the 4 experimental conditions/avatars. There were three sessions per day where we aggregated unique tweets from the morning, the afternoon and night. An automated email was sent to the teammates in order to start sending out tweets that were designated to each twitter avatar. Twitter API doesn't allow the same twitter user to send the same words in a tweet within a 24 hr window, so we sent the tweets by hand, automatically populating each avatar into a different web browser (i.e Chrome, Firefox, Opera, etc) in order to reduce the likelihood of tweeting the wrong tweet per Avatar.

Figure 3: Treatment Flowchart



The experiment selected hashtags as per what was trending, creating a series of tweets that were

provocative but not inflammatory in order to elicit a response. Each avatar always maintained its stance (either pro or anti the subject of the hashtag). The potential number of tweets sent per day were the same, however other factors such as spill over (avatars having already sent to the same user or tweet), confederates not tweeting in a timely fashion, broken links, etc. meant that the actual tweets sent per avatar is not identical but similar in volume. In addition, spillage was often observed. A tweet would have a hashtag for possibly two or more of the hashtags we were studying. For example, a tweet might be attacking sexual misconduct with a statement and then hashtagging both KevinSpacey and harveyWeinstein (etc.). There were other spillage problems with time lapse not long enough and same tweets from previous sessions making it on two or more sessions, due to over all low volume of tweets on a subject. We simply did not tweet if another avatar already posted to the tweet. Some examples of experimental conditions and avatars are as follows for the hashtag KevinSpacey:

Darnel - Black, pro
Tweet- These allegations have blown Spacey's "house of cards" down
Jake - White, pro
Tweet - These allegations have blown Spacey's "house of cards" down
Tyrone - Black, anti
Tweet - Weren't we the ones entertained by Spacey's portrayal of sexual deviant in American Beauty and House of Cards? Take a look at yourself!
Wyatt - White, anti
Tweet - Weren't we the ones entertained by Spacey's portrayal of sexual deviant in American Beauty and House of Cards? Take a look at yourself!

Figure 4: Sample Tweets

4 Measurement of variables

The main draw of Twitter in running this experiment was the availability of reported metrics via Twitter Analytics. Unfortunately, the functionality of a full corpus of tweets and their response metrics via CSV download was not available to us during the trial period. At best, we could pull 75 tweets per account over the past 28 days, a far cry from the 1140 per account that were actually sent. This seems to be a longstanding issue with Twitter, despite their advertising, user beware.

However, due to the low response rate achieved in our study, we were able to use the summary statistics available on the Twitter Analytics site to categorize the engagements for the 30 tweets that received responses during the main study focused on sexual allegations. Likewise, we were able to use these statistics to categorize our distributions and some other method checks, though limited.

For example, without the CSV printout available, compliance was not measurable by tweet (the unit of observation). We were able to see daily compliance rates per avatar (which allowed us to build the randomization inference distributions in the following sections), but while we originally intended to compute the complier average causal effect (CACE), we had to take on faith that a tweet sent was a tweet viewed (in Twitter lingo, potential views are noted as "impressions"), limiting us to an intention to treat analysis for our average treatment effect.

If we had been able to access the total impressions per tweet, we could have divided our outcome of number of "engagements" (likes + retweets + replies) by total impressions to normalize the effects witnessed from each of our tweets. This would scale our dependent variable accordingly, adding precision to our treatment effect estimate. Even with this data available, there is a caveat, as potential impressions do not necessarily indicate compliance in our trial, as there is no guarantee that the people to whom our tweets were visible actually read them, lending credence to an ITT approach.

Two scenarios that limit compliance include the small double digit percentage of Twitter users who choose to make the replies to their tweets invisible to other users. This would intentionally limit the viewership of our replies to the @user who we originally replied to. Given that they hide replies to their tweets, these users are unlikely to engage in back and forth conversations, which would constitute noncompliance. Likewise, Twitter has instituted hiding abusive tweets. We intended our replies not to be abusive in nature, however given that we sent the same Tweets multiple times Twitter might have hidden our replies without us knowing. This would encourage noncompliance and bias our average treatment effect towards 0.

Attrition in our study was only possible if we had deleted our own tweets, which did not occur. While certainly some of the tweets that we replied to were removed over time, our replies stayed live and we could measure the outcome, though in this situation, it would also bias our result towards 0, though this effect should be balanced throughout our 4 treatment groups due to randomization.

