Virtue Signals

Deivis Angeli* Matt Lowe†

Abstract

We study whether tweets about racial justice predict costly related behaviors. Academics that tweet about racial justice are more likely to favor minority students in an audit experiment, receive higher teaching ratings, work with more Black co-authors, and are more likely to subsequently leave Twitter. Non-academics that tweet about racial justice make larger private donations towards racial justice efforts. However, three pieces of evidence suggest that higher returns to tweeting reduce the predictive value of racial justice tweets. First, tweets became almost completely uninformative during the aftermath of the murder of George Floyd, when more people were tweeting about racial justice. Second, the informativeness of tweets is driven by low-visibility tweet types, like retweets. Third, racial justice retweets are somewhat less informative of donation behavior than private statements of support. Finally, we find that roughly half of surveyed graduate students are overly cynical, believing tweets to be close to uninformative.

*Global Talent Lab.

†University of British Columbia. The audit experiment was made possible by a team of over 100 research assistants led by Carla Colina, Jordan Hutchings, Noor Kumar, Ines Moran, Saloni Sharma, Aurellia Sunarja, Chihiro Tanigawa, Akash Uppal, and Kevin Yu (see Appendix A for a full list), while we thank Catalina Garcia Valenzuela for outstanding research assistance for the other parts of the paper. Our experiment was approved by UBC's Behavioural Research Ethics Board (H20-03758 and H23-02866) and pre-registered in the AEA RCT Registry (AEARCTR-0009359). This research was undertaken, in part, thanks to funding from the Canadian Institute for Advanced Research (CIFAR) and the Canada Excellence Research Chairs program awarded to Dr Erik Snowberg in Data-Intensive Methods in Economics. We thank Vincent Pons and Edgard Dewitte for helping with the FEC data. We also thank Luca Braghieri, Leo Bursztyn, Claudio Ferraz, Patrick Francois, Alex Frankel, Nick Hagerty, Emir Kamenica, George Loewenstein, Rachael Meager, Jamie McCasland, Sam Norris, Nathan Nunn, Chris Roth, Heather Sarsons, Frank Schilbach, Munir Squires, and Carolyn Stein for helpful feedback. First draft: November 2022. This draft: May 2025.

1 Introduction

Social signaling is a central aspect of human behavior. Humans signal their ability to potential employers (Tyler et al. 2000), their ambition to potential partners (Bursztyn et al. 2017), and their customs and parenting prowess to their neighbors (Bursztyn et al. 2020a; Karing 2023). A key question about social signaling is whether, and when, signals are informative. This question is relevant to many economically important domains – whether the informativeness of a job candidate's interview answers for later on-the-job performance (Jacob et al. 2018), or of messages traded on a dating app for later marital prospects, or of an academic's policy advice for their true belief about the policy in question (Morris 2001).

These questions take on extra importance when people signal online. With the advent of social media platforms, a huge swathe of signaling is unverifiable, takes place at little cost (Zhuravskaya et al. 2020), and is observed by large audiences – precisely the conditions that theory suggests should reduce the informativeness of communication (Kartik 2009; Frankel and Kartik 2019). Compounding this, individuals may have particularly strong incentives to misrepresent their values when signaling about politically-charged moral virtues, like opposition to racism or xenophobia (Bursztyn et al. 2020b). This raises a question: how informative are these posts about costly moral behaviors?

In principle, social media posts provide a high-frequency barometer of the attitudes and moral values of an individual's network, potentially influencing second-order beliefs and behaviors (Bursztyn et al. 2020a), whether who to vote for, when to protest, who to work for, or whether to get vaccinated (Fujiwara et al. 2023; Cantoni et al. 2019; Alatas et al. 2019). However, theory makes ambiguous predictions about whether statements on social media reflect the true values of posters. If social media users have preferences for truth-telling (Abeler et al. 2019), or self-persuade (Schwardmann et al. 2022), social media posts will reflect underlying values. Instead, if reputational concerns are sufficiently strong, posts may be completely uninformative of private values (Morris 2001; Frankel and Kartik 2019). In fact, if social media users engage in moral licensing or conscience accounting (Mazar and Zhong 2010; Gneezy et al. 2014), the most vocal users may even be the least likely to behave in line with their social media-reported values offline. Which case describes real-world equilibria? And do social media users recognize what equilibrium we are in?

In this paper, we study communication on Twitter,[1] a platform used by 23% of Americans (PRC 2021), and 450 million users worldwide (BOA 2022). We first focus on the informativeness of the racial justice tweets of 18,514 US-based non-Black academics. We focus on US academia for three main reasons. First, academia plays an important role in shaping the ideology of future elites, and the direction of social movements, particularly in the US (Kaur and Yuchtman 2024; Yuchtman 2025). Second, academics have broad reach and influence on social media – the average academic in our sample has 3,027 followers, while the top-1% have over 32,000. Third, critics argue that US academia's strong left-wing skew is costly (Langbert et al. 2016; McArdle 2025), and that left-leaning campuses breed "virtue signaling and discrimination" (NBC San Diego 2025), culminating in the US's current presidential ad-

[1]The platform is now named X. We use the moniker Twitter throughout, given that the bulk of our data assembly and experimentation occurred before the name change.

ministration targeting elite US institutions for funding cuts (Yang 2025). Each of these points motivates our study of the informativeness of the communication of academics. In particular, such a study can inform whether the influence of specific members of elite academia is toward the private values of those academics, or instead toward a distorted representation of those values. Nevertheless, we show that our core results with academics generalize to a sample of non-academics recruited through an online panel.

For both academics and non-academics, we primarily measure informativeness as the mean difference in behavior between people that do versus do not tweet about racial justice. We verify that in our setting, practically all tweets about racial justice are in support of racial justice-related efforts.[2] Otherwise, we note that we do not take a normative stance on the discriminatory, and other, behaviors we measure. We report correlations between tweets and behavior; we leave the reader to judge which behaviors they deem desirable.[3]

Our analysis proceeds in three steps. First, we characterize informativeness. Using an audit experiment with 11,450 US academics, we compare the discriminatory behaviors of those that tweet about racial justice (Vocal) with those that do not (Silent); then with a broader sample of 18,514 US academics, we study the predictive value of tweets for teaching ratings, co-author and research topic choice, and whether the academic left Twitter. Complementing the analysis of academics, we use an online study with 1,704 Americans, and compare the private donations in support of racial justice made by the Vocal and Silent. Second, we use three approaches to test for how tweet informativeness varies with signaling stakes, including a comparison of the informativeness of tweets made immediately after the death of George Floyd (a high-visibility period) versus those made some months later (a low-visibility period). Third, we explore whether graduate students have sophisticated beliefs about the informativeness of the tweets of academics.

We classify 62% of academics as Vocal – those that have tweeted in support of racial justice. We ran an audit experiment, sending one email to each academic. Each email came from a fictitious student requesting an online meeting about graduate studies. We randomly assigned half of the emails to be sent from a student with a distinctively Black name, and half to be sent from a distinctively White name. We cross-randomized whether the name was distinctively female or distinctively male, and whether the email also mentioned that the sender is a first-generation college student.

We find no evidence of racial discrimination in the full sample of audited academics. Academics accept 30.6% of meeting requests from distinctively White names, and 29.8% from distinctively Black names ($p = 0.31$ for the difference). The lack of racial discrimination in the full audit sample masks heterogeneity: Silent academics are 5.3 percentage points (18%) less likely to accept a meeting with a Black student than with a White student ($p < 0.001$). Vocal academics are 1.9 percentage points (6%) more likely to accept a meeting with a Black student than with a White student ($p = 0.07$). Racial justice tweets are highly informative – we firmly reject equality of the racial gaps in meeting acceptance between

[2] With our focus on informativeness, we assess the *predictive* value of tweets, rather than the *motives* behind tweets. The informativeness of tweets is not necessarily connected to the motives for tweeting: an informative equilibrium may be consistent with all racial justice-tweeters being motivated by social image concerns or all being motivated by prosociality.

[3] Any remaining normative language is unintentional – we would be grateful for readers to bring such language to our attention so that we can revise accordingly.

Vocal and Silent academics ($p < 0.001$). Among the Vocal, the intensive margin is also informative – those that tweet about racial justice below the median are unbiased, while those that tweet above the median are more likely to accept meetings with Black students. More generally, tweets signal broader support for students from disadvantaged groups – Vocal academics are also more likely to favor female students over male students than Silent academics ($p = 0.03$).

One possible interpretation of these results is that public tweets predict meeting acceptance because of subsequent anticipated social image gains: for example, if the academic successfully recruits the Black or female graduate student, they can then signal to others that their team is diverse. We think this channel unlikely, given that the probability of a meeting with an unknown student turning into a hire is low. Even so, to more carefully test for this, we turn to our first-generation student randomization – taking advantage of the fact that a student's first-generation status is much less visible than their race or gender. Against the social image story, we find that Vocal academics also show more favoritism towards first-generation students than Silent academics ($p < 0.001$).

To summarize the treatment of minority-group students, we pool the 7/8 of the audit emails from any minority group (Black, female, or first-generation) and compare them with the 1/8 of the audit emails from White males with no mention of first-generation status. Vocal academics are 27% less likely to accept a meeting with a White male student than with an underrepresented minority student ($p < 0.001$), while we cannot reject the null that Silent academics treat the two groups equally. Going further, while minority students are more likely to get a positive response from Vocal academics than from Silent academics, the reverse is true for White male students. This gives a force for sorting – with White males advised more by Silent academics, and minority students advised more by Vocal academics.

Our finding that racial justice tweets are unconditionally informative extends to conditional comparisons. In particular, it is not just the case that Vocal academics behave differently than Silent academics, they also behave differently conditional on their gender, race/ethnicity, academic position, university, department, political views, how much they tweet, and the vocality of academic accounts they follow. Controlling for these factors, Vocal academics have a Black-minus-White meeting acceptance gap that is roughly five percentage points larger than Silent academics ($p < 0.01$). Racial justice tweets provide information about discriminatory behaviors over and above these observables, making tweets informative even for students that know a given professor reasonably well. In fact, Vocality is more predictive of racial discrimination than each of these other observable factors.

Detection is a particular concern when auditing academics. Encouragingly, our findings are similar when we consider subsamples of our audit data where detection is less plausible – for example, our results are similar when dropping all social scientists. Otherwise, we turn to a broader set of behavioral outcomes, where detection is no longer a concern.

We link our full population of 18,514 tweeting academics to Rate My Professor teaching evaluations, and to their research topics and co-authors using the OpenAlex bibliographic catalogue. We also track whether each academic has since left Twitter, and we document that academics in more left-wing departments are much more likely to have left. Vocality is highly predictive of these behaviors. Relative

to Silent academics, Vocal academics have 0.11 SD higher teaching ratings, they are more than twice as likely to work on racial justice-related research, they have 41% more Black co-authors, and they are 60% more likely to have left Twitter. These differences attenuate, though remain statistically significant, after adding controls.

Our results for academics show that racial justice tweets are highly predictive of a range of important behaviors. We ran a supplementary donation study to explore external validity, and to have a setting in which our measured behavior is fully private – connecting more tightly with models focused on the signaling of private types. We recruited 1,704 non-Black Twitter-using Americans to answer a series of questions on demographics, social media usage, and well-being, followed by a question asking how much of their $5 survey incentive they would like to donate to the National Association for the Advancement of Colored People (NAACP, a civil rights organization), and a link to an NAACP tweet, asking if they would like to retweet. The results from the donation study corroborate those in the audit experiment: participants that retweet the NAACP tweet donate 53% more to the NAACP than those that do not ($p < 0.001$). Results using past tweets are identical.

Having demonstrated the informativeness of racial justice tweets, we explore whether high signaling stakes are associated with information loss, as predicted by theory (Appendix B, Frankel and Kartik 2019). For our first test, we use time-series variation in visibility induced by the murder of George Floyd on May 25th, 2020. We distinguish between two periods: the period immediately after the murder in which tweets about racial justice are common, and in which the pressure for academics to speak out was high (the during-period); and the period several months later in which racial justice tweets were less common (the post-period). We find that post-period racial justice tweets are consistently more informative about audit-measured discriminatory behavior than during-period racial justice tweets. Most starkly, in a regression in which we include both during- and post-period vocality as predictors of racial discrimination, we find that (i) post-period vocality is highly informative, and (ii) during-period vocality is completely uninformative. This suggests that tweets about racial justice reveal more about an academic's offline discriminatory behaviors when made in a lower-scrutiny environment.

For our second test of information loss, we use variation in the visibility of different types of tweets. In particular, original tweets are fully composed by the author, and thus attribute high credit and visibility to the author. In contrast, when an academic retweets, they amplify the visibility of the original poster much more than the academic's own visibility. In a regression in which we allow both types of tweets to predict racial discrimination, we find that only retweets are informative. Conditional on the racial justice retweets, original tweets about racial justice are completely uninformative. This suggests that differences in credit and visibility drive the differences in informativeness across types of tweets, as opposed to differences in costs – since the effort costs of crafting an original tweet are higher than those for retweeting.

For our third test, we use an additional feature of the donation study. Besides deciding whether to retweet the NAACP tweet, participants also stated their private support for the activities of the NAACP described in that same tweet. The two questions about retweeting and private support were shown in

random order. We transform the private support variable so that we have two binary variables with the same mean: a dummy variable equal to one if the participant retweeted the NAACP retweet, and a dummy variable equal to one if the participant privately reported support for the NAACP above a threshold. This allows us to test for whether behavior on social media is more or less informative about the NAACP donation than a private statement of support. While both are highly informative, we find suggestive evidence in support of the model: the NAACP retweet is 23% less informative than the binary measure of private support ($p = 0.19$). This information loss appears to come from low types pooling with high (i.e. low-support types deciding to retweet), rather than "shy" high types pooling with low (by deciding not to retweet).

While each of our three tests of information loss have weaknesses, collectively they provide consistent support for the key prediction of theory. In particular, the aspects of social media that serve to increase visibility and the strength of signaling concerns also serve to reduce informativeness. Nevertheless, our results show that social media posts in support of racial justice span the full range from fully uninformative to highly informative.

In the final part of the paper, we characterize beliefs – racial justice tweets may generally be informative, but do audiences realize this? We ask 1,752 US-based graduate students to predict the racial discrimination of Silent and Vocal academics. In our survey we find that the average student has somewhat accurate beliefs about informativeness – the Vocal-Silent difference in racial discrimination. However, roughly half of students guess below the lower end of our 95% confidence interval, believing that tweets are close to uninformative. According to theory, such beliefs can help sustain an informative equilibrium, by reducing signaling stakes.

Our paper builds on three strands of research. First, in emphasizing an informational benefit of social media, we build on the broader evidence of the welfare effects of social media – whether effects on happiness and depression (Mosquera et al. 2020; Allcott et al. 2020; Braghieri et al. 2022), or political outcomes (Zhuravskaya et al. 2020; Song 2022). Some papers demonstrate that learning from peers on social media influences economic decision-making (Bailey et al. 2018, 2022); we show that the posts of peers can be used to learn about peers' private moral behaviors. The scope for learning is meaningful: a Black student that learns that a given professor has tweeted about racial justice should rationally increase their expectation of securing a meeting with that professor by 37% (assuming the student's prior was that the professor had not tweeted about racial justice). A fundraiser from the NAACP should expect those that tweet about racial justice to donate roughly 50% more, when prompted, than those that do not tweet.[4]

Second, by characterizing a real-world communication equilibrium, we relate to models of strategic information transmission (Crawford and Sobel 1982; Austen-Smith 1990; Loury 1994; Morris 2001; Kartik 2009; Frankel and Kartik 2019). While most existing tests of these models use lab experiments (Cai and Wang 2006; Blume et al. 2020), Braghieri (2024) uses an online experiment to explore how

[4]While we focus on the informativeness of social media posts, a separate question is whether having the option to declare support on social media *causally* reduces costly support offline. On this question Bonheur (2023) finds in a lab experiment that the option to post support for racial justice reduces donations to racial justice charities. Our donation study does not permit a similar test, as all participants answered the donation question before the retweet question.

visibility affects the self-reported sensitive political attitudes of college students. He finds that students' public statements are less predictive of demographics and incentivized behaviors than their private statements, and that audience naïveté amplifies the information loss. We find complementary results in the naturalistic context of social media and support for racial justice, suggesting that the core findings of Braghieri (2024) extend to real-world communication platforms.

