Signalling Virtue: Pre-Analysis Plan

Matt Lowe* Deivis Angeli†

May 24, 2022

1 Background

Social media platforms are not only a technology for social networking, but also an important space to signal support for social causes. Are these signals informative about the private behaviors of the poster? Or do they instead reflect some aspirational ideal? We bring data to this debate, clarifying both whether (a) public signals correlate with private virtue, and (b) whether consumers of social media *think* that public signals signal virtue. Parts (a) and (b) together speak to a game-theoretical debate: whether in equilibrium signallers disclose useful information about private behaviors or not. While part (a) uncovers the behaviors of signallers, part (b) speaks to whether people have sophisticated beliefs about what signalling equilibrium we are in, or whether they instead misperceive the equilibrium. We test these ideas in the context of US academics' support for racial justice on Twitter.

In this pre-analysis plan, we describe our experimental approach to part (a). In particular, we will run an audit experiment with US-based academics. We will link the data from the audit experiment with the tweets of each academic. We will use the linked data to test whether racial discrimination differs between those that have supported racial justice on Twitter versus those

*Assistant Professor of Economics, University of British Columbia.
†PhD Candidate in Economics, University of British Columbia.

that have not done so.

Following the experiment, we plan to survey US-based graduate students to have them predict the results of the experiment – a test of whether students have accurate beliefs about the signalling equilibrium. In the survey we also hope to elicit an additional measure of academics' private behaviors, based on the third-party reports of the students themselves. We do not prespecify the student survey part here, since the questions we plan to ask students depend somewhat on the results from the audit experiment itself.[1]

2 Data and Sample

2.1 Tweeting Academics

We began assembling the data in early-2021. With the help of a rotating team of over 100 undergraduate volunteers (the team, from now), we first listed all research academics in the top-150 universities according to the 2019 US News Universities Rankings.[2] The team found lists of all departments granting Ph.D. degrees. Next, they inspected the faculty pages of each department, recording the name of each academic in a research position. The team found over 125,000 research academics in 5,113 departments from the top-150 universities in this step. We also collected data for eight large universities ranked outside the top-150 to participate in a pilot audit experiment (see Section 4).

In the second step, we found the subset of academics with Twitter accounts. We used a search engine with automated searches to create a shortlist of possible Twitter handles for each academic. For the academics with at least one possible matching Twitter account, the team manually found the correct Twitter account. For these academics, the team also recorded the academic's email address and position (Assistant, Associate, or Full Professor), and guessed

[1] For example, if we find that discrimination in the audit is predicted by the content of original tweets but not by retweets, we may want to elicit student predictions separately for these two types of tweets.

[2] The 2022 rankings can be found here.

their gender (Male, Female, Other) and race or ethnicity (Caucasian, Black, East Asian, South Asian, Hispanic, Other, Uncertain). We drop any academics without an email address available online.

In early-2022 the team double-checked the entire list of tweeting academics. We used this check to drop any non-research-active or non-professor academic – e.g. those on leave, post-docs, emeritus professors, and adjunct professors. We also dropped professors with websites mentioning a policy of not responding to emails from prospective students. This leaves us with a sample of 28,302 tweeting research academics.

For the experimental sample, we impose six final eligibility criteria. First, we drop Black academics and those with ethnicity coded as "Uncertain," given that the core research question is about how *non-Black* academics signal support for Black people in America. Second, we keep only the academics that joined Twitter on May 1, 2020 (just prior to George Floyd's murder) or before, ensuring that our experimental sample were on Twitter during the height of tweeting about racial justice. Third, we require a minimal level of public Twitter activity, keeping only the academics with at least five public tweets in 2020. Fourth, we drop academics with lab-oriented Twitter accounts with no personal tweet content. Fifth, we drop a few academics in departments for which our email templates do not fit well (e.g. Theater, Education departments oriented towards practitioners). Sixth, we drop a handful of academics with whom we had discussed the project with. This leaves us with a final experimental sample of 18,514 academics.

2.2 Twitter Data

We used Twitter's official API to download user-level and tweet-level data for each of the academics with a Twitter account.

Among other variables, the user-level data includes the user's screen name, number of followers, number of accounts following, number of tweets, and the date that the account was created. We have weekly snapshots of this user-level data from March 18 to May 6, 2022 (as of

writing).

The tweet-level data includes all tweets, replies, quote tweets, and regular retweets from January 1, 2020 to March 27, 2022. We use January 1, 2020 as the start date to cover the tweets before and around the time of the murder of George Floyd. We use March 27, 2022 as the last date as we began to pull the tweets shortly after. Among other variables, for each tweet we have the full text of the tweet, the date it was created, and the number of likes, replies, quote and regular retweets.