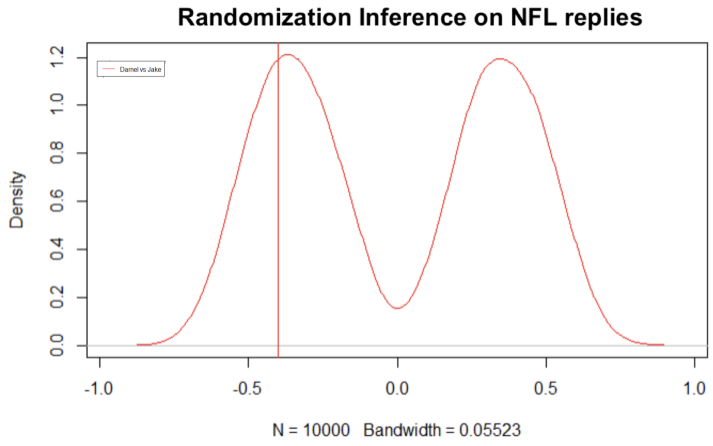
5 Pretest

A pretest was performed for our experiment in order to figure out what platform we were going to use and to see if we were able to get the sample size needed for our statistical analysis. The platform choices were Facebook news pages and Twitter hashtags. We discovered very quickly that just posting to the article in Facebook post yielded no results. We moved to randomly selecting users that had already posted to the news article and compared it with randomly selecting user's

posts that have tagged a specific hashtag we were investigating. For the first week of the pretest, we found out that we had very small amount of interaction on Facebook. We had very few comments and likes vs Twitter which had more interactions with the users. We decided to go with Twitter moving forward with the experiment because of interaction rate and more tools in order to gather data on each message (Twitter Analytics) as well as the ability of automating tweet processing.

The initial category we were attacking was the "NFL take a knee" subject that flooded the media for the first half of the 2017 NFL season. Our response rate was fairly good, but after a week of an additional pretest on the Twitter hashtag subject, we realized that we were potentially biasing the experiment with the choice of an intrinsically racial subject. Our main research question was racial bias on Twitter and we were using a serious racial issue that exists in America (systematic racism against black Americans). We were worried about biasing the experiment because of the degree of racial problem inside the hashtag we were studying. For example, will the experimental subject be more likely to react to a black avatar on a black/white issue vs a white? In our opinion we theorized that many people will be hesitant to react because the experimental subjects wouldn't want to appear racist online. We decided to move to a new set of hashtags and the shape of the random inference comparison of white vs black of engagement rate for this subject, figure shown below.

Figure 5: Randomization Inference for NFL Pre-Test



As you can see in the above image, we see a bi-modal random inference mode shape, showing that there was high variance between the number of people who engaged with the African-American avatar vs the white avatar. We would ideally want a smooth bell shaped curve of potential values in order to make apple-to-apple comparisons of our treatment vs control subjects. Given the racial focus of the NFL national anthem issue, there may be other predictive covariates that we can't account for. For example, we would need to control for the racial focus of the given debate. Given this potential bias, we decided to move to another subject on Twitter.

We then switched to another pretest of switching the subject from NFL white/black issue to tax reform. After one week and sending almost 1000 tweets, we only received 1 engagement (reply, retweet, like, etc). We were attacking tax reform and having no engagement, which our theory was that Twitter subjects posted about it, but didn't care or know enough about tax reform to react

to our tweets. After that failed attempt, we finally switched to responding to comments on twitter hashtags surrounding sexual misconduct allegations in Hollywood and politics.

We were tweeting on the subject of allegations surrounding the sexual misconduct and not the actual act of sexual misconduct. We thought that allegations are sometimes built on weak claims so we had potential for strong reactions. We were running a 2x2 experiment or white vs black and pro vs anti allegation, so we had the potential to have engagement from both sides of the sexual misconduct allegations. The hashtags we were tweeting against were: #RoyMoore, #AlFrankenResign, #GeorgeTakai, #HarveyWeinstein, #KevinSpacey, #LouisCK, #Conyers, and #CharlieRose. One thing to note is that we had to pay attention to the actual random tweets we selected to post to because a lot of tweets had multiple hashtags on the same tweet, so spillage was a main concern as well with subjects attached to each tweet. This resulted in us not sending a few tweets here and there, but the balance of tweets sent per avatar remained quite even among the four groups.

Figure 6: 2x2 Table

Avatar Sentiment	Anti - Accused		Pro - Accused	
Avatar Race	Darnel (Black)	Jake (White)	Tyrone (Black)	Wyatt (White)
Engagement Rate	1.7%	0.79%	0.95%	0.62%
Tweets Sent (N)	1137	1140	1155	1121

A larger loss of tweets sent was due to not completing them in a timely fashion, due to the sheer volume of tweets we collected everyday. Without Twitter analytics functioning, we were only able to get summary statistics and replies to our tweets sent, and were unable to get exact counts of how many tweets were posted by avatar per hashtag. However, we ended up sending a total of 4553 out of the total corpus of 8320 collected. The randomization retained balance (25 +/- 0.5 %) because using our script we sent fully randomized CSVs for a given hashtag at a time. The python package webbrowser that we used to open the tweets sometimes dropped the browser for Wyatt, resulting in slightly fewer tweets sent from him as seen above.