Third, in showing that posts on social media predict heterogeneity in discrimination, we contribute to the large economics literature on discrimination. Recent contributions emphasize heterogeneity in discriminatory behaviors across firms (Kline et al. 2022), while our student survey relates to work that emphasizes the importance of sorting away from discriminators (Becker 1957; Charles and Guryan 2008). In focusing on academia, we complement Milkman et al. (2012), who find that academics are more likely to discriminate against minorities when students request a meeting in one week rather than that day, and Ajzenman et al. (2023), who find that economists on Twitter discriminate against low-ranked and Black students, and in line with our results, male students. We also contribute new evidence on favoritism towards first-generation students, an understudied dimension of disadvantage (Stansbury and Schultz 2023; Stansbury and Rodriguez 2024). Above all, we build on Pager and Quillian (2005). With a sample of 156 employers, they find that self-reported attitudes toward hiring ex-offenders are not predictive of hiring behavior in an audit experiment. Some of our analysis is similar in spirit, though we increase power with a sample size that is over 70 times larger, and we focus on statements made on social media, where social signaling concerns are more important. Despite these social signaling motives, we reach the opposite qualitative conclusion to Pager and Quillian (2005): words predict behavior.

2 Data, Context, and Methods

2.1 Audit Experiment

To assemble a sample of academics for the audit experiment, we first listed all research academics in PhD-granting departments in the top-150 universities according to the 2019 US News University Rankings. We found over 125,000 research academics in this step. In the second step, we found the subset of academics with Twitter accounts. We then recorded the academic's email address, position (Assistant, Associate, or Full Professor), gender (Male, Female, Other), and a best guess of the academic's race or ethnicity (White, Black, East Asian, South Asian, Hispanic, Other, Uncertain). We dropped any academics without an email address available online, those with an explicit policy of not answering prospective student's emails listed on their website, and non-research-active or non-professor academics (e.g. emeritus professors, post-docs). This leaves us with a sample of 28,302 tweeting research academics.

We imposed six final eligibility criteria. First, we kept 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. Second, we required a minimal level of public Twitter activity, keeping only the academics with at least five public tweets in 2020. Third, we dropped academics with lab-oriented Twitter accounts with no personal tweet content. Fourth, we

dropped a few academics in departments for which our email templates do not fit well (e.g. Theater, Education departments oriented only towards practitioners). Finally, we dropped Black academics and those with race/ethnicity coded as Uncertain, given that the core research question is about how *non-Black* academics signal support for Black people in America. This leaves us with a final experimental sample of 18,514 academics.[5]

We planned to audit the full set of 18,514 academics, staggering the experiment according to the end of the spring term in each university. Due to a detection-related concern explained below, we stopped the experiment after emailing 11,450 academics – these 11,450 academics comprise the final audit sample.[6] Audited academics are similar to non-audited academics on most observables (Table A1), suggesting that our audit results might generalize to the broader population of tweeting academics. The exceptions are that, on average, the audited academics are at slightly lower-ranked universities (59th of 150 vs. 52nd), tweet somewhat less (1,022 tweets vs. 1,207), have fewer Twitter followers (2,781 vs. 3,426), and make smaller political contributions.

Twitter Data and Political Preferences. We used Twitter's academic API to download user-level and tweet-level data for each of the academics with a Twitter account. We mostly use user-level data as of May 10, 2022, just prior to the launch of the audit experiment. We also scraped the full list of accounts followed by each academic during July 2022. We use this data to create a proxy for the vocality of an academic's network: the fraction of in-our-sample academics followed that are vocal about racial justice (with vocality defined below).

The tweet-level data includes the full text of all original tweets, replies, quote tweets, and 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 May 25, 2020, the date of the murder of George Floyd. We use March 27, 2022 as the last date as we began to download the tweets shortly after.

We collected three measures of political preferences. First, we web-scraped Blindspotter to get a measure of the political slant of the news each user interacts with on Twitter. Second, we used and updated data from UCSD on the Twitter accounts of politicians in the US Senate and House of Representatives. We then linked this data with the accounts followed by each academic to calculate (i) the number of political accounts each academic follows, and (ii) the percentage of those accounts that are Democrats. Third, building on Bouton et al. (2022), we linked each academic with their FEC-reported political contributions made from January 1, 2020 to March 27, 2022 (see Appendix D for details), mirroring the period for which we have tweets. We link 29.9% of our academics with at least one FEC contribution, while Bouton et al. (2022) find that 8.5% of the adult US citizen population contributed in 2019 or 2020. Vocal, Silent, Black and non-Twitter-using academics each overwhelmingly support the Democratic party (Figure 1), consistent with other work on the liberal slant of academia (Langbert et al. 2016). Even so,

[5] Our ex ante power calculations motivated us to form such a large sample – with a sample of this size, we estimated that we would have 83% power to detect overall racial discrimination of two percentage points (in line with Kline et al. 2022), and 77% power to detect a difference in the racial gap of Vocal relative to Silent academics of 3.5 percentage points.

[6] The reduction in sample size for the audit is the most important deviation from our pre-analysis plan, see full details in Appendix C.

vocality about racial justice is a strong signal of support for Democrats – Vocal academics are 81% more likely than Silent academics to have given to Democrats (36.2% vs. 20%). We show below that vocality is predictive of discriminatory behaviors even conditional on these measures of political preferences.

Going beyond the audit experiment, we link our academics to teaching ratings, research topics and co-authors, and whether they have left Twitter. We defer the description of these measures to Section 3.2.

Measuring Racial Justice Tweets. A key part of our paper is determining which individuals tweeted in support of racial justice on Twitter and which did not. Given the large sample size of academics, we automated this classification based on the words and phrases included in tweets. We pre-registered our core measure, Vocal_i (with $\text{Silent}_i = 1 - \text{Vocal}_i$), 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:[7]

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 / justiceforbreonnataylor / 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. In the full experimental sample of 18,514 academics, we automatically classify 63% as vocal.

Our automated approach 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 to be vocal 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 used a richer manual measure of signaling status for a random subset of our experimental sample ($N = 450$). For this random subsample, one or two team members spent up to five minutes scrolling through the academic's tweets, coding 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. Encouragingly no academic in this random subsample ever tweeted in opposition of racial justice, and only two academics questioned efforts to promote racial justice (Figure A1). It follows that there are practically no false positives – academics auto-classified as supporting racial justice that in fact question or oppose efforts to promote racial justice.

[7] For retweets and quote tweets, we include the text in the tweet being retweeted. Throughout, we also include words and phrases found in any expanded hyperlinks included in tweets.

Second, while 13% of those auto-classified as Silent are manually-coded as having shown any support on Twitter for racial justice, the figure is 74% for the Vocal. This validation shows that false negatives are not too common, and that our automated measure of signaling has a large first stage for the richer manual measures of signaling.

Vocal_i is our pre-registered measure of whether an academic has signalled support for racial justice on Twitter. Nevertheless, we also use a continuous measure (the percentage of tweets auto-classified as racial justice-related) and measures that treat each type of tweet separately (original tweets, replies, quote tweets, and retweets).

We show what factors predict tweeting behavior in Figure 2, using one multivariate regression. Female academics are almost 10 percentage points more likely to tweet about racial justice, as are academics that gave money to Democrats. Comparing fields, social scientists are the most likely to tweet about racial justice, and engineers the least. Comparing universities, racial justice tweeting does not differ much by rank, nor is it more common in universities with more Black undergraduate students. Unsurprisingly, academics who tweet more in general are far more likely to ever tweet about racial justice. We show below that racial justice tweets are informative even after conditioning on how much academics tweet.

Audit Experiment Design. We chose 120 racially distinctive names largely following the approach of Kessler et al. (2019). We created one gmail account for each of the 120 full names. We sent one email to each academic in our audit sample, purporting to be an undergraduate student interested in graduate studies at the academic's university. The core randomizations were: distinctively Black vs. distinctively White name of sender (50:50), distinctively male vs. distinctively female name of sender (50:50), and a sentence mentioning that the sender is a first-generation college student or not (50:50), all stratified on university-by-department.

We randomly chose 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 ensured that all tweeting academics in the same department at the same university assigned to receive an email from, for example, a Black male, would receive an email from the exact same Black male.

In the final step, we randomized 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 chose the main text of the email from 12 possible variants. We then randomly chose 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 randomized 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.[8] For an example email format, see Appendix E, while for a discussion of the ethics of the experiment, along with further audit experiment details, see Appendix F.

Coding replies. While we automated the sending of emails, the team classified each email reply manually as either accepting or declining the meeting request. We pre-registered meeting acceptance as our main outcome, rather than whether the academic replies, given that meeting acceptance is more welfare-relevant for students than replies. We consider meeting acceptance to be a costly prosocial behavior – costly because acceptance effectively commits the academic to a 20-minute or so online meeting, and prosocial given that meetings with students outside of an academic's institution tend to benefit the student rather than the academic.

Detection on Twitter. We began sending emails in May 2022 with the intention of emailing all 18,514 academics over a two-week period. Following the launch of the experiment, we monitored Twitter for any conversation about the audit experiment. On May 19th, an economist wrote a tweet thread mentioning their suspicion of an audit experiment as well as advice on running audit studies. This tweet got some traction among economists, with 46 retweets and 133 likes by May 24th.[9] To minimize the possibility of mass detection (particularly among fields outside of economics), we decided on May 19th to not send any further emails. Given this pause, the final audit sample includes 11,450 academics. As reported earlier, audited academics are similar on most observables to the non-audited academics (Table A1). We also use a series of analyses below to show that our findings are unlikely to be explained by academics detecting that the email was part of an audit experiment. In addition, we later turn to a broader set of academic outcomes (including teaching ratings), and the donation study, where deception and detection are ruled out by design.

2.2 Donation Study

We ran the donation study in late-2023 and early-2024. While less naturalistic, we designed the donation study to address three weaknesses of the audit experiment: (i) our donation measure of racial justice support is fully private, as opposed to email replies which are visible to the recipient, (ii) there is no deception, and (iii) we can cleanly compare the informativeness of tweets with private statements of support.

We partnered with CloudResearch to recruit participants. We made a screener survey available to non-Black American Twitter users in CloudResearch's pool. 6,100 participants completed the screener and passed an attention check. Of these participants, we sent the donation survey to the 2,096 participants that (i) are 18 years or older, (ii) consented to sharing their Twitter handle and having their tweets linked

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

[9] As of May 23, 2023 (one year later), the tweet had 44 retweets and 133 likes.

to their survey response, and (iii) have a public Twitter account with at least 10 tweets. 1,704 (81.3%) of the 2,096 participants completed the donation survey and passed two attention checks (see Table A2 for summary statistics). We presented our aim as being "to understand the relationship between social media, attitudes, and well-being."

The survey begins with demographic questions, followed by questions on social media usage, and then questions on psychological well-being. The survey ends with the three most important questions. First, we tell the participant that we are supporting the NAACP, explain that the NAACP's "vision is to have a world without racism where Black people enjoy equitable opportunities in thriving communities," and give a short description of their work (Figure A2). We then say that "As compensation for answering our survey, you will receive an additional $5. You can choose to give any or none of that $5 to the NAACP. As with the other questions, your answer will not be shared with anyone." The amount donated serves as our private measure of support for racial justice. Participants donate $1.56 on average, with 61.3% donating a positive amount.

Following the donation question, we give participants the opportunity to retweet an NAACP tweet (Figure A3), and ask the participant how much they support the NAACP on a scale from 0 to 100 (Figure A4). These two questions are asked in random order.[10] Each question has a 30-second timer before participants can advance. We did this to give participants enough time to go into Twitter and retweet, and we applied the timer to both questions to avoid attention-based confounds – for example, retweet behavior could differ from private statements of support if participants pay more attention to the retweet question, rather than due to forces emphasized by our model. In the week following the survey, we manually verify on Twitter which participants retweeted the NAACP tweet (13% retweeted). As in the audit experiment, this gives us a binary variable, Vocal NAACP$_i$, equal to one if the participant retweeted the NAACP tweet and zero otherwise.

Improving on the audit experiment, the donation study allows us to compare the informativeness of tweeting with that of a private measure of support. To create a measure of private support comparable to the binary variable Vocal NAACP$_i$, we define a binary variable, Support NAACP$_i$, equal to one if support is greater than a cutoff, with the cutoff chosen such that the number of participants at or above the cutoff is as close as possible to the number of participants that retweeted the NAACP retweet. In our case, the cutoff is 79, with 221 participants declaring support at or above the cutoff, as compared with 222 participants that retweeted the NAACP tweet.

We designed this approach of comparable binary variables to address two concerns. First, a continuous measure of support could be more predictive of donations than a binary retweet variable because the measure is richer (being continuous), rather than through the forces emphasized in the model. Second, if we had asked participants a binary yes/no question about support for the NAACP, we would face problems if the fraction that answered yes was different from the fraction that retweeted the NAACP tweet – in particular, we could find that one variable is more informative than the other even in the absence of

[10] We do not estimate statistically significant order effects ($p = 0.75$ for the binary retweet variable, $p = 0.4$ for the continuous support variable). As a result, we pool the responses in our analysis.

information loss.[11]

In addition to the NAACP retweet measure of vocality, ideally we would also create a measure that exactly parallels the measure we used for the audit experiment. However, changes to Twitter's academic API made it prohibitively costly to replicate the same measure. Instead, we used a manual approach. Among the participants eligible to receive the donation survey, we manually scrolled through tweets made between May and October 2020 (the height of tweeting about racial justice) for the 854 participants that (i) joined Twitter before May 2020, (ii) have at least five tweets during May to October 2020, (iii) do not tweet in a foreign language, and (iv) do not have so many tweets that we are prevented by Twitter from scrolling back to October 2020. We coded whether each user ever opposed, questioned, tweeted neutrally, tweeted some support, or tweeted heavy support for racial justice efforts. We then constructed Vocal_i as a binary variable equal to one if the user had any tweet in support of racial justice efforts and no tweets opposing or questioning.

2.3 Specifications and Outcomes

Audit Experiment. To estimate overall racial discrimination of academics on Twitter, we 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 accepted the meeting invitation,[12] $\alpha_{d(i)}$ 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.

Black_i is a dummy variable equal to one if academic i received an email from a purportedly Black student. We cluster standard errors at the university-by-department-by-sender name-level, the level of treatment, with up to four clusters per university-by-department. Balance checks are consistent with the randomization being carried out successfully (Table A3).