For proxies of political preferences, we are in the process of assembling data from three sources. The first two use Twitter data: first, Blindspotter uses the news diet of Twitter users to classify each on a left-right scale. Second, we will use the following lists of each user to measure the fraction of political accounts followed that are Republican vs. Democrat-affiliated. Third, we will link each academic with public data from the FEC on political donations.

2.3 Measuring Signalling

A key part of the project is determining which academics tweeted in support of racial justice on Twitter and which did not. Given the large sample size, we automate this classification based on the words and phrases includes in each academic's tweets. In particular, we define Signaller_i as a binary variable equal to one if academic i has at least one tweet (of any type) from January 1, 2020 to March 27, 2022 that mentions at least one of these racial justice-related words or phrases:

1. *Racism-related*: racism / racist / racial bias / racial discrimination / racial justice / racial prejudice / anti black / white supremacy

2. *Black Lives Matter movement-related*: BLM / black lives / blackintheivory

3. *References to Black individuals killed*: george floyd / ahmaud arbery / breonna taylor / daunte wright / justiceforgeorgefloyd / justiceforgf / justiceforahmaudarbery / justicefor-

breonnataylor / justicefordauntewright / sayhername / sayhisname / nojusticenopeace / icantbreathe

We chose these words and phrases to cover the most popular racial justice-related hashtags and to explicitly reflect racial justice themes. For example, we do not include phrases like 'affirmative action' given that a tweet that references this may not necessarily be referring to affirmative action around race specifically. Otherwise, we allow for slight variants of the above terms (e.g. "breonnataylor"). For "BLM" we require the term to not be part of a longer word. This way we avoid classifying medieval manuscript lovers (who may refer to the account @BLMedieval) as racial justice signallers.[3]

The automated signalling measure raises two main concerns. First, an academic auto-classified as a signaller may have tweeted about racial justice, but not necessarily *in support* of racial justice. For example, the auto-classification would consider an academic a signaller if they have ever tweeted "the racial justice movement has gone too far." This case would be a false positive. Second, there may be academics that have tweeted in support of racial justice without using one of the words or phrases above. These cases would be false negatives.

To test for these concerns, we use a richer manual measure of signalling status for a random subset of our experimental sample ($N = 443$). In this random sample, we automatically classify 58% (259 of 443) as signallers. In the full experimental sample, we automatically classify 59% as signallers.

For each academic in the random sample, one or two team members each spent up to five minutes scrolling through the academic's tweets, beginning around May 25, 2020 (the date of George Floyd's murder). After doing so, they selected as many options that apply from the following:

- I did not find any tweets or retweets related to racial justice for Black people

[3]Currently our classification ignores words and phrases in hyperlinks in tweets, as the hyperlinks are shortened by Twitter to the format beginning t.co/, losing any meaningful content in the link address. We are working on recovering the full link address, and will then update the algorithm to look for the same words and phrases as above in the link itself.

- I found at least one tweet or retweet **opposing** efforts to promote racial justice for Black people
- I found at least one tweet or retweet **questioning** efforts to promote racial justice for Black people
- I found at least one **neutral** tweet or retweet about racial justice for Black people (e.g., mentioning a neutral statistic or a study)
- I found at least one tweet or retweet **showing some support** of efforts to promote racial justice for Black people (e.g. one short tweet with #BLM)
- I found at least one tweet or retweet **heavily supporting** efforts to promote racial justice for Black people (e.g., a long and thoughtful tweet describing why we should support racial justice for Black people, or problems with how police treat Black people)

The data from this user-level manual classification exercise increases confidence in our cruder automated measure, Signaller$_i$. In particular, no academic in this random subsample has ever tweeted in opposition of racial justice, and only four academics questioned efforts to promote racial justice (Figure 1). It follows that there are practically no false positives – in the sense of academics auto-classified as supporting racial justice that in fact question or oppose the movement.

Second, while roughly 17% of those not auto-classified as signallers have shown any support on Twitter for racial justice, the figure is 77% for the signallers. The difference is even larger for heavy support, with our signallers roughly 12 times more likely to have tweeted in heavy support of racial justice efforts. This validation shows that false negatives are not too common, and that our automated measure signalling has a large first stage on the richer manual measures of signalling.

Signaller$_i$ is our primary measure of whether an academic has signalled support for racial justice on social media. Nevertheless, in our empirical work we will also explore what *types* of

racial justice posts are more versus less informative of racial discrimination in the audit study. This empirical exploration necessitates the breaking up of Signaller_i into different components, as we discuss in more detail below.