At this rate of engagement, to get 30 measurements per outcome group in the simple 2X2 model which was our original goal, we would have needed to send 4X (n=18,000) tweets, which would have meant automating or turking the process.

6 Modeling choices

In the Twitterverse in which people comment on sexual allegations of well known male public figures, most of the comments come from white men and women who are expressing sentiment against the person accused of sexual misconduct (Anti-the-accused). With this in mind, we established that our

tweets originating from an African-American avatar (B_i) that are pro-the-accused (P_i) constitute both our treatment groups. Dummy variables were set up so that our avatars reflected this, with Tyrone Jackson ($B_i = 1$, $P_i = 1$), etc.

1.) 2X2 Model with experimental variables Black and Pro-the-accused status:

$$Y_i = \alpha + \beta_1 B_i + \beta_2 P_i + \beta_{12}(B_i P_i) + u_i$$

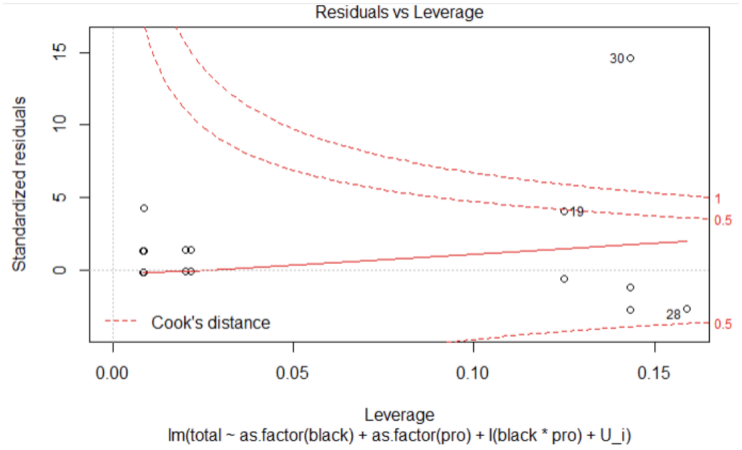
Figure 7: 2x2 Regression Summary

```
Call:
lm(formula = total ~ as.factor(black) + as.factor(pro) + I(black *
    pro), data = df_accusation)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7143 -0.0039 -0.0031 -0.0031  9.2857

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.0039387  0.0035417   1.112    0.266
as.factor(black)1  2.7103470  0.0640876  42.291   <2e-16 ***
as.factor(pro)1   -0.0008726  0.0050099  -0.174    0.862
I(black * pro)    -1.3384131  0.0877646 -15.250   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1693 on 4579 degrees of freedom
Multiple R-squared:  0.3356,    Adjusted R-squared:  0.3352
F-statistic: 770.9 on 3 and 4579 DF,  p-value: < 2.2e-16
```

30 tweets received some type of interaction from other users, resulting in a < 1 % engagement rate for the overall experiment (n=4553). This data is clustered by individual tweet sent. As can be seen by the regression coefficients, there was a noticeable uptick in number of responses when the avatar was African-American. However, when combined with being on the side of the accused sex offender, this response rate died down a bit, with the interaction coefficient being about half as large but in the opposite direction of the coefficient on being black. While no effect was witnessed on the sentiment of the avatar's response (the coefficient on "pro" treatment was non-significant), the interaction of pro and black seemed to have an effect. Given the small amount of data to support this regression, we do not put much weight on these effects, except to say that they are worthy of a fuller, larger study which engenders a higher response rate. Some of the methods to achieve this are detailed in the two remaining model descriptions below.

Due to its small size, this data set happened to be badly right skewed, due to one of the tweets (observation 30) having 10 likes and 2 retweets. As this is a normal amount of activity for a tweet, in a future experiment with more tweets sent and engagement rate increased, this skew would come back towards center.

Figure 8: Residuals vs. Leverage Plot



2.) 2X2 Model with Covariate for the User Intention of the original tweet:

$$Y_i = \alpha + \beta_1 B_i + \beta_2 P_i + \beta_{12}(B_i P_i) + \beta_3 U_i + u_i$$

For our exogenous covariate of interest, the sentiment of the original tweet that we collected with our scraper and replied to under randomization was marked with a dummy variable, with $U_i = 1$ for tweets that are pro-the-accused. We labeled the user intention of a random subset of the tweets that we replied to and pushed our 30 tweets with responses into the same dataframe to mimic the 10 % of responses we found in the pretest, finding the following results.