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

[11] To see this, consider the following example. Suppose that participants answer "yes" to the support question whenever their racial justice support type $\eta > k_1$, and fewer people tweet about racial justice, only those with $\eta > k_2 > k_1$. Here tweet activity is a simple function of racial justice types – there is no information loss due to different types pooling on a given action. Nevertheless, the informativeness of the two binary variables (whether retweeted and whether supported) can differ mechanically: $\mathbb{E}[\eta|\eta > k] - \mathbb{E}[\eta|\eta \leq k]$ can increase in the cutoff k, decrease, or stay unchanged, depending on the distribution of η. For example, measured informativeness is constant in k when η is uniformly distributed. When normally distributed, measured informativeness is decreasing in k when $k < \mathbb{E}[\eta]$, but increasing in k when $k > \mathbb{E}[\eta]$. In this sense, the difference in informativeness using the binary measure may reflect the distribution of underlying racial justice support, rather than the information loss described in our model.

[12] We last checked the email accounts eight weeks after we sent emails. The vast majority of responses came much earlier.

estimate:

$$\begin{aligned}\text{Accepted}_i = \alpha_{d(i)} + \alpha_{e(i)} + \gamma_1 \text{Black}_i \\ + \gamma_2 \left(\text{Black}_i \times \text{Vocal}_i\right) + \gamma_3 \text{Vocal}_i \\ + \sum_j \theta_j \left(\text{Black}_i \times X_i^j\right) + \sum_k \eta_k X_i^k + \varepsilon_i\end{aligned} \tag{2}$$

where γ_2 is the key coefficient, Vocal_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 with Black_i) varies 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, γ_2 answers the question: what is the unconditional difference in discriminatory behavior between Vocal and Silent academics? With interacted controls, γ_2 delivers the Vocal-Silent difference in discriminatory behavior conditional on those controls. This is the more relevant measure of informativeness in many practical applications, given that audiences will often already know a set of observables about the tweeting academic (e.g. their gender, race, and field of study).

To allow for different possible information sets of onlookers, we estimate specification 2 with different sets of interacted controls, with the full set of controls including measures of Twitter activity, basic demographics, university and department fixed effects, measures of political preferences, and vocality of the accounts the academic follows.

Donation Study. To estimate the informativeness of racial justice tweets for donations, we use the following specification:

$$\text{Donation}_i = \alpha + \phi \text{Vocal NAACP}_i + \sum_k \zeta_k X_i^k + \varepsilon_i \tag{3}$$

where Donation_i is the amount in dollars donated to the NAACP (from \$0 to \$5) and Vocal NAACP_i is a dummy variable equal to one if participant i retweeted the NAACP tweet. We also estimate the specification using the manually-coded measure of vocality based on past tweets, Vocal_i, instead of Vocal NAACP_i. As with the audit analysis, we estimate unconditional informativeness by omitting the covariates X_i^k, and for conditional informativeness, we include the following covariates: number of posts per month since joining Twitter (replaced with the number of posts made from May to October 2020 for the specification with Vocal_i), a dummy variable for female, dummy variables for six income categories (less than \$25,000, \$25,000 to \$49,999, etc., with the final category being \$150,000 or more), dummy variables for five categories of political views (very liberal, liberal, moderate, conservative, and very conservative), a dummy variable for non-White, a dummy variable for Hispanic, and age. We estimate robust standard errors.

To test for the informativeness of the binary support variable, we estimate the same specification, replacing Vocal NAACP_i with Support NAACP_i (recall that the latter is the binary version of the 0 to 100

private measure of support for the NAACP). We then use a stacked specification to test for equality of the $\hat{\phi}$ coefficients estimated separately for Vocal NAACP$_i$ and Support NAACP$_i$.

3 What Do Racial Justice Tweets Signal?

3.1 Discrimination in the Audit Experiment

Overall Discrimination. We do not detect racial discrimination in the full sample – academics accept 30.6% of emails from distinctively White names, and 29.8% of emails from distinctively Black names ($p = 0.31$ for the difference, Figure 3). Tweeting academics discriminate against Black individuals less than the 2.1 percentage points found among large US employers (Kline et al. 2022), and less than the 8 percentage points found in a representative sample of over 6,000 academics audited in 2010 (Milkman et al. 2012).

Unconditional Informativeness. Silent academics are 5.3 percentage points (18%) less likely to accept a meeting with a Black student ($p < 0.001$, Figure 3), whereas Vocal academics are 1.9 percentage points (6%) *more* likely to accept a meeting with a Black student ($p = 0.07$). The difference in discrimination is then 7.2 percentage points ($p < 0.001$). In this setting, academic Twitter is racially unbiased overall because the pro-Black bias of the Vocal academics almost exactly offsets the anti-Black bias of the Silent academics.[13]

Vocal and Silent academics treat emails from distinctively White names similarly – while the raw White student acceptance rate is 0.8 percentage points higher for Vocal than for Silent academics, we cannot reject the null of no effect ($p = 0.65$). If these results generalize to other faculty-student interactions, we would conclude that White students have similar experiences with Silent and Vocal academics, while Black students are 37% more likely to secure a meeting with a Vocal than with a Silent academic.

Less Bias or More Support? Racial justice tweets signal greater support for Black students. They also signal less overall bias – the absolute racial gap in meeting acceptance is 1.9 percentage points for Vocal academics versus 5.3 percentage points for Silent academics. Are tweets primarily a signal of support for minorities and disadvantaged groups? Or a signal of less bias? To answer this question, we turn to the variation in student gender, given the underrepresentation of women across many academic fields.

Academics overall are 4.3 percentage points more likely to accept meetings from distinctively female names ($p < 0.001$, panel (a), Figure 4), similar to evidence of gender gaps elsewhere in academia – in the past 20 years, women were more likely to be selected as members of prestigious national academies than men with similar records (Card et al. 2023). Vocal academics discriminate more in favor of women than

[13] While we pre-registered meeting acceptance as our main outcome, our findings are similar if we look at effects on whether the academic replied at all (Figure A5). The one qualitative difference is that we cannot reject the null hypothesis that Vocal academics replied to Black and White students equally. This suggests that the favoritism towards Black students in Figure 3 comes from the margin of academics accepting a meeting rather than declining while still replying.

Silent academics ($p = 0.03$) – unlike racial discrimination, here the Vocal academics are *more* biased than Silent academics.

Visibility. One key conceptual question is whether Vocal academics behave differently than Silent academics because of differences in underlying private types, or because of differences in social image concerns. The latter may play a role given that (i) the academic's email response is visible to the student, and (ii) there is a small probability that the meeting would lead to the student joining the academic's team as a research assistant or graduate student, enabling the academic to signal to others that they care about diversity. We use the donation study to address (i), in Section 3.3. To address (ii), we note that this story predicts that informativeness should fall when the student's disadvantaged status is less observable. For this, we can use the first-generation status randomization – as first-generation status is less observable to colleagues than a student's gender and race.[14][15]

Overall, tweeting academics are 3.2 percentage points more likely to accept meetings with students that reference their first-generation status ($p < 0.001$, panel (b), Figure 4). This finding echoes recent audit evidence that minorities benefit from explicitly mentioning their demographic identity when requesting help (Kirgios et al. 2022). The first-generation advantage is driven entirely by the Vocal academics, who favor first-generation students by 5.8 percentage points ($p < 0.001$). This speaks against the argument that Vocal academics behave differently solely because of anticipated social image returns of working with a diverse set of graduate students.

White Males and Sorting. One way to summarize our results across the three dimensions (race, gender, and first-generation status) is to group together the 7/8 of students that belong to any underrepresented group (Black, female, or first-generation) and compare their meeting success rate with the remaining 1/8 of students: White males with no mention of first-generation status. While this exercise unmasks new findings, we caveat that the comparison of 1/8 versus 7/8 of the emails is lower-powered than our earlier comparisons of 1/2 versus 1/2.

Vocal academics are 9.1 percentage points (27%) less likely to accept a meeting with a White male student than with an underrepresented minority student ($p < 0.001$, Figure 5). This large gap was masked in the previous figures, where given the design of the experiment, 75% of the disfavored category (e.g. White) belonged to at least one underrepresented group. In contrast, we cannot reject the null that Silent professors treat the two groups equally ($p = 0.64$), although given the lack of statistical power, the 95% confidence interval is large – ranging from 5.1 percentage points discrimination against underrepresented minorities to 3.1 percentage points discrimination against White males.[16]

[14] And while in the real-world first-generation status is likely correlated with visible attributes like race, in our experiment the first-generation status randomization is orthogonal to the racial and gender-distinctiveness of names.

[15] Of course, an academic could tell their colleagues that a research assistant is first-generation, making that private characteristic public. But this argument applies to *any* private action that the academic is consciously aware of – i.e. an academic that takes some private action in support of racial justice can always later make that action visible to others (and credibly, in some cases).

[16] We also note that the estimate of the gap from the specification without strata and email type fixed effects ($28.7 - 26.6 = 2.1$) is somewhat larger in this case than the estimate after including strata and email type fixed effects (1 percentage point).

Next, we compare the overall level of meeting acceptances between Vocal and Silent academics. Recall that Vocal academics accept more meeting requests from White students (Figure 3), male students, and non-first-gen students (Figure 4), than Silent academics. These differences are driven by the presence of underrepresented minority students in each of those three groups. In particular, when we look at only White male non-first-generation students, Vocal academics accept 4.4 percentage points fewer meeting requests than Silent academics (Figure 5, $p = 0.06$).

Our pattern of results reveal a force toward sorting. Other things equal, in the real world White male students would be more likely to actually meet with Silent professors, while underrepresented minority students would be more likely to meet with Vocal professors.

The Intensive Margin. Our results show signaling along the extensive margin: racial justice ever-tweeters discriminate more in favor of students from underrepresented groups than racial justice never-tweeters. To explore signaling on the intensive margin, we split the set of Vocal academics into two groups: those with a percentage of racial justice tweets below versus above the median. We call these two sets of academics the "Rarely Vocal" and the "Regularly Vocal."

Rarely Vocal academics are close to unbiased, accepting 30.4% of meeting requests from White students, and 31% from Black students ($p = 0.39$ for the difference using the specification with strata and email fixed effects, Figure A6). Regularly Vocal academics favor Black students by 3.2 percentage points ($p = 0.01$). Hence, academics who tweet more often about racial justice show more support for Black students than academics who tweet less often about racial justice.

In contrast, the Regularly Vocal are no more likely than the Rarely Vocal to favor female and first-generation students (Figure A7). So while the extensive margin of racial justice tweets signals support for three types of marginalized students, the intensive margin carries a narrower informativeness: only signaling about racial bias.

Conditional Informativeness. Our results so far make unconditional comparisons between Vocal and Silent academics. But as referred to earlier, Vocal and Silent academics differ along many dimensions other than their racial justice tweets: for example, Vocal academics are more likely to be female and Democrat-leaning (Figure 2). More mechanically, they tweet more often. Is vocality merely proxying for these other dimensions, or does vocality predict racial discrimination above and beyond these other observables? We answer this question in Figure 6.

The far-left coefficient replicates our earlier result: Vocal academics discriminate against Black students 7.3 percentage points less than Silent academics (when including strata and email fixed effects). The coefficient falls slightly, to 7 percentage points, when conditioning on the number of tweets. Vocality is then not just capturing the fact that Vocal academics tweet more, and that those that tweet more discriminate less. The coefficient falls gradually as we add more controls: gender, race/ethnicity, position, university, and department fixed effects. With these controls the coefficient falls to 5.3 – smaller,

As with the previous figures, the p-value of 0.64 comes from the specification with strata and email type fixed effects, while without strata and email type effects, the p-value is 0.3.

though still statistically significant at the 1% level. These controls likely constitute a common information set of prospective students. Given this, we will refer back to this particular measure of conditional informativeness below.

While students are perhaps less likely to know the political views of prospective advisors, controlling for our three measures of political views barely changes the informativeness of racial justice tweets. Nor does controlling for the racial justice vocality of the academic's network – proxied by the fraction of academics followed classified as Vocal. So even onlookers with richer information sets can learn from the tweets of academics.

Most other factors we measure do not predict discriminatory behaviors enough for us to reject the null hypothesis of no effect. This is the case when including each interacted factor one-by-one or when including all interacted factors at once (Figure A8). Focusing on the latter, only two variables are statistically significant predictors of discrimination at the 5% level – Vocal academics discriminate against Black students 5.6 percentage points less ($p = 0.01$) and academics at universities ranked in the top-50 discriminate 5.2 percentage points less ($p = 0.02$), other things equal. In terms of the magnitude of the point estimate, Vocality is in fact the single most predictive variable of racial discrimination, whether considering unconditional or conditional predictiveness. However, we cannot statistically reject equality of the Vocal interaction with several other interaction terms.

Detecting Detection. Suspicion of the fictitious nature of our emails is more likely in our setting than others, since academics essentially invented the audit method. Two patterns of detection and behavior would make our results particularly misleading. First, if emails from distinctively Black names raise more suspicion than those from distinctively White names,[17] and if academics respond less when suspicious, racial gaps in responses may be unrelated to actual discriminatory behaviors. Second, even if suspicion is not affected by race, it could be the case that (i) Vocal academics are more likely to be suspicious (perhaps because they are more familiar with audit studies, being more interested in racial justice), and (ii) when suspicious, social signaling concerns drive these academics to reply more to Black names than White names. Either of these two cases could explain our findings, even if Silent and Vocal academics are equally likely to discriminate against Black students in daily interactions. Given this concern, we report a series of checks.

First, our core findings are very similar when we use samples and outcomes less subject to detection concerns (Figure A9). In particular, the patterns of discrimination of the Silent and Vocal are similar when we drop academics in fields more familiar with audit studies: either those in Economics, Political Science, Sociology, and Business (12.4% of the sample), or all of those in the Social Sciences (25% of the sample). The results are also similar if we drop the 7% of academics to whom we sent more generic emails – not mentioning the specific field of the academic, which could arouse more suspicion. Next, assuming that suspicion is more likely when academics see that colleagues have received similar emails, we show that the results are similar when we drop university-departments that received either more than

[17] Given that Black students are underrepresented in graduate studies, a Bayesian should think that emails from Black students are more likely to be fictitious than emails from White students.

ten, or more than five, emails. Finally, based on the same idea that discussion between academics might increase suspicion, we show that the findings are similar when we define the outcome as accepting the meeting within one day (likely without having the time to discuss the email with other academics), or within three days.

Second, we use an accounting exercise to ask: assuming that there is no true difference in discrimination between Silent and Vocal academics, what percentage of academics would need to be suspicious to fully explain a difference in racial discrimination of 7.3 percentage points? For this exercise, we make the following assumptions: (i) Silent academics never detect the audit (a conservative assumption if detection makes academics weakly more likely to avoid anti-Black discrimination), (ii) some percentage X of the Vocal detect the audit, while the remaining (100-X)% act as they would in real life, (iii) true racial discrimination is the same for Vocal and Silent academics, at 5.3 percentage points, and (iv) Vocal detectors accept meetings only when the student is Black, with 30% overall acceptance (a particularly extreme assumption, giving the Vocal detectors a 60 percentage point discrimination rate). Even under these conservative assumptions, we would need at least 11.2% of the Vocal academics to have detected the audit to fully explain our core unconditional signaling result.[18] We find this number relatively unlikely, especially for fields outside of the Social Sciences.

Our checks give some evidence against detection driving our results. However, since the true prevalence and patterns of detection are unobserved, we cannot fully rule out the possibility of detection affecting our results. We turn then to a broader set of outcomes that are not subject to the detection concern.