3 Audit Experiment

3.1 Experiment Design

Name Selection. We chose 120 racially distinctive names largely following the approach of Kessler et al. (2019). For first names, we use data on baby names from New York City and Massachusetts.[4] We keep only those for birth years 1995 to 2004, making the individuals around late-college age today. We drop distinctively Jewish- and Italian-sounding names, any first names used in Bertrand and Mullainathan (2004) (since these names may be distinctively fictitious-sounding to some academics), and eight first names used in our pilot experiment.[5] We impose a popularity threshold, keeping only the first names used by at least 0.01% of a gender-race cell (e.g. white men). We then keep the top-36 most distinctive for each gender-race cell. For example, for white men, we keep the 36 first names with the highest probability of being a white man conditional on the first name being used. This leaves us with 144 potential first names.

For last names, we use the 2010 US Census, and a within-race popularity cutoff of 0.1% (exactly as in Kessler et al. 2019). We drop eight last names used in our pilot.[6] We keep the 72 most racially distinctive last names – 36 for Black last names, 36 for white.

In the final step, we randomly match each first name to a last name, with each last name used twice – once for a male-sounding and once for a female-sounding name. This leaves us with 144 full names. To select the 120 most racially distinctive names from among these, we paid

[4] For the data for New York City, see here. We received the Massachusetts data from the Massachusetts Registry of Vital Records and Statistics.

[5] Iyanna, Tyra, Latrell, Tyreek, Jaclyn, Molly, Graham, and Jonah.

[6] Washington, Glover, Ware, Clay, Collins, Peterson, Ward, and Phillips

MTurkers to guess the race of the names. We drop the six names with the least accurate guesses in each gender-race cell, leaving us with 120 full names to use for the full audit experiment (see Figures 2 and 3).

Email Addresses. We created one gmail account for each of the 120 full names. Stratifying by race-gender cell, we randomly assigned each name to one of four possible email formats: [firstname].[lastname][X]@gmail.com, [firstname][lastname][X]@gmail.com, [lastname].[firstname][X]@gmail.com, or [lastname][firstname][X]@gmail.com, where X is a number.

To choose the number X we use a protocol that ensures that the number of digits in X is balanced between Black- and white-sounding names. In particular, we first randomly pair each full name with a full name from the same gender but different race. We then find the lowest X such that a gmail account with that X is available for both the Black and white full name in a given pair. We then randomly pick two numbers above that number, with the same number of digits, and assign one to the Black name and one to the white name.

Main Randomization. For the audit experiment, we will send one email to each academic, purporting to be an undergraduate student interested in graduate studies at the academic's university. The core randomization is to:

1. Black- vs. white-sounding name (50:50), stratifying on university-by-department
2. Male- vs. female-sounding name (50:50), stratifying on university-by-department-by race of sender
3. Sentence mentioning first-generation college student or not (50:50), stratifying on university-by-department

We use the first-generation college student randomization to test for whether support for racial justice on Twitter is informative of support for underrepresented students in general, be-

yond support for racial minorities. We use the gender randomization primarily for a companion project on the signalling of gender bias, inside and outside of academia.

After these steps, we know the purported race, gender, and first-generation student status of the sender for each academic recipient. Next we randomly assign the student's name. For this we randomly choose one Black-male name, one Black-female name, one white-male name, and one white-female name to be used for each university-department. This ensures that all tweeting academics in the same department at the same university assigned to receive an email from, for example, a Black male, will receive an email from the exact same Black male.

In the final step, we randomize the subject and main text of each email at the level of the sender-by-university-by-department, subject to the constraint that the same email type is not used by more than one sender for the same university-by-department. This constraint minimizes the possibility of academics detecting the deception by comparing emails and seeing two identical-looking emails from different senders.

We choose the main text of the email from 12 possible variants. We then randomly choose a minor variant of the email from three options for each of the 12 main text variants. The minor variants involve small changes to minimize the chances of our emails being detected as spam (e.g. "final year of undergrad" instead of "final-year undergraduate"). We randomize the minor variant at the level of sender-by-university-by-department, meaning that a given fictitious student uses the same minor email variant for all of their emails to a given university-department.[7]

Minor Randomization Details. For randomization stratified on university-by-department, we make sure that all strata have at least four observations (covering the four race-by-gender treatments) by joining together small strata (usually creating a strata that includes all of the small departments of a given university).

We split the universities into nine groups according to the final exam dates for the last

[7]For some departments in which the academic would not work on research per se (e.g. because they compose music), we use a fourth minor variant which replaces the term "research" with "work" throughout.

semester. We will email the universities according to this order – i.e. we will first email the set of universities with the earliest exam date. Within these nine groups, we randomize the order in which we email each academic.