Figure 9: 2x2 w/ User Intention Regression Summary

```
Call:
lm(formula = total ~ as.factor(black) + as.factor(pro) + I(black *
    pro) + U_i, data = df_user_intentions)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7213 -0.0788 -0.0631 -0.0299  9.2787

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.07885    0.06419   1.228 0.220356
as.factor(black)1  2.64243    0.26697   9.898  < 2e-16 ***
as.factor(pro)1   -0.01570    0.08465  -0.186 0.852972
I(black * pro)    -1.33057    0.36618  -3.634 0.000333 ***
U_i               -0.04892    0.10140  -0.482 0.629888
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6881 on 275 degrees of freedom
  (50 observations deleted due to missingness)
Multiple R-squared:  0.3163,    Adjusted R-squared:  0.3063
F-statistic:  31.8 on 4 and 275 DF,  p-value: < 2.2e-16
```

Again, being African-American had an positive impact on the number of engagements yet being black and pro the debate reduced this effect due to a heterogeneous treatment effect, whereas neither the sentiment of the original tweet nor the sentiment of our response had a noticeable effect on

number of replies. The low response rate and sub-sampled nature of this data make drawing any conclusions fruitless here.

3.) Saturated 2X2X2 Model with User intentions and all interaction terms:

$$Y_i = \alpha + \beta_1 B_i + \beta_2 P_i + \beta_3 U_i + \beta_{12}(B_i P_i) + \beta_{13}(B_i U_i) + \beta_{23}(P_i U_i) + \beta_{123}(B_i P_i U_i) + u_i$$

Figure 10: Saturated Model Table

Audience Sentiment	Anti Accused (Siding w/ Allegations)				Pro Accused (Against Allegations)			
Avatar Sentiment	Anti - Accused		Pro - Accused		Anti - Accused		Pro - Accused	
Avatar Race	Darnel (Black)	Jake (White)	Tyrone (Black)	Wyatt (White)	Darnel (Black)	Jake (White)	Tyrone (Black)	Wyatt (White)
Engagement Rate								
(N)								

Above is our ideal model, including interaction terms to measure all heterogeneous treatment effects among the three variables noted here.

In terms of User sentiment 75 % are Anti the Accused , 25 % Pro the Accused in this debate judging from the sample of 150 tweets that we marked manually, so the boxes on the left are over weighted. This is due to endogeneity unique to this debate. In order to block on Pro-accused and over sample (3:1) to balance the groups, it would have required reading the tweets beforehand and marking their user intentions. This is a complicated problem to solve (it takes about 10-20 seconds for a human to judge the intention of a tweet), with 10 % of all tweets coming out as non-biased/ambiguous and the potential to influence the double blindness due to the increased knowledge of the confederates sending the tweets. Conversely, to train an NLP algorithm to do this task, it would require at least as large of a labeled training data set as our current total corpus of tweets and potentially still not be that accurate, so was not very feasible in our case.

With enough data to run these models, Models 2 and 3 can be compared using an F-test in order to see if there are differential treatment effects between how users grouped on pro/anti intentions to their tweets react to the race and sentiment of the specific avatar's tweets. Given that we did not randomize on user sentiment and did not find any effect of avatar sentiment in the data we collected, we would first attempt to use Model 2, treating sentiment as a covariate in our model to help reduce variance and make a more accurate ATE.

7 Discussion

Given the low response rate in our study, we are unable to quantify the amount of racial bias that exists on a trending national debate on Twitter. With the few observations that we had, our model

was susceptible to outlying and leveraged observations. In future studies, we would not only require larger sample sizes, but may also want to add covariates (i.e. number of followers or engagement of the targeted tweet) to increase precision in our coefficients.

However, even within this small number of observations collected, there is some evidence that white and black Twitter users receive differing amounts of engagements or attention on Twitter. If we could reproduce these results with a larger scale study, this would indicate evidence of racial bias even in short-term internet exchanges through Twitter handles where there is no physical interaction. At a minimum, this warrants a fuller study as described in the Models section.

8 Sources

Gerber, A. S., Green, D. P. (2012). Field experiments: design, analysis, and interpretation. New York: W.W. Norton.

Pager, D., Western, B. (2012). Identifying Discrimination at Work: The Use of Field Experiments. Journal of Social Issues, 68(2), 221-237. doi:10.1111/j.1540-4560.2012.01746.x