3.2 High Stakes Behaviors

The audit experiment delivers clean identification of discrimination – allowing us to answer the first-order question of whether racial justice tweets are predictive of racial discrimination. But how generalizable is this informativeness? Are racial justice tweets also informative of higher stakes academic behaviors? To explore this, we collect data for our 18,514 tweeting academics on four additional behaviors: (i) average teacher ratings scraped from `ratemyprofessors.com` (1 to 5), (ii) a dummy variable for whether the academic has worked on a racial justice-related topic since 2000, using data on academic works from `openalex.org`, (iii) the percentage of co-authors with Black-sounding names, using OpenAlex data on co-authors and GPT-4o for guessing the race of each name, and (iv) a dummy variable for whether the academic has left Twitter, using scraped data from `nitter.net` (for full data details, see Appendix G). These additional variables capture multiple high-stakes domains of an academic's career: their teaching, choice of topics, and choice of co-authors. The outcome of leaving Twitter is more open to interpretation, though plausibly it reflects a response to the Elon Musk takeover and resultant changes to the platform. Consistent with this, academics began to leave Twitter when Elon took over (Figure A10), and

[18] Note that this is a much stronger statement than requiring 11.2% of the Vocal academics to be familiar with the audit experiment methodology. We are requiring them to recognize that our email was fictitious. This reaction is plausible for those for whom the audit method is top-of-mind, but less so for those who are familiar with the method but work in a field that does not use audit studies.

the probability of leaving Twitter is much higher in left-leaning departments (Figure A11).

Compared with Silent academics, Vocal academics have 0.11 SD higher teaching ratings, they are more than twice as likely to work on racial justice-related topics, their share of Black co-authors is 41% higher, and they are 60% more likely to have left Twitter by March 2025 (all $p < 0.01$, panel (a), Table 1).[19] These differences attenuate (by 24 to 67%), though remain statistically significant, after adding our plausible student information set of controls – the number of tweets, the academic's gender, position and race/ethnicity, and university and department fixed effects – and some outcome-specific controls, like dummy variables for deciles of the number of ratings in the case of teacher ratings (panel (b)). In the case of teacher ratings, robustness to controls also speaks a little against the interpretation that the Vocal-Silent gap in teacher ratings is due to bias in the ratings, as opposed to objective differences in the quality of teaching, since our controls include the main dimensions along which bias has been documented, like gender (Mengel et al. 2019). Otherwise, the results are robust to alternative outcome definitions, including a dummy variable for any Black co-author, rather than the percentage (Table A4).

3.3 Tweeting and Racial Justice Donations

Our results so far establish that racial justice tweets are highly predictive of discriminatory and broader academic behaviors, even once we condition on other factors observable to students. We replicate this result in the donation study, which bolsters external validity in two ways. First, we show that our informativeness results extend to non-academics. Second, the donation outcome is more private than an academic's email response, and much more private than our four high stakes behaviors – and thus a closer proxy to the private types of individuals emphasized by theory.

We replicate the unconditional informativeness result in Figure 7. Participants that did not retweet the NAACP tweet (87% of participants) donated \$1.46, or 29%, of their survey incentive to the NAACP. Participants that retweeted the NAACP tweet gave \$2.24, 53% more than non-retweeting participants.

One concern with this result is that the retweeting behavior is not naturalistic – it follows an explicit nudge in an online survey, which itself follows a question asking for a donation decision. The selection into retweeting the NAACP, based on private types, may differ from the selection into tweeting about racial justice outside of our study. Related, it may be that experimenter demand effects influence both the retweeting and the donation in the same direction, generating a correlation between the two. To explore this, we use our manually coded measure of racial justice-tweeting behavior during March to October 2020. We find the same result. Participants that were Vocal in 2020 donate 53% more to the NAACP several years later than participants that were Silent in 2020 (right-hand panel, Figure 7).

As with the audit-measured behaviors, tweets are also conditionally predictive: racial justice tweeters donate more to the NAACP even after conditioning on gender, income, politics, race/ethnicity, and age (Figure A12).

[19] Recall too that racial justice tweets are highly predictive of political contributions (Figure 1), which could be thought of as an additional higher-stakes measure of behavior.

3.4 When Are Tweets More Informative?

We have rejected the possibility of an uninformative equilibrium – racial justice tweets predict racial discrimination, high stakes academic behaviors, and charitable donations, both unconditionally and conditionally. Next, we consider a key comparative static from theory: that informativeness is lower when the returns to signaling support for racial justice are higher. We derive this comparative static in the simplest possible model in Appendix B, though the same prediction holds in more general models, like Frankel and Kartik (2019).

In our simple model, individuals have either low or high support for racial justice-related efforts. Both low and high types have social image concerns, receiving utility from signaling that they are a high type, while only low types face a cost of tweeting in support of racial justice. Equilibrium informativeness depends crucially on the stakes of the signaling concerns. When stakes are low (for example when tweet visibility is low, or when tweets are not used by audiences for decision-making), there is full information revelation: only high types tweet in support of racial justice. When stakes are high, an uninformative pooling equilibrium prevails, with both types tweeting in support of racial justice. With intermediate stakes, we have a partial-pooling equilibrium, with some low types tweeting in support of racial justice. In this region, informativeness is strictly decreasing in signaling concerns.

We test this comparative static in three ways, with a focus on our most private outcomes – audit-measured behavior and NAACP donations – to hue closely to the theoretical object of private types. First, we compare tweet informativeness immediately after the murder of George Floyd (a period of high signaling returns to tweeting), with informativeness some months later. Second, we compare the informativeness of the most visible type of tweets (original tweets), to the informativeness of other types of tweets (replies, retweets, quote retweets). Third, using the donation study, we compare the informativeness of NAACP retweets with that of private statements of support. While each of the three sources of variation has weaknesses, they tell a consistent story: higher signaling stakes are associated with lower informativeness.

George Floyd. Racial justice tweets among non-Black academics were rare in early-2020, but spiked following the murder of George Floyd by a White police officer on May 25th (Figure A13). After a month or so, racial justice tweets became less common again, though more common than in early-2020. We split 2020 into three periods: the pre-period (January 1 to May 24), the during-period (May 25 to June 30), and the post-period (July 1 to November 30). We think of the during-period as a period of high social pressure to tweet about racial justice, or more generally, as a period of higher incentives to signal low racial bias. For example, this was a period in which academics would sometimes be publicly called out for remaining silent on the topic, and in which racial justice tweets would receive more praise (e.g. in terms of likes and retweets).

Our model predicts that informativeness should be weakly higher (and strictly higher using the more general model of Frankel and Kartik (2019)) in the post-period than in the during-period, given that the incentives to signal are weaker. The prediction for informativeness in the during-period relative to the

pre-period is more ambiguous: while signaling incentives are surely higher in the during-period than the pre-period, the distribution of underlying racial bias of academics is also likely to have changed, with academics getting exposure to many more cases of racial injustice (Reny and Newman 2021). Thus we focus on the during- versus post-period comparison.

With our focus now on testing for *differences* in informativeness, our primary tests use a continuous measure of racial justice signaling – the winsorized percentage of a user's tweets that are about racial justice – rather than the binary measure Vocal_i.[20] We do this for the reason discussed in footnote 11 – an issue with the binary measure is that we could find that empirical informativeness changes even in the absence of information loss as defined in the model. The continuous measure is less subject to this critique, though it is not immune, since it is censored at zero. In any case, we also report similar findings using the binary measure.

During the pre-period, 17% of audited academics tweeted at least once about racial justice. The low percentage makes this period's estimates the most imprecise, though there is evidence for unconditional and conditional informativeness – a one percentage point increase in racial justice tweets during this period is associated with roughly three percentage points less anti-Black discrimination in the subsequent audit experiment ($p < 0.02$, Figure 8). The point estimates for the during-period, when 49% of academics are Vocal, are much smaller, though more precisely estimated – an unconditional informativeness of 0.17 percentage points ($p = 0.02$), and a conditional difference in discrimination of 0.11 percentage points ($p = 0.16$). Hence, during the height of racial justice tweeting, we cannot reject the null hypothesis that racial justice tweets are uninformative when basic observable characteristics are known.

Consistent with the model's partial pooling equilibrium, informativeness is substantially higher in the post-period – unconditional informativeness is 1.2 percentage points ($p < 0.01$) and conditional informativeness is 0.94 percentage points ($p = 0.01$). We can reject that during- versus post-period informativeness is equal ($p < 0.01$ for unconditional, $p = 0.02$ for conditional). When using Twitter, rational belief updating about racial discrimination should take into account the current popularity of racial justice tweets. A Bayesian updates much more from seeing one racial justice tweet when such tweets are unpopular than when they are popular.

The key pattern of during- versus post-period informativeness is similar with the binary measure (Figure A14). Unconditional informativeness in the post-period is 57% higher than in the during-period, while conditional informativeness is roughly twice as high, though we have less power to reject the null that during- and post-period informativeness are equal ($p = 0.13$ and $p = 0.19$). The conclusion is also unchanged when we horse-race during-period vocality against post-period vocality in the same regression (Table 2). Whether considering unconditional informativeness (columns 1 and 3) or conditional informativeness (columns 2 and 4), racial justice tweets immediately after the murder of George Floyd are uninformative of discriminatory behavior, while racial justice tweets months later are highly informative.

To the extent that the increase in informativeness is due to the fall in social pressure and scrutiny, the results here highlight a tradeoff: while scrutiny increases attempts to transmit information, through

[20] We winsorize continuous measures at the 99th percentile to reduce the influence of outlying observations.

prompting more academics to tweet and retweet, scrutiny also reduces the informational benefits of social media.

Tweet Type. Some types of tweets give more credit and visibility to the agent that tweets. Original tweets are composed fully by the agent (high-credit), whereas retweets involve the agent amplifying an original tweet composed by someone else (lower credit, and lower visibility given that the agent's name appears less prominently to others in the Twitter feed). Quote retweets are similar, although they include extra text added by the agent, and tweet replies are similar in that they are less visible to others, as they are less likely to appear on the Twitter feed. Given this, we compare the informativeness of original tweets (high-credit) with non-original tweets (low-credit) in Table 3, decomposing the total percentage of racial justice tweets into its additive components by tweet type.[21] In addition, to ensure that informativeness is driven by the numerator (the number of racial justice tweets), rather than the denominator (the number of total tweets), we include the full set of Twitter activity levels and interactions throughout.

Columns 1 and 2 replicate our earlier result with the continuous measure. For each additional 1% of racial justice tweets of any type, academics discriminate against Black students 1.62 percentage points less (column 1). This difference falls to 1.33 percentage points after adding the preferred set of controls and interactions – gender, race/ethnicity, position, university, and department (column 2) – as well as our measure of the vocality of the academic's network. Consistent with the model, this informativeness is fully driven by the non-original tweets, with original tweets being completely uninformative (columns 3 and 4), although we lack power to reject equality of the two coefficients ($p = 0.18$ and $p = 0.26$).

Breaking up the racial justice tweets into their four components, we see that original tweets are the least predictive (columns 5 and 6), and we can almost reject the null hypothesis that original tweets are as informative as retweets ($p = 0.1$ and $p = 0.16$). Quote tweets and replies are the most informative, although the least precisely estimated. Importantly, since these results hold conditional on our measure of the vocality of the academic's network (and its interaction with Black), the informativeness of retweets is not just reflecting that (i) academics with more vocal networks see (and thus retweet) more racial justice tweets, and (ii) academics that choose to follow more vocal accounts are those that favor Black students.

The results in columns 5 and 6 suggest that credit is the distinguishing feature of types of tweets – tweets composed fully by the agent are the least informative. Informativeness instead comes only from the tweets that amplify the voices of others (retweets and quote retweets) or contribute to conversations started by others (tweet replies).[22]

[21] The continuous measure of racial justice signaling has an additional advantage here. In particular, a comparison of the informativeness of a dummy variable for any racial justice retweet with a dummy variable for any racial justice non-retweet is not comparing like-for-like. For example, if the mean number of racial justice retweets among those with at least one was 20, and the equivalent mean for racial justice non-retweets was 10, we would effectively be comparing the informativeness of 20 racial justice retweets with that of 10 racial justice non-retweets.

[22] While we focus on differences in visibility between retweets and original tweets ($v_{original} > v_{retweet}$ in our simple model in Appendix B), we might also think that there are differences in costs. For example, suppose that the high racial justice support type has $c = 0$ for both retweets and original tweets, but the low-support type has $c_{retweet} > 0$ for retweets, and $c_{original} > c_{retweet}$ for original tweets – perhaps because original tweets require an additional cost of thinking through the appropriate language to use when talking about racial justice; language that is more familiar to high-support types. In the partial pooling equilibrium of our model, the informativeness of retweets is greater than that of original tweets only if $\frac{v_{original}}{v_{retweet}} > \frac{c_{original}}{c_{retweet}}$. Our findings in

Private Statements of Support vs. Retweets. We have shown that two types of visibility on Twitter are associated with information loss: the visibility of the tweet type, and the visibility of tweets in general. These results motivate a third question: in the best-case-scenario of low-scrutiny and low-credit retweets, is there still information loss relative to a counterfactual of private statements of support for racial justice? For this, we turn to our donation study – a study explicitly designed to give comparable binary measures of support and retweet behavior, enabling a horse race between the informativeness of retweets versus private statements.[23]

Both the binary retweet and support variables are highly unconditionally informative of NAACP donations (panel (a), Figure 9). Retweeters give \$0.78 more than non-retweeters, and supporters give \$1.01 more than non-supporters. Retweets are then 23% less informative than private support, although we do not have the power to reject equality at conventional levels ($p = 0.19$). Once we condition on other observables, informativeness of both retweets and support drops somewhat, with retweets remaining 25% less informative ($p = 0.3$).

Conceptually, information loss could be driven by low types pooling with high types – tweeting about racial justice in the absence of true support – or by "shy" high types pooling with low types – by deciding not to tweet about racial justice despite being supportive.[24] Such behavior could be motivated by a desire to take altruistic action without seeking any public acclaim. To test this, we compare the NAACP retweeting and private statements of support of the 39% of participants that gave nothing to the NAACP with the 15% of participants that gave all \$5 to the NAACP (panel (b), Figure 9). The results show suggestive support for the model: low types are 36% more likely to retweet the NAACP retweet than to privately report support above the calibrated threshold ($p = 0.11$), whereas high types are similarly likely to retweet and support ($p = 0.66$ for the difference).

Discussion. Consistent with theory, we find three pieces of evidence of signaling stakes reducing informativeness: racial justice tweets in the high-scrutiny aftermath of George Floyd's murder become nearly-uninformative, high-credit original tweets are less informative than low-credit retweets, and NAACP retweets are somewhat less informative than private statements of support. While each of the three tests is imperfect, collectively they suggest that (i) tweets span nearly the entire spectrum of informativeness, and (ii) visibility and scrutiny of tweets drives information loss.

3.5 What Do Audiences Believe About Informativeness?

Having characterized informativeness, we finish the paper by characterizing audience *beliefs* about informativeness. Such beliefs matter in theory: if audiences perceive racial justice tweets to be uninformative

Table 3 can then be taken as evidence that the relative visibility of original tweets to retweets is larger than the relative cost.

[23] Our exercise here is inspired by Braghieri (2024), who finds that public statements on politically sensitive topics are less informative than private statements. Here we test for the effect of the *bundle* of attributes that comprise social media, with one attribute being signaling concerns, and others including social influence and awareness-raising. I.e. it could be the case that signaling concerns alone reduces informativeness, but that signaling concerns bundled with influence has no net effect on informativeness.

[24] Our model in Appendix B predicts the former – low types mimicking high types – but not the latter.

virtue signaling, the returns to signaling are in turn lower, and counterintuitively, the equilibrium becomes *more* informative. While we cannot test this comparative static directly, we use a survey of graduate students to characterize beliefs.