Since each email mentions the undergraduate institution of the fictitious sender, we assign this institution randomly at the level of the sender-by-university-by-department. For the set of possible institutions, we start with the same top-150 US News ranked institutions as for our sample of academics. We use NCES data from Fall 2020 to keep the 90 institutions that satisfy these eligibility criteria: (i) least 4% Black or African-American undergraduate enrollment, (ii) at least 20% White undergraduate enrollment, (iii) 20 to 80% Female undergraduate enrollment, (iv) undergraduate degrees offered, (v) at least 4,000 undergraduates enrolled, and (vi) no technology focus (i.e. we drop institutions like MIT). For each university we email, we keep the eight of the 90 institutions that are closest in rank to be considered as the institution of the fictitious student.

Ethics. We received full ethics approval for the audit experiment from UBC's Behavioural Research Ethics Board (ID: H20-03758). The audit experiment necessitates deception to give a plausible measure of actual racial discrimination over email. We take several steps to minimize ethical concerns. First, to reduce the burden on academics, we will send the emails during May when most research academics are not teaching. Second, given that the core research question is about how *non-Black* academics signal support for Black people in America, we exclude Black academics from the experiment entirely. This means that we do not burden the Black academics who are already underrepresented in academia. Third, whenever an academic accepts a meeting invite, we will cancel Zoom meetings promptly and politely. This limits the burden on a given academic to just deciding on, and writing to, the reply to one email. Fourth, and as suggested by UBC's ethics board, we will not debrief academics on the fictitious nature of the email after the experiment is over. This reduces the possibility that academics become more suspicious of future emails from genuine students.

3.2 Specifications and Outcomes

Racial Discrimination. To estimate overall racial discrimination for the sample of academic Twitter users, we will use the following specification

$$\text{Accepted}_i = \alpha_{d(i)} + \alpha_{e(i)} + \beta \text{Black}_i + \varepsilon_i \tag{1}$$

where Accepted_i is a dummy variable equal to one if academic i accepts the meeting invitation sent to them, $\alpha_{d(i)}$ are university-by-department of academic i fixed effects (equivalent to randomization strata), and $\alpha_{e(i)}$ are major-by-minor email type fixed effects.[8]

Black_i is a dummy variable equal to one if academic i received an email from a purportedly Black student.[9] We will cluster standard errors at the university-by-department-by-sender name-level, with up to four clusters per university-by-department.

For the more important test of whether discriminatory behavior differs by racial justice tweeting, we will estimate:

$$\begin{aligned} \text{Accepted}_i = {} & \alpha_{d(i)} + \alpha_{e(i)} + \gamma_1 \text{Black}_i \\ & + \gamma_2 \left(\text{Black}_i \times \text{Signaller}_i\right) + \gamma_3 \text{Signaller}_i \\ & + \Sigma_j \theta_j \left(\text{Black}_i \times X_i^j\right) + \Sigma_k \theta_k X_i^k + \varepsilon_i \end{aligned} \tag{2}$$

where γ_2 is the key coefficient, Signaller_i is a dummy variable equal to one for those automatically classified as having signalled support for racial justice, and the set of controls X_i^j (with levels and interactions) will vary across specifications.

The interpretation of γ_2 depends on the set of interacted controls we include in the regression. In particular, the coefficient tells us the signal conveyed by racial justice tweets over and above the information contained in the controls. Without any interacted controls, the signal is

[8] We will test for balance on recipient characteristics. In the event of imbalances, we will check the robustness of our results to adding these characteristics as controls.

[9] We could also control for Female_i and $\text{First-Generation}_i$, but these controls are not necessary given that these randomizations are assigned orthogonally to race.

unconditional – what is the unconditional difference in discriminatory behavior between signallers and non-signallers? In most cases, the more relevant comparison would be conditional. For example, as we have equated signalling status to having *ever* tweeted about racial justice, academics that tweet more are more likely to be signallers than those who tweet less. For a student scrolling Twitter, the relevant question may be: knowing that professors X and Y both tweet a similar amount, what additional information about discriminatory behavior is conveyed by the fact that X tweets about racial justice while Y doesn't? Going further, the student may know the gender and race of X and Y, the institution at which they teach, and perhaps even finer details. What is the signal of a racial justice tweet over and above this information?[10]

The question of whether racial justice tweets provide information on discriminatory behavior depends on what the onlooker already knows about the tweeting academic. To allow for different possible information sets of onlookers, we will estimate the specification with different sets of interacted controls, including:

1. No interacted controls (a test of unconditional signalling – a benchmark with the least real-world relevance given that onlookers will have at least some other information about the academic)

2. Only measures of Twitter activity (number of tweets, number of quote tweets, number of regular retweets, number of replies)

3. Only basic demographics (gender, race/ethnicity, academic position)

4. Twitter activity and demographics

5. (4) + university and department

6. (5) + political measures

[10] A similar logic applies to the coefficient γ_3, where in this the case we are predicting differences in overall response rates to white-sounding emails, rather than differences in differential response rates by race. We will explore and interpret these signals too, though our focus is on γ_2.

We expect (4) and (5) to be the most common information sets of onlookers – once you follow someone on Twitter, these pieces of information are hard to miss. (6) is a less common information set, given that it incorporates less easily visible data like political donations. It may nevertheless match the information set of a student that works with a given academic closely, or reads their tweets regularly.

While Accepted_i is our primary outcome from the audit experiment, we will also estimate the specifications for two secondary outcomes: whether the academic replied or not, and whether the academic replied 'helpfully' or not. A 'helpful' reply could be a reply that turns down the meeting but offers to put the student in touch with someone else, or shares other useful information over email. An acceptance will count as a helpful email, whereas a reply need not be a helpful one.

First-Generation Favoritism. After estimating racial discrimination and its predictors, we will re-estimate equations 1 and 2, replacing Black_i with $\text{First-Generation}_i$. By doing so we can test for discrimination against (or favoritism toward) first-generation students, and more importantly, we can test for whether this behavior differs by signalling status. This tests for whether racial justice support signals broader support for underrepresented groups in academia. While we could carry out a similar exercise with Female_i, for now we plan to report the gender analysis in a companion paper.[11]

Types of Signals. After establishing the informativeness of our automated measure of signalling, we want to ask the question: what types of social media signals are the most informative about discriminatory behavior? For this exercise, we plan to re-estimate equation 2 with alternative definitions (and horse races) of Signaller_i. In particular, we will explore differences in the 'cost' of tweets (e.g. original tweets vs. retweets, longer vs. shorter tweets), differences in social

[11] Relatedly, while first-generation students are underrepresented across academia, the representation of women varies dramatically across fields.

pressure to tweet (e.g. tweets during periods when many people are discussing racial justice vs. quieter periods), differences in visibility (e.g. tweets vs. replies), and variation in the intensive margin (i.e. more vs. fewer racial justice tweets).

For other analyses, we will consider user-level variation. For example, are racial justice tweets less informative for those with greater social image returns (e.g. more followers, working in Democrat-voting states),[12] or for those that are more left-leaning politically (who may be under more pressure to tweet in support of racial justice)?

Beyond the Audit Experiment. After the audit experiment, we plan to survey US-based graduate students to elicit their predictions about the experiment, and to provide a third-party report of the behavior of academics at their institution. We hope to use this additional behavioral measure to repeat the analysis above. We will use the student predictions for an entirely different part of the paper on perceptions of the signalling equilibrium.

4 Pilot

We ran an ethics-approved pilot audit experiment in December 2021 with 1,157 academics from eight US universities just outside of the top-150. Focusing on the 806 academics that tweeted at least once in 2020, 29.8% accepted the meeting, 41.1% replied helpfully, and 45.0% replied at all. These numbers are high enough to persuade us that our emails tend to be taken seriously, despite some lack of personal detail. These numbers would also be high enough to give us reasonable statistical power to detect effects in the full experiment.

While the pilot experiment itself is under-powered, we estimate $\hat{\beta} = -0.022$ (p-value = 0.48, 95% CI: -0.082 to 0.039), similar to the 2.1 percentage point racial gap found for US employers in Kline et al. (2021). More interestingly, we estimate $\hat{\gamma}_2 = 0.073$ (p-value = 0.28, 95% CI: -0.060 to 0.207) without any interacted controls – i.e. signallers discriminate 7.3 percentage

[12] For a data-driven approach to social image returns, we are hoping to estimate a production function for 'likes', allowing racial justice tweets to differentially attract likes across locations.

points less than non-signallers.[13]

While we are agnostic about the results of the full experiment, these pilot results shift our priors a little in the direction of expecting at least some unconditional informativeness of racial justice tweets, with signallers discriminating less against Black students (and perhaps showing more favoritism) than non-signallers.

Power Calculations. We use simulations to predict the statistical power of our full experiment. We assume a sample size of 18,500, 60% of academics auto-classified as signallers (Figure 1), a base meeting acceptance rate of 30% (similar to the pilot), and a 5% test size. This gives us 83% power to detect overall racial discrimination ($\hat{\beta}$) of 2 percentage points (as in Kline et al. 2021) and 77% power to detect a difference in the racial gap of signallers relative to non-signallers ($\hat{\gamma}_2$) of 3.5 percentage points. We have over 99% to detect the $\hat{\gamma}_2$ of 7.3 percentage points estimated in the pilot.