Using publicly available email addresses, we sent a survey in late-2022 to 10,839 doctoral students at top-80 US universities (see Appendix H for details). 1,752 students (16.2%) completed the survey (515 Black and 1,237 non-Black). In the survey, we described our audit experiment in detail, and asked each student to predict the meeting acceptance rate for distinctively Black names, separately for the full sample, the Vocal academics, and the Silent academics (see Appendix I for survey wording). In each of the three cases, we gave the true acceptance rate for White students as a benchmark. To incentivize predictions, we randomly assigned half of the students to receive one additional lottery ticket for one of four $250 cash prizes for each accurate guess. We report summary statistics of the graduate student sample in Table A5. Similar to our academics, the students skew left-wing – only 4% of Black students and 4% of non-Black students self-identify as conservative or very conservative.

Predicting Discrimination. Most students overestimate how much academics discriminate against Black students: 84% predict anti-Black discrimination to be above the upper bound of our estimate's 95% confidence interval, 5% predict it to be within our confidence interval, and 11% predict it to be below its lower bound, meaning that they predict that Black students will be favored by at least 0.77 percentage points (Figure 10). We will call these three types of people overestimators, accurate, and underestimators, from now.

When guessing separately for Silent and Vocal academics, on average students correctly anticipate that Vocal academics discriminate less against Black students than Silent academics, but in both cases, they again tend to overestimate anti-Black discrimination. For Silent academics, 72% overestimate discrimination, while for Vocal academics, 74% overestimate. Though we find that Vocal academics discriminate in favor of Black students, 75% of students predict that Vocal academics discriminate in favor of White students.

Non-Black students predict less anti-Black discrimination than Black students, and center- or right-leaning students predict less than left-leaning students. Even so, the median non-Black and the median right-leaning student overestimates anti-Black discrimination.

Predicting Informativeness. Students make more accurate predictions about informativeness: 29% make a guess for unconditional informativeness within our confidence interval. Among the remaining guesses, students are over twice as likely to underestimate than overestimate informativeness. This finding suggests that the fundamental attribution error (Jones and Harris 1967; Andre 2021) – the overattribution of behavior (e.g. tweets) to personality traits (e.g. racial bias) rather than situational factors (e.g. signaling incentives) – is not the key source of biased beliefs in our context. Instead, taking theory at face value, the high fraction of skeptical receivers can help rationalize why the equilibrium is informative overall.

Unlike predictions about discrimination, predictions about informativeness are similar by race. In

addition, despite our prior that "virtue signaller" is a pejorative used more by the political right, liberals and moderates make similar predictions about informativeness.

4 Conclusion and Future Directions

This paper characterizes the informativeness of tweets about racial justice. Despite the cheap talk and large audience features of social media, we rule out the possibility of an uninformative equilibrium. Academics that tweet about racial justice behave substantially differently from those that do not – they discriminate less against Black students, and more in favor of both female and first-generation students; they receive higher teaching ratings, work more on race-related topics, and collaborate more with Black co-authors; and they are more likely to have left Twitter in recent years. Similarly, Americans that tweet about racial justice donate 53% more to the NAACP than those that do not. Nevertheless, consistent with theory, we do find that visibility-enhancing features of social media reduce informativeness, ruling out the other extreme case of a fully-separating equilibrium. Tweets are not created equal: low-visibility retweets are more informative than high-visibility original tweets; racial justice tweets during low-pressure periods are more informative than those when racial justice support is under the spotlight. In addition, we see suggestive evidence that private statements of support remain more predictive than retweets. Collectively, these results demonstrate that the *design* of social media platforms will matter for information revelation in equilibrium. For example, the informativeness of retweets suggests that platforms can increase informativeness without needing to resort to anonymity (which would be self-defeating, since the signal of a tweet could no longer be connected with the author), by reducing the salience of the name of a post's author.

Several questions merit further study. First, our analysis has taken a social media platform as given, but future work might study the effects of changes to platform features on informativeness. For example, algorithms that reduce the reach of tweets, and thus visibility, might reduce information loss. Related, researchers might study how tweet informativeness has evolved now that academic Twitter is less active (Figure A10), or how informativeness differs on competing platforms, like Bluesky. Second, informativeness may change if informativeness is publicized, increasing the fraction of sophisticated onlookers. Theoretically, this weakly reduces informativeness, as more low racial justice-support individuals start tweeting about racial justice. Future work could explore this by randomizing information about equilibrium informativeness and measuring subsequent social media activity. Third, work could explore the source of audience misperceptions. One hypothesis is that hypocrisy is more memorable than consistency, leading people to underestimate the informativeness of moral statements.

References

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond, "Preferences for truth-telling," *Econometrica*, 2019, *87* (4), 1115–1153.

Ajzenman, Nicolás, Bruno Ferman, and Pedro C Sant'Anna, "Discrimination in the Formation of Academic Networks: A Field Experiment on# EconTwitter," 2023.

Alatas, Vivi, Arun G Chandrasekhar, Markus Mobius, Benjamin A Olken, and Cindy Paladines, "When celebrities speak: A nationwide twitter experiment promoting vaccination in Indonesia," Technical Report, National Bureau of Economic Research 2019.

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow, "The welfare effects of social media," *American Economic Review*, 2020, *110* (3), 629–76.

Andre, Peter, "Shallow meritocracy: An experiment on fairness views," Technical Report, ECONtribute Discussion Paper 2021.

Austen-Smith, David, "Information transmission in debate," *American Journal of political science*, 1990, pp. 124–152.

Bailey, Michael, Drew Johnston, Theresa Kuchler, Johannes Stroebel, and Arlene Wong, "Peer effects in product adoption," *American Economic Journal: Applied Economics*, 2022, *14* (3), 488–526.

_ , Ruiqing Cao, Theresa Kuchler, and Johannes Stroebel, "The economic effects of social networks: Evidence from the housing market," *Journal of Political Economy*, 2018, *126* (6), 2224–2276.

Becker, Gary Stanley, *The Economics of Discrimination*, University of Chicago Press, 1957.

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.

Blume, Andreas, Ernest K Lai, and Wooyoung Lim, "Strategic information transmission: A survey of experiments and theoretical foundations," *Handbook of experimental game theory*, 2020, pp. 311–347.

BOA, "Twitter Revenue and Usage Statistics 2022," Technical Report, Business of Apps, https://www.businessofapps.com/data/twitter-statistics/ 2022.

Bonheur, Amanda, "Is Slacktivism Harmless? Unintended Consequences of Social Media Activism," in "2023 APPAM Fall Research Conference" APPAM 2023.

Bouton, Laurent, Julia Cagé, Edgard Dewitte, and Vincent Pons, "Small Campaign Donors," Technical Report, National Bureau of Economic Research 2022.

Braghieri, Luca, "Political correctness, social image, and information transmission," *American Economic Review*, 2024, *114* (12), 3877–3904.

_ , Ro'ee Levy, and Alexey Makarin, "Social media and mental health," *American Economic Review*, 2022, *112* (11), 3660–93.

Bursztyn, Leonardo, Alessandra L González, and David Yanagizawa-Drott, "Misperceived social norms: Women working outside the home in saudi arabia," *American economic review*, 2020, *110* (10), 2997–3029.

_ , Georgy Egorov, and Stefano Fiorin, "From extreme to mainstream: The erosion of social norms," *American economic review*, 2020, *110* (11), 3522–48.

_ , Thomas Fujiwara, and Amanda Pallais, "'Acting Wife': Marriage Market Incentives and Labor Market Investments," *American Economic Review*, 2017, *107* (11), 3288–3319.

Cai, Hongbin and Joseph Tao-Yi Wang, "Overcommunication in strategic information transmission games," *Games and Economic Behavior*, 2006, *56* (1), 7–36.

Cantoni, Davide, David Y Yang, Noam Yuchtman, and Y Jane Zhang, "Protests as strategic games: experimental evidence from Hong Kong's antiauthoritarian movement," *The Quarterly Journal of Economics*, 2019, *134* (2), 1021–1077.

Card, David, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri, "Gender gaps at the academies," *Proceedings of the National Academy of Sciences*, 2023, *120* (4), e2212421120.

Charles, Kerwin Kofi and Jonathan Guryan, "Prejudice and wages: an empirical assessment of Becker's The Economics of Discrimination," *Journal of political economy*, 2008, *116* (5), 773–809.

Crawford, Vincent P and Joel Sobel, "Strategic information transmission," *Econometrica: Journal of the Econometric Society*, 1982, pp. 1431–1451.

Culbert, Jack H, Anne Hobert, Najko Jahn, Nick Haupka, Marion Schmidt, Paul Donner, and Philipp Mayr, "Reference coverage analysis of OpenAlex compared to Web of Science and Scopus," *Scientometrics*, 2025, *130* (4), 2475–2492.

Economist, The, "Has Twitter (now X) become more right-wing?," Technical Report, https://www.economist.com/graphic-detail/2023/12/20/has-twitter-now-x-become-more-right-wing 2023.

Frankel, Alex and Navin Kartik, "Muddled information," *Journal of Political Economy*, 2019, *127* (4), 1739–1776.

_ and _ , "Improving information from manipulable data," *Journal of the European Economic Association*, 2022, *20* (1), 79–115.

Fujiwara, Thomas, Karsten Müller, and Carlo Schwarz, "The effect of social media on elections: Evidence from the United States," *Journal of the European Economic Association*, 2023, p. jvad058.

Gaddis, S Michael, Edvard Larsen, Charles Crabtree, and John Holbein, "Discrimination against black and Hispanic Americans is highest in hiring and housing contexts: A meta-analysis of correspondence audits," *Available at SSRN 3975770*, 2021.

Gneezy, Uri, Alex Imas, and Kristóf Madarász, "Conscience accounting: Emotion dynamics and social behavior," *Management Science*, 2014, *60* (11), 2645–2658.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart, "Designing Information Provision Experiments," *Journal of Economic Literature*, forthcoming.

Jacob, Brian A, Jonah E Rockoff, Eric S Taylor, Benjamin Lindy, and Rachel Rosen, "Teacher applicant hiring and teacher performance: Evidence from DC public schools," *Journal of public economics*, 2018, *166*, 81–97.

Jones, Edward E. and Victor A. Harris, "The attribution of attitudes," *Journal of Experimental Social Psychology*, 1967, *3* (1), 1–24.

Karing, Anne, "Social Signaling and Childhood Immunization: A Field Experiment in Sierra Leone," *Working Paper, University of California, Berkeley*, 2023.

Kartik, Navin, "Strategic communication with lying costs," *The Review of Economic Studies*, 2009, *76* (4), 1359–1395.

Kaur, Harnoor and Noam Yuchtman, "Protests on campus: the political economy of universities and social movements," *Comparative Economic Studies*, 2024, pp. 1–18.

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.

Kirgios, Erika L, Aneesh Rai, Edward H Chang, and Katherine L Milkman, "When seeking help, women and racial/ethnic minorities benefit from explicitly stating their identity," *Nature Human Behaviour*, 2022, *6* (3), 383–391.

Kline, Patrick, Evan K Rose, and Christopher R Walters, "Systemic discrimination among large US employers," *The Quarterly Journal of Economics*, 2022, *137* (4), 1963–2036.

Langbert, Mitchell, Anthony J Quain, Daniel B Klein et al., "Faculty voter registration in economics, history, journalism, law, and psychology," *Econ Journal Watch*, 2016, *13* (3), 422–451.

Lerner, Josh, Henry J Manley, Carolyn Stein, and Heidi Williams, "The wandering scholars: Understanding the heterogeneity of university commercialization," Technical Report, National Bureau of Economic Research 2024.

Loury, Glenn C, "Self-censorship in public discourse: A theory of "political correctness" and related phenomena," *Rationality and Society*, 1994, *6* (4), 428–461.

Mazar, Nina and Chen-Bo Zhong, "Do green products make us better people?," *Psychological science*, 2010, *21* (4), 494–498.

McArdle, Megan, "Abandoning DEI won't fix academia's left-leaning problem," *The Washington Post*, 2025.

Mengel, Friederike, Jan Sauermann, and Ulf Zölitz, "Gender bias in teaching evaluations," *Journal of the European economic association*, 2019, *17* (2), 535–566.

Milkman, Katherine L, Modupe Akinola, and Dolly Chugh, "Temporal distance and discrimination: An audit study in academia," *Psychological science*, 2012, *23* (7), 710–717.

Morris, Stephen, "Political correctness," *Journal of political Economy*, 2001, *109* (2), 231–265.

Mosquera, Roberto, Mofioluwasademi Odunowo, Trent McNamara, Xiongfei Guo, and Ragan Petrie, "The economic effects of Facebook," *Experimental Economics*, 2020, *23* (2), 575–602.

NBC San Diego, "Trump Administration Cancels $60 Million in Harvard Grants Over Campus Anti-semitism Allegations," *NBC San Diego*, 2025.

Pager, Devah and Lincoln Quillian, "Walking the talk? What employers say versus what they do," *American Sociological Review*, 2005, *70* (3), 355–380.

PRC, "Social Media Use in 2021," Technical Report, Pew Research Center, Washington, D.C., https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ 2021.

Reny, Tyler T and Benjamin J Newman, "The opinion-mobilizing effect of social protest against police violence: Evidence from the 2020 George Floyd protests," *American Political Science Review*, 2021, *115* (4), 1499–1507.

Schwardmann, Peter, Egon Tripodi, and Joël J Van der Weele, "Self-persuasion: Evidence from field experiments at international debating competitions," *American Economic Review*, 2022, *112* (4), 1118–46.

Song, Lena, "Essays on Information Technology and Inequality," 2022.

Stansbury, Anna and Kyra Rodriguez, "The class gap in career progression: Evidence from us academia," Technical Report, mimeo MIT 2024.

_ **and Robert Schultz**, "The economics profession's socioeconomic diversity problem," *Journal of Economic Perspectives*, 2023, *37* (4), 207–230.