5 Post-Experiment Launch Updates

May 24, 2022. We launched the experiment, sending the first emails, on May 13th. Following the launch, we monitored Twitter for any conversation about the audit study – e.g. an academic who suspects having been audited may post publicly about it, leading others to be aware of the audit. On May 19th, an economist wrote a tweet thread mentioning his suspicion of our audit study as well as advice on running audit studies. This tweet got some traction among economists (e.g. as of today it has 46 retweets and 133 likes).

To minimize the possibility of mass detection (particularly among fields outside of economics), we decided on May 19th to not send any further emails. We made this decision without looking at the data (i.e. without estimating $\hat{\beta}$ and $\hat{\gamma}_2$). At this point we had sent almost 11,500 emails. Given the concern of detection, in the final paper we will check for robustness of the

[13] This uses a definition of Signaller$_i$ defined prior to the pilot, differing slightly from the one described above.

results to excluding academics in fields more likely to have detected the audit (like economics), along with a range of other robustness checks.

We faced one smaller issue when emailing professors in three departments at one of the universities. A few emails in our dataset were incorrectly matched with the academic's remaining data, leading us to automatically send ten emails addressed to the wrong person in the same university. After a coder promptly alerted us of the mistake, we decided not to send emails to the remaining 13 academics in those same departments.

References

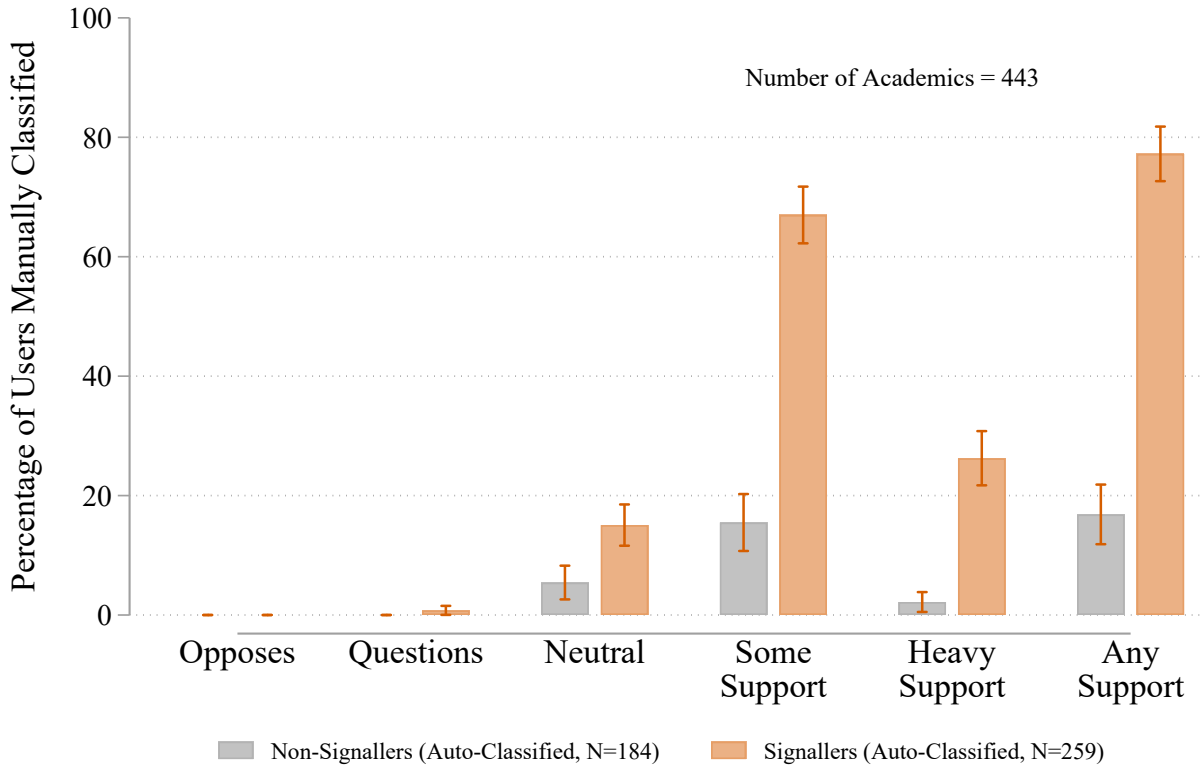
Bertrand, Marianne and Sendhil Mullainathan, "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review*, September 2004, *94* (4), 991–1013.