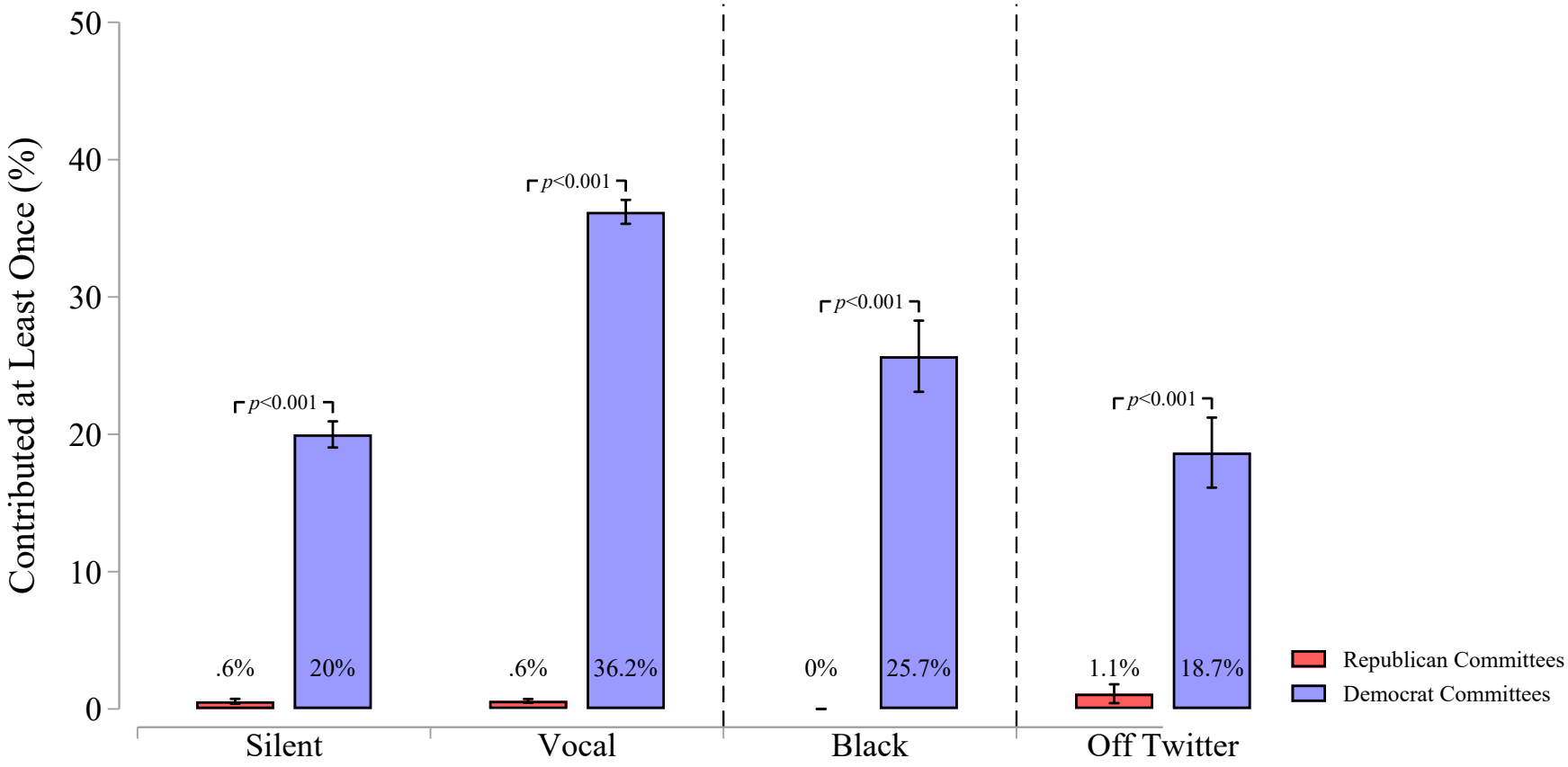
Tyler, John H, Richard J Murnane, and John B Willett, "Estimating the labor market signaling value of the GED," *The Quarterly Journal of Economics*, 2000, *115* (2), 431–468.

Yang, Maya, "Trump administration halts Harvard's ability to enroll international students," *The Guardian*, 2025. Accessed: 2025-05-25.

Yuchtman, Noam, "Universities and the Contested Creation of the Elite," *The Manchester School*, 2025.

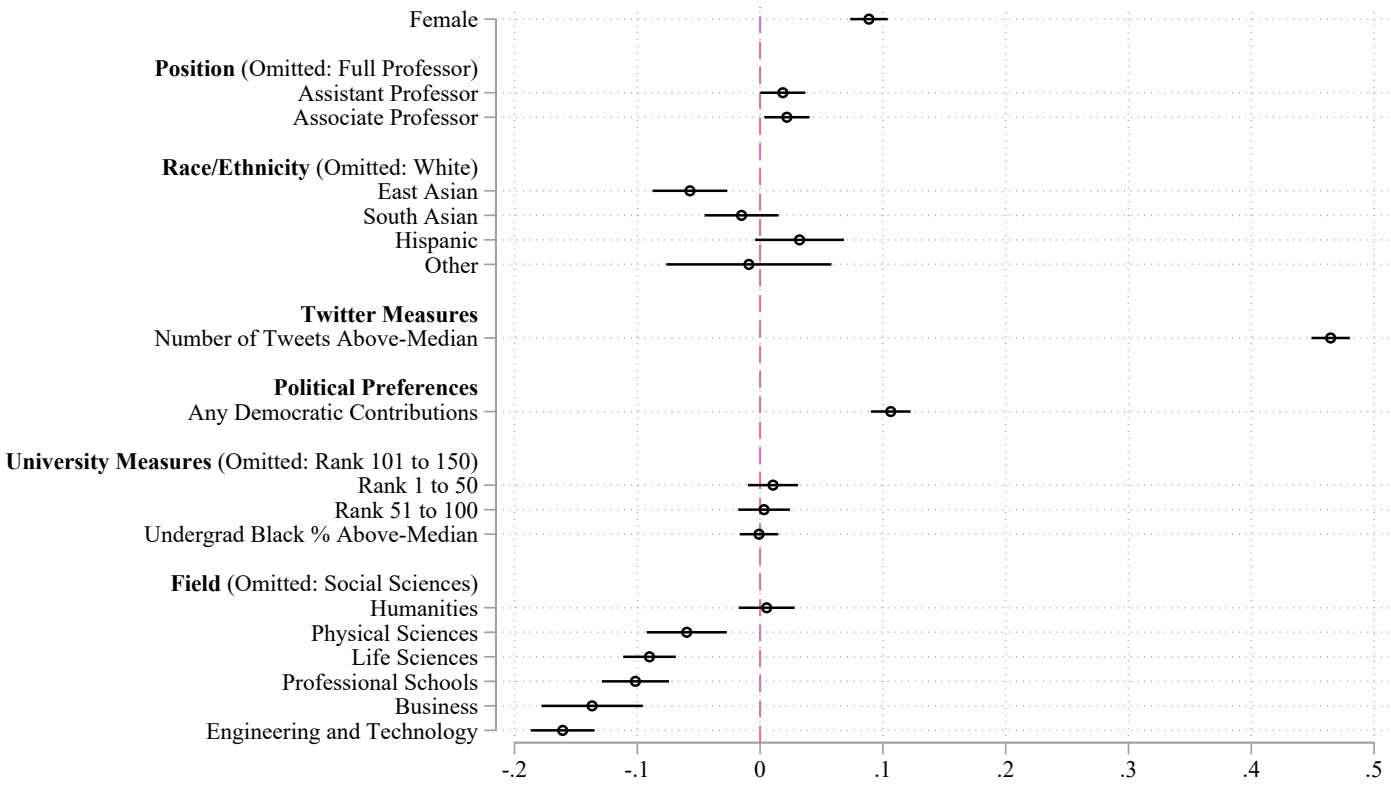
Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov, "Political effects of the internet and social media," *Annual review of economics*, 2020, *12*, 415–438.

Figure 1: Academics Almost Never Contribute to Republicans, Vocal Academics Are More Active



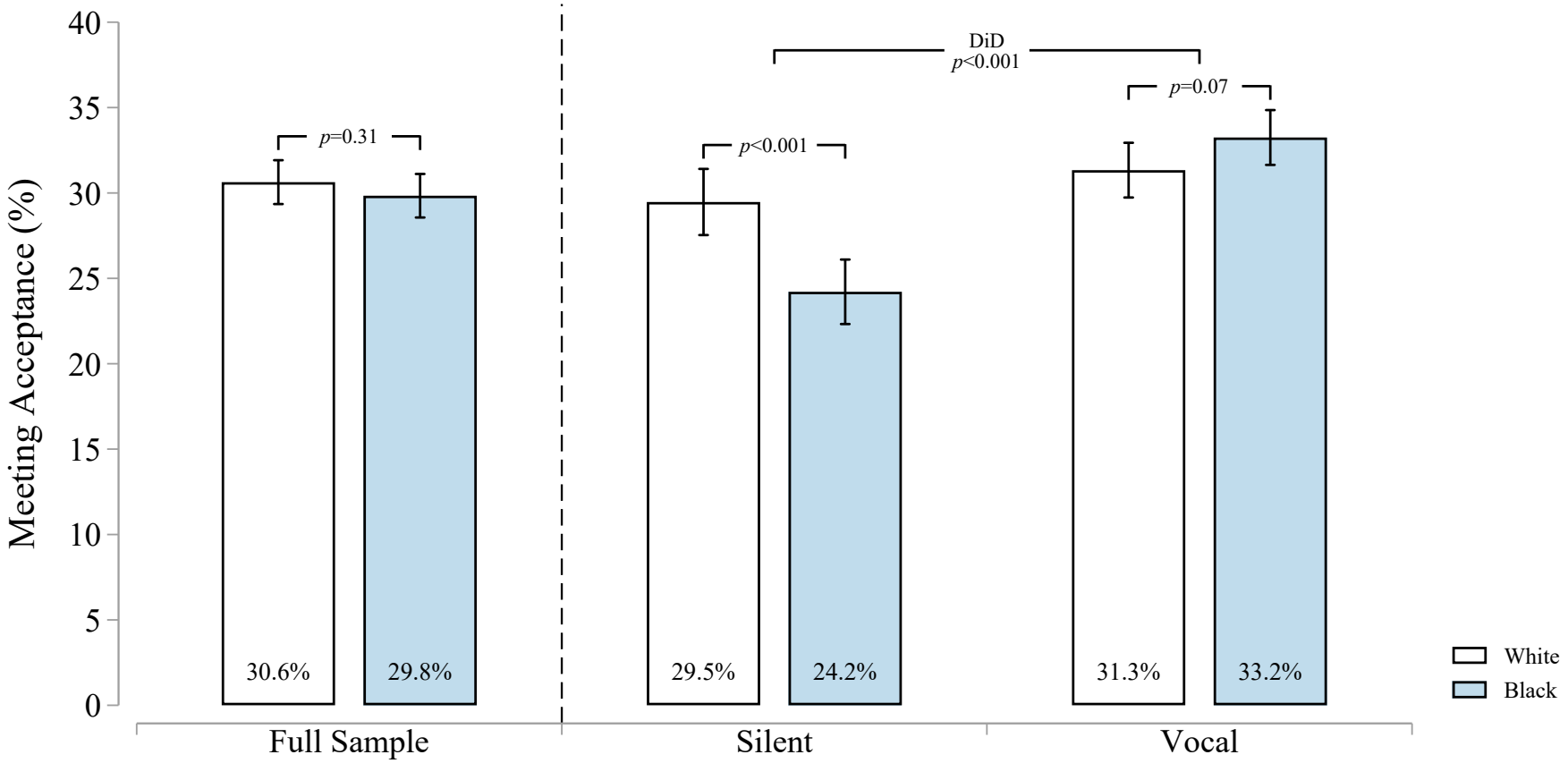
Notes: The bars show what percentage of academics made FEC-reported political contributions to Republican- and Democrat-related committees from January 2020 to March 2022. Silent includes the 6,784 non-Black academics that did not tweet about racial justice during the same time period, and Vocal includes the 11,730 non-Black academics that did tweet about racial justice. Black includes the 1,094 tweeting Black academics satisfying the same sample criteria as our audit sample. Off Twitter includes a random sample of 900 non-Black academics without Twitter accounts, but otherwise satisfying the same criteria as our audit sample. Unconditional raw means with 95% confidence intervals are shown.

Figure 2: Who Tweets About Racial Justice?



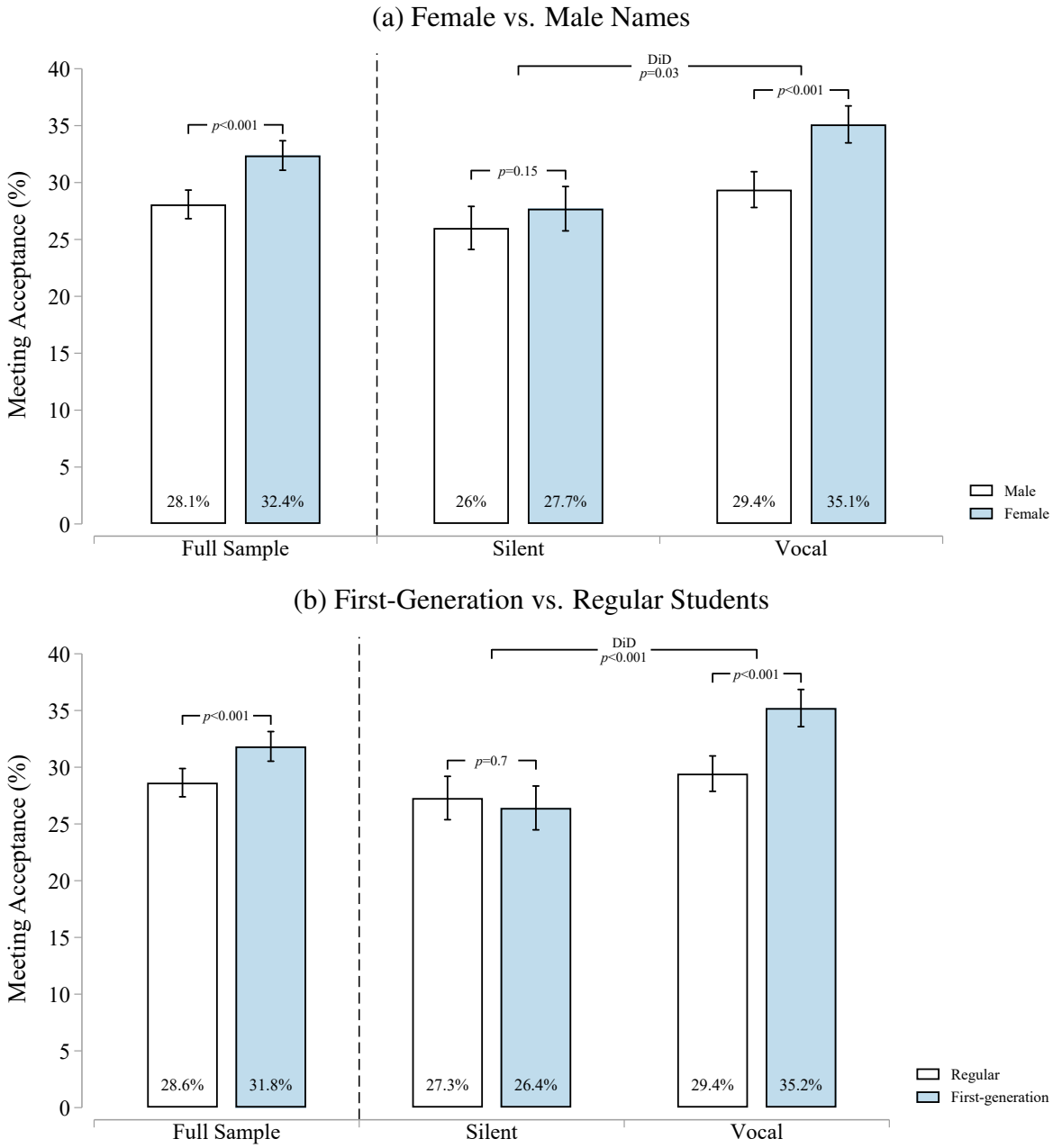
Notes: The figure displays coefficients and 95% confidence intervals from a regression of Vocal_i on a set of covariates, for the sample of 11,450 non-Black academics included in the audit experiment. The covariates are: dummy variable for female, dummy variables for Assistant and Associate Professor, race/ethnicity dummy variables, dummy variable for above-median total tweets from January 1, 2020 to March 27, 2022, dummy variable for any contributions to Democratic-related FEC committees from January 1, 2020 to March 27, 2022, dummy variables for university ranked 1 to 50 and 51 to 100, dummy variable for undergraduate Black student share above-median, and dummy variables for broad academic fields. Standard errors are HC3 robust.

Figure 3: Silent Academics Discriminate Against Black Students, Vocal Academics Discriminate (Somewhat) Against White Students



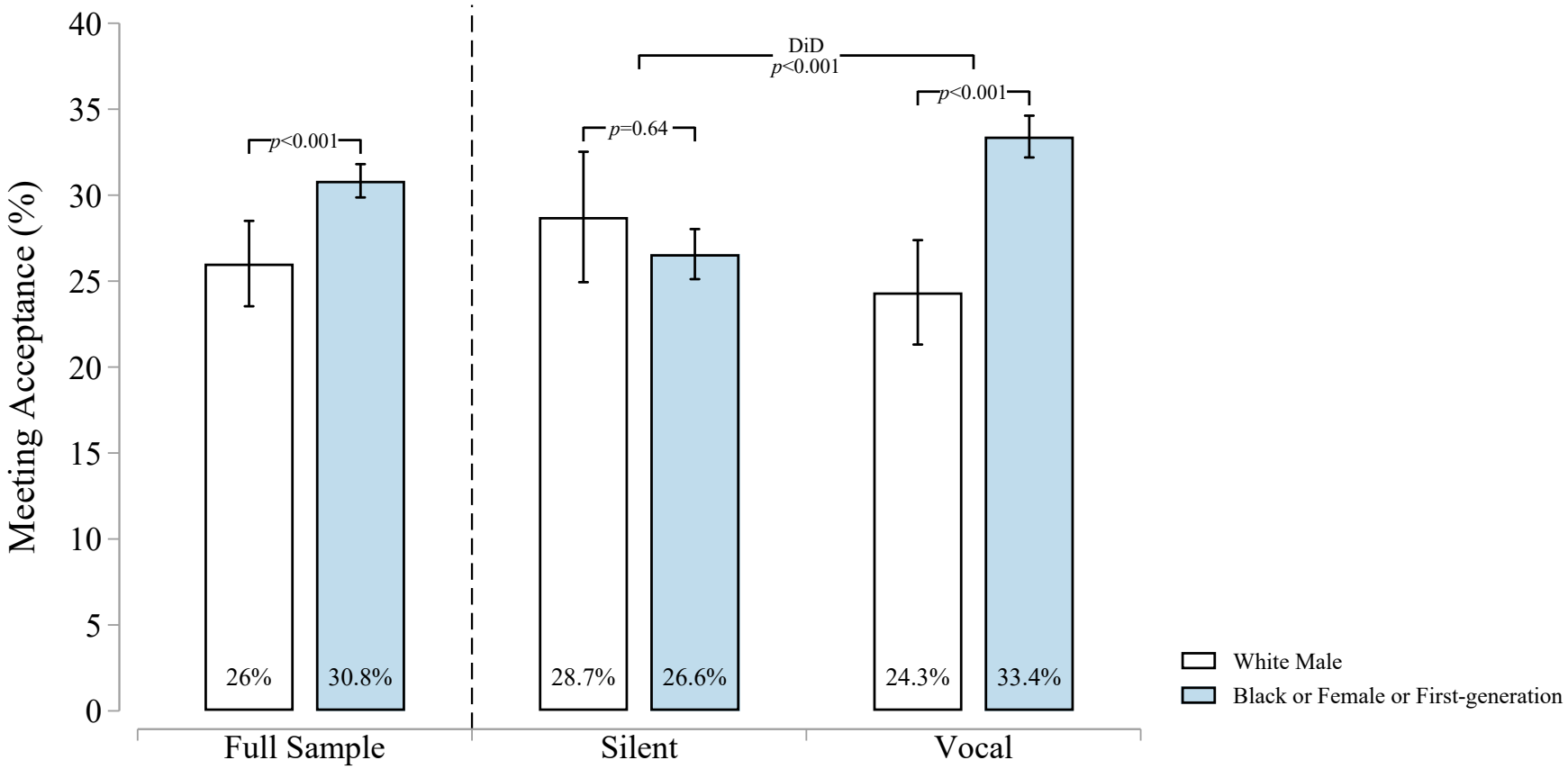
Notes: The bars show what percentage of audited academics accepted meeting requests from distinctively White and distinctively Black names. The Full Sample includes the 11,450 audited academics (4,318 Silent and 7,132 Vocal). Vocal academics are those that tweeted at least once about racial justice from January 2020 to March 2022. Silent academics are those that did not. The raw means and 95% confidence intervals come from a regression of Accepted_i on dummy variables for White and Black email sender (to the left of the vertical dashed line), and a regression on dummy variables for White email sender to Silent academic, White email sender to Vocal academic, and the same for Black email sender (to the right of the vertical dashed line). The p-values come from the specification that also includes strata and email type fixed effects. The DiD (diff-in-diff) p-value is from a test for equal discrimination rates across Vocal and Silent academics ($\hat{\gamma}_2$ in specification 2). Standard errors are clustered at the university-by-department-by-sender name-level.

Figure 4: Vocal Academics Favor Female and First-Generation Students, Silent Academics Show Little or No Bias



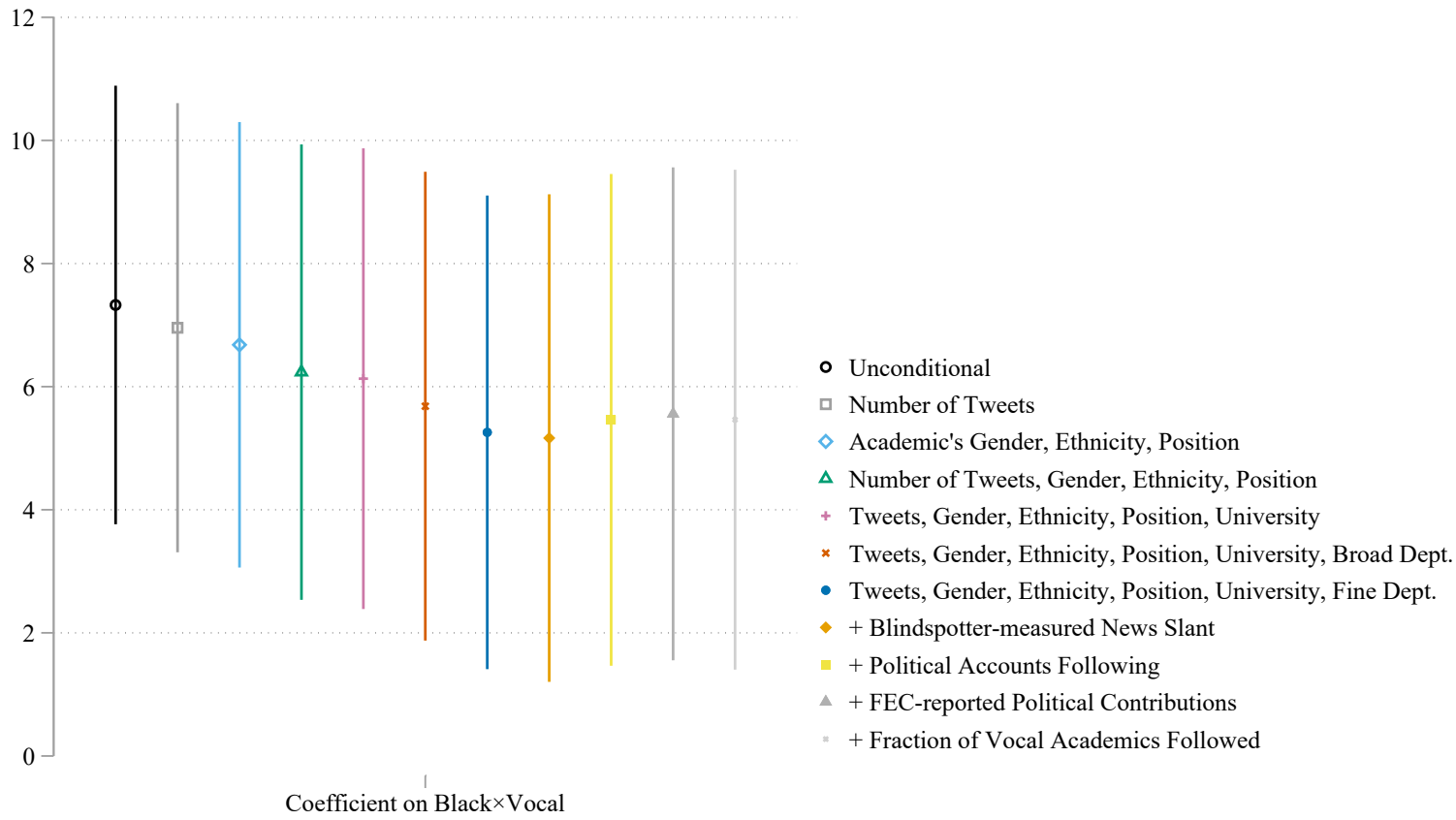
Notes: The bars show what percentage of audited academics accepted meeting requests from distinctively male and female names (panel (a)), and from students that mention their first-generation status and students that did not ("Regular" students, panel (b)). The Full Sample includes the 11,450 audited academics (4,318 Silent and 7,132 Vocal). Vocal academics are those that tweeted at least once about racial justice from January 2020 to March 2022. Silent academics are those that did not. The raw means and 95% confidence intervals to the left of the vertical dashed line come from a regression of Accepted_i on dummy variables for female versus male email sender (panel (a)), or from first-generation student and regular student email sender (panel (b)). Those to the right of the dashed line come from a regression on dummy variables for female (or first-generation) sender to Silent academic, female (or first-generation) sender to Vocal academic, and the same for male (or regular) sender. The p-values come from the specification that also includes strata and email type fixed effects. The DiD (diff-in-diff) p-value is from a test for equal discrimination rates across Vocal and Silent academics ($\hat{\gamma}_2$ in specification 2). Standard errors are clustered at the university-by-department-by-sender name-level.

Figure 5: Vocal Academics Discriminate Against White Males



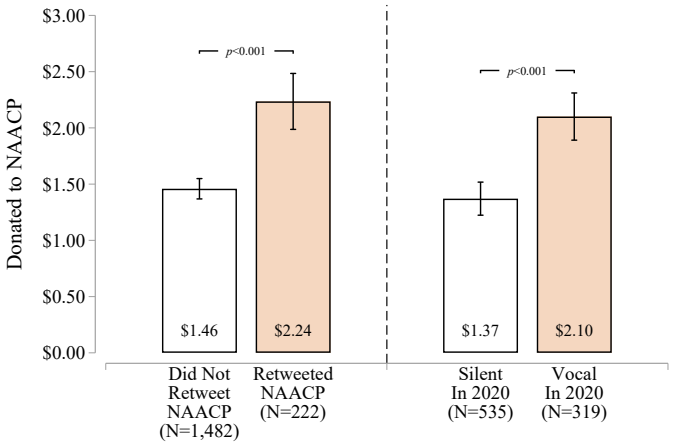
Notes: The bars show what percentage of audited academics accepted meeting requests from distinctively White male names with no mention of first-generation status (1/8 of emails) vs. emails from distinctively Black names or female names, or emails that mention first-generation status (7/8 of emails). The Full Sample includes the 11,450 audited academics (4,318 Silent and 7,132 Vocal). The raw means and 95% confidence intervals come from a regression of Accepted_i on dummy variables for White male and Black/female/first-generation email sender (to the left of the vertical dashed line), and a regression on dummy variables for White male email sender to Silent academic, White male email sender to Vocal academic, and the same for Black/female/first-generation email sender (to the right of the vertical dashed line). The p-values come from the specification that also includes strata and email type fixed effects. The DiD (diff-in-diff) p-value is from a test for equal discrimination rates across Vocal and Silent academics ($\hat{\gamma}_2$ in specification 2). Standard errors are clustered at the university-by-department-by-sender name-level.

Figure 6: Racial Justice Tweets Predict Racial Gap in Meeting Acceptance, Even After Adding Controls



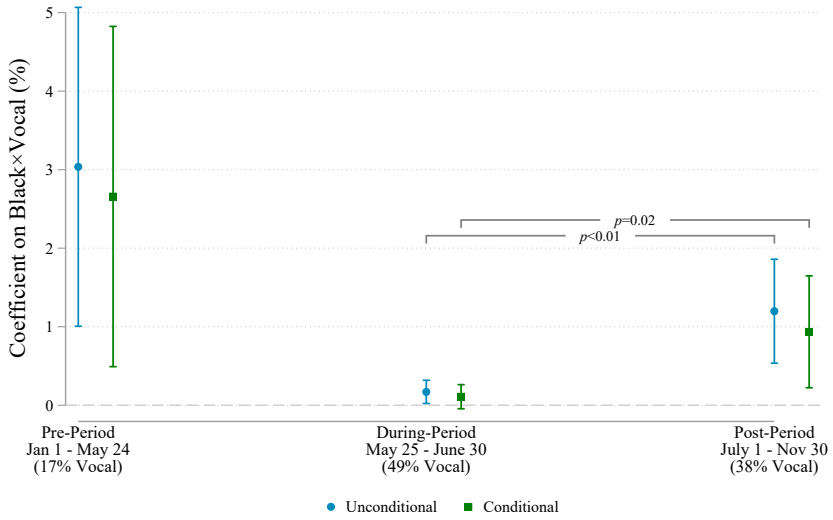
Notes: The figure shows how the difference in discrimination between Vocal and Silent academics ($\hat{\gamma}_2$ from specification 2) changes as the set of X_i^j covariates change. $\hat{\gamma}_2 > 0$ indicates that Vocal academics discriminate against Black students less (or equivalently, more in favor of) than Silent academics. The covariates are: (1) Number of Tweets: the number of original tweets, reply tweets, retweets, quote tweets, and quote reply tweets, all for the period January 1, 2020 to March 27, 2022; (2) dummy variable for female, dummy variables for Assistant Professor and Associate Professor, dummy variables for race/ethnicity; (3) includes both the tweet variables from (1) and the demographics variables from (2); (4) adds university fixed effects; (5) adds broad department dummy variables (seven departments, e.g. Social Sciences); (6) replaces broad departments with narrowly defined department dummy variables (75 departments, e.g. Economics); (7) adds Blindspotter-measured percentage left, percentage center (percentage right is omitted), dummy variable for missing because profile is private, dummy variable for missing because of insufficient content; (8) adds number of political accounts followed, percentage of political accounts followed that are Democrats, dummy variable for at least one political account followed, dummy variable for follow zero accounts, and dummy variable for missing following data (e.g. because profile is private); (9) adds dummy variable for contributed to Democrats, dummy variable for contributed to Republicans, total contributions to Democrats, and total contributions to Republicans, all for the period January 1, 2020 to March 27, 2022, and (10) adds the fraction of academics followed that are Vocal, and a dummy variable for missing following data. Standard errors are clustered at the university-by-department-by-sender name-level. 95% confidence intervals are shown.

Figure 7: Vocal Non-Academics Donate 53% More to the NAACP



Notes: The figure shows the mean NAACP donation (ranging from $0 to $5) for participants in the donation study that (i) did versus did not retweet the NAACP tweet, and (ii) tweeted in support of racial justice during May to October 2020 versus did not. The latter is shown only for the subsample for which we could manually code tweets made from May to October 2020. 95% confidence intervals are shown.