Kessler, Judd B, Corinne Low, and Colin D Sullivan, "Incentivized resume rating: Eliciting employer preferences without deception," *American Economic Review*, 2019, *109* (11), 3713–44.

Kline, Patrick M, Evan K Rose, and Christopher R Walters, "Systemic discrimination among large US employers," Technical Report, National Bureau of Economic Research 2021.

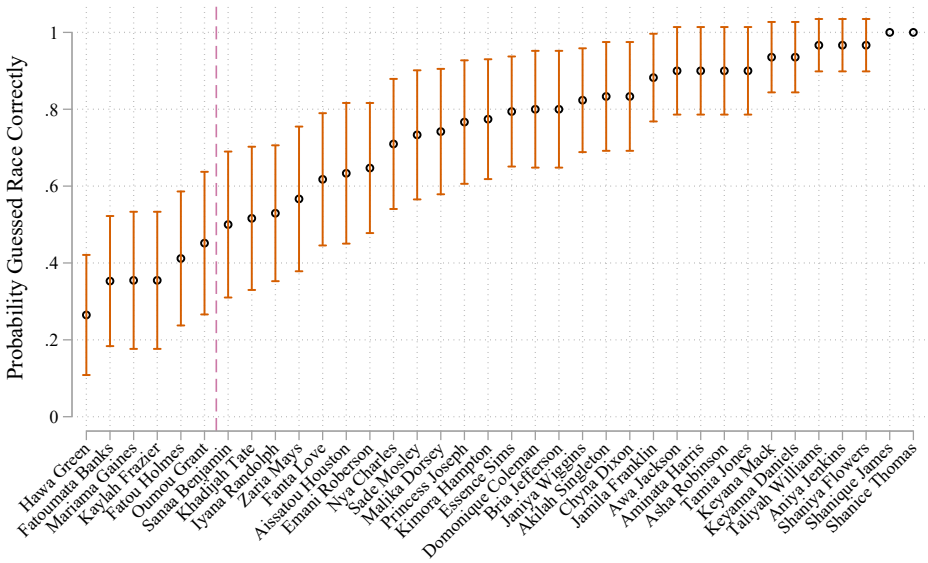
Figures and Tables

Figure 1: Validating the User-level Signaller Measure

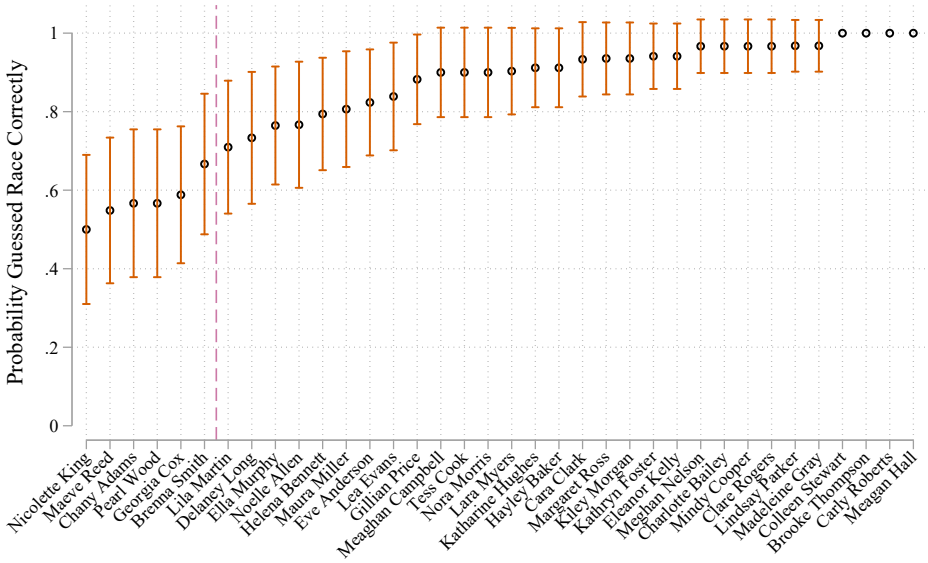


Notes: The figure validates our automated signalling algorithm. The team scrolled through the post-May 2020 Twitter feeds of a random subset of our experimental sample, recording for each user whether they ever: (i) opposed racial justice, (ii) questioned racial justice, (iii) tweeted neutrally about racial justice, (iv) tweeted some support for racial justice, or (v) tweeted heavy support for racial justice. The orange bars include data for the academics we automatically classify as "Signallers." The grey bars include data for the academics we automatically classify as "Non-Signallers." As an example, the orange "Neutral" bar shows the percentage of users that the team *manually* find to have ever tweeted neutrally about racial justice, only among the academics that we *automatically* classify as "signallers." The "Any Support" category shows the percentage of users that ever tweeted some or heavy support. 95% confidence intervals are shown.

Figure 2: MTurk Validation of Racial Distinctiveness: Female Names



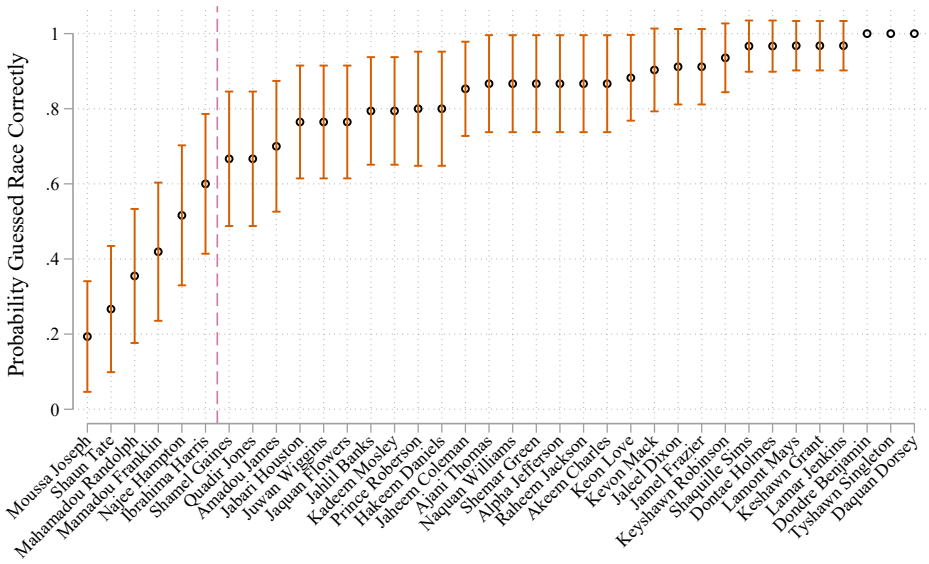
Name: Black, Female



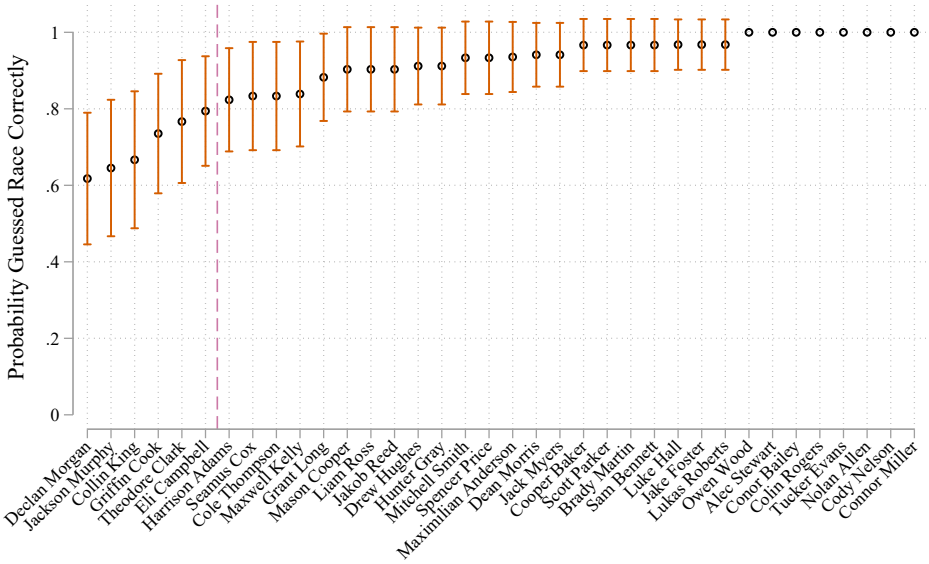
Name: White, Female

Notes: Each full name was guessed by at least 30 MTurkers. The figure shows the fraction of MTurkers that guessed the race correctly for the female names along with 95% confidence intervals. Each MTurker chose one answer from Black, White, Hispanic, or Other. They did not know that no names were chosen to be Hispanic- or Other-sounding. We drop the six names to the left of the vertical dashed line.

Figure 3: MTurk Validation of Racial Distinctiveness: Male Names



Name: Black, Male



Name: White, Male

Notes: Each full name was guessed by at least 30 MTurkers. The figure shows the fraction of MTurkers that guessed the race correctly for the male names along with 95% confidence intervals. Each MTurker chose one answer from Black, White, Hispanic, or Other. They did not know that no names were chosen to be Hispanic- or Other-sounding. We drop the six names to the left of the vertical dashed line.