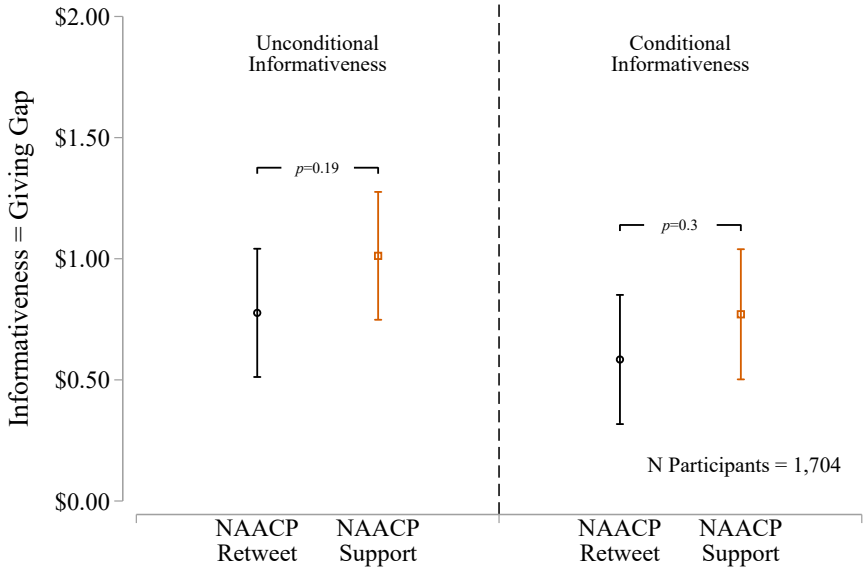
Figure 8: Tweets About Racial Justice Made When Fewer People Are Tweeting Are More Informative of Later Audit-Measured Behavior



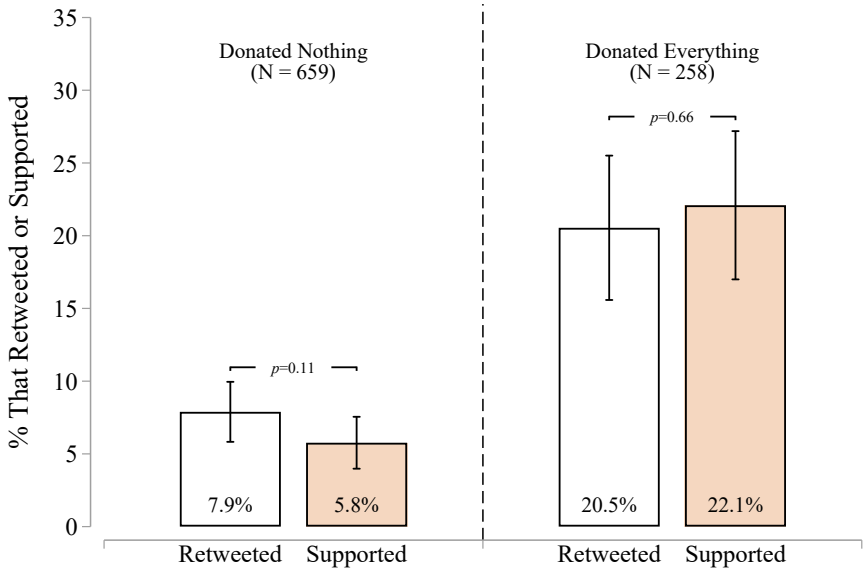
Notes: The figure shows the predictive value of racial justice tweets made during different time periods for subsequent audit-measured behavior (with the latter always measured at the same time point). In particular, the figure shows how tweet informativeness ($\hat{\gamma}_2$ from specification 2, though with Vocal $(\%)_i$ replacing Vocal_i) changes for tweets made before, during, versus after the murder of George Floyd on May 25, 2020. Vocal $(\%)_i$ is the percentage of academic i's tweets that are about racial justice, winsorized at the 99th percentile, and set to zero (rather than undefined) for academics that did not tweet at all in a given period. 17% of the 11,450 audited academics tweeted about racial justice at least once during January 1 to May 24, rising to 49% during May 25 to June 30, and falling to 38% during July 1 to November 30. The unconditional estimate denotes the reduction in subsequent audit-measured anti-Black discrimination associated with a one percentage point increase in the percentage of racial justice tweets (Vocal $(\%)_i$). The conditional estimate denotes the conditional difference in discrimination, using the fifth-from-the-right specification from Figure 6. Standard errors are clustered at the university-by-department-by-sender name-level. 95% confidence intervals are shown.

Figure 9: NAACP Retweets Are Somewhat Less Informative Than Private Statements of Support

(a) Informativeness of Tweets Versus Private Support

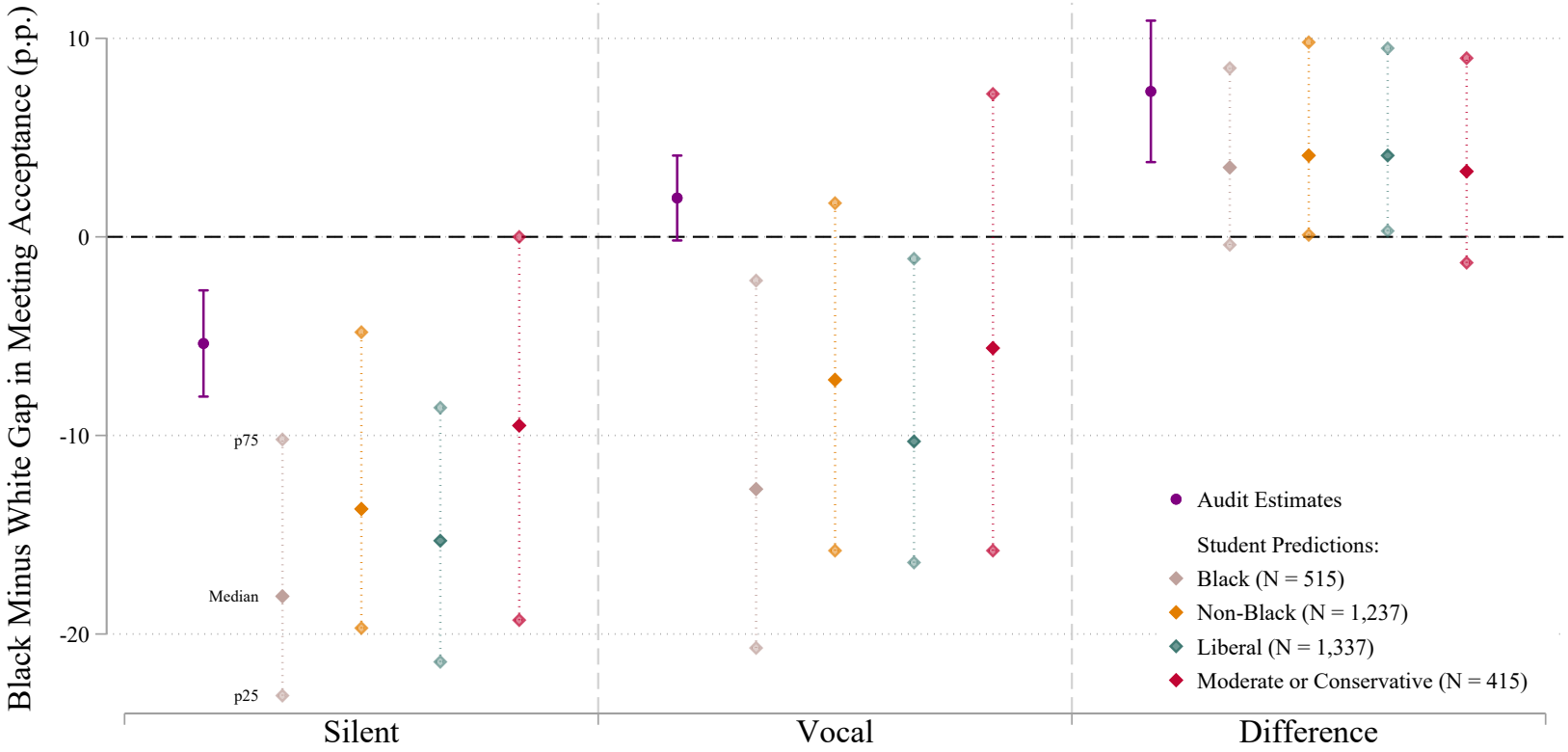


(b) Source of Information Loss



Notes: Panel (a) shows how the informativeness of NAACP retweets in the donation study ($\hat{\phi}$ from specification 3) compares with the informativeness of the privately-reported binary measure of support for the NAACP. As an example, the left-hand-side coefficient tells us that participants that retweeted the NAACP retweet donated roughly $0.78 more of their survey incentive to the NAACP than participants that did not retweet. The two right-hand-side coefficients consider informativeness conditional on the following controls: number of posts per month since joining Twitter, dummy variable for female, dummy variables for income categories, dummy variables for political views, dummy variables for non-White and Hispanic, and age. Panel (b) explores the source of the information loss, whether low-donation types (those that gave nothing to the NAACP) retweeting more than supporting, or high-donation types (those that gave the full $5 to the NAACP) retweeting less than supporting. 95% confidence intervals are shown for both panels, based on robust standard errors.

Figure 10: Students Overestimate Discrimination and Tend to Underestimate Informativeness



Notes: The figure shows our audit experiment estimates and 95% confidence intervals in purple. Difference is the unconditional difference in racial discrimination between Vocal and Silent academics, which is positive if Vocal academics are more pro-Black than Silent academics. The diamonds denote the 25th, 50th, and 75th percentile of student predictions, separately by (i) students that self-identify as Black or African American versus students that do not, and (ii) students that describe their political views as liberal or very liberal versus those that describe their political views as moderate, conservative, or very conservative. Before making predictions, students were informed of the meeting acceptance rate for distinctively White student names, separately for Silent and Vocal academics.

Table 1: Racial Justice Tweets Predict High Stakes Behaviors

	Teacher Rating (1)	Work on Racial Topic (2)	Black Coauthor (%) (3)	Left Twitter (4)
		(a) Without Controls		
Vocal	0.11***	0.12***	0.84***	0.09***
	(0.02)	(0.01)	(0.08)	(0.01)
		(b) With Preferred Controls		
Vocal	0.05**	0.04***	0.60***	0.06***
	(0.02)	(0.01)	(0.09)	(0.01)
Observations	9,197	16,753	15,801	18,514
Silent Mean	3.72	0.08	2.05	0.15
Silent SD	1.02	0.28	4.15	0.36

Note: Sample is on-Twitter non-Black academics. Outcomes: (1) Average answer to question "Rate your professor" on a scale from 1-5, (2) dummy variable indicating whether the academic has any work on a race-related topic since 2000, (3) percentage of the academic's coauthors that are Black, based on their name and for works between 2020 and 2022, and (4) dummy variable indicating that the academic left Twitter, based on their account not being found in March 2025. Vocal is a dummy variable indicating whether the academic tweeted at least once about racial justice from January 2020 to March 2022. Silent Mean and Silent SD are the mean and standard deviation of the outcome among those with Vocal = 0. The controls included in panel (b) follow those in the fifth-from-the-right coefficient in Figure 6, and are: number of each type of tweet, dummy variable for female, dummy variables for Assistant Professor and Associate Professor, dummy variables for race/ethnicity, university fixed effects, and narrowly defined department dummy variables. Column 1 of panel (b) also includes dummy variables for decile of the number of ratings in Rate My Professor. Column 2 in panel (b) also includes dummy variables for decile of the number of research topics, dummy variables for decile of works count, dummy variables for decile of name commonness, and year of first work fixed effects. Column 3 in panel (b) also includes dummy variables for decile of the number of coauthors' names coded and a dummy variable indicating whether the academic had no works or no coauthors between 2020 and 2022. Standard errors are robust.

Table 2: Racial Justice Tweets Are More Informative When Fewer People Are Tweeting About Racial Justice

	Meeting Accepted (%)			
	(1)	(2)	(3)	(4)
Black × Vocal (%) During George Floyd	0.10	0.06		
	(0.08)	(0.08)		
Black × Vocal (%) After George Floyd	1.08***	0.89**		
	(0.36)	(0.38)		
Black × Vocal (Binary) During George Floyd			1.53	0.71
			(2.10)	(2.15)
Black × Vocal (Binary) After George Floyd			6.87***	5.33**
			(2.17)	(2.32)
Observations	11,393	11,393	11,393	11,393
Strata Fixed Effects	Yes	Yes	Yes	Yes
Email Type Fixed Effects	Yes	Yes	Yes	Yes
Interacted Controls	No	Yes	No	Yes
p-value (Black × During = Black × After)	.012	.044	.15	.22

Notes: All regressions also include level variables for the interaction terms. Meeting Accepted is equal to 100 if the academic accepted the meeting request, and zero otherwise. Black is equal to one if the academic received an email from a distinctively Black-sounding name, and zero otherwise. Vocal (%) is the percentage of an academic's tweets that are about racial justice, winsorized at the 99th percentile, and set to zero (rather than undefined) for academics that did not tweet at all in a given period. 49% of academics tweeted about racial justice immediately after the murder of George Floyd on May 25, 2020 (during May 25 to June 30, which we call "During George Floyd"), while 38% of academics tweeted about racial justice later, during July 1 to November 30 ("After George Floyd"). Even columns include the variables used in the fifth-from-the-right coefficient in Figure 6, fully interacted with Black. Standard errors are clustered at the university-by-department-by-sender name-level. *** p<0.01, ** p<0.05, * p<0.1.

Table 3: High-Credit Original Tweets Are Less Informative Than Low-Credit Retweets

	Meeting Accepted (%)					
	(1)	(2)	(3)	(4)	(5)	(6)
Black × Racial Justice Tweets (%)	1.62***	1.33***				
	(0.41)	(0.45)				
Black × Original Racial Justice Tweets (%)			-0.25	-0.21	-0.89	-0.79
			(1.61)	(1.65)	(1.66)	(1.68)
Black × Non-Original Racial Justice Tweets (%)			2.27***	1.93***		
			(0.53)	(0.56)		
Black × Racial Justice Retweets (%)					2.21***	1.87***
					(0.66)	(0.69)
Black × Racial Justice Quote Retweets (%)					4.68*	4.27*
					(2.55)	(2.58)
Black × Racial Justice Tweet Replies (%)					4.29	4.11
					(3.35)	(3.40)
Observations	11,393	11,393	11,393	11,393	11,393	11,393
Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Email Type Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Twitter	All	Twitter	All	Twitter	All
p-value			0.18	0.26	0.10	0.16

Notes: Each variable shown in the table interacted with Black is winsorized at the 99th percentile. Sample mean of Racial Justice Tweets (%) = 1.34; Original Racial Justice Tweets (%) = 0.23; Non-Original Racial Justice Tweets (%) = 1.09; Racial Justice Quote Tweets (%) = 0.16; Racial Justice Tweet Replies (%) = 0.09. The denominator for each of these measures is the total number of tweets made from January 1, 2020 to March 27, 2022. Before winsorizing, Racial Justice Tweets is equal to the sum of Original Racial Justice Tweets and Non-Original Racial Justice Tweets; while Non-Original Racial Justice Tweets is equal to the sum of Racial Justice Retweets, Quote Tweets, and Tweet Replies. All regressions also include level variables for the interaction term (winsorized for those shown in the table). Odd columns additionally include the number of original tweets, reply tweets, retweets, and quote reply tweets, all for the period January 1, 2020 to March 27, 2022, and each of these variables interacted with Black. Even columns include the additional variables used in the fifth-from-the-right coefficient in Figure 6, as well as the fraction of academics followed that are Vocal and a dummy for this variable being missing; with each interacted with Black. The bottom row p-value tests for equality of the Original and Non-Original interactions in columns 3 and 4, and equality of Original and Retweets interactions in columns 5 and 6. Standard errors are clustered at the university-by-department-by-sender name-level. *** p<0.01, ** p<0.05, * p<0.1.

Online Appendix

Table A1: Audited Academics Are Similar on Most Observables to Non-Audited Academics

	Audited *N = 11,450*	Not Audited *N = 7,064*
	Mean	Mean
Female	0.42	0.42
Rank of University	59.08	51.50
Full Professor	0.36	0.40
Assistant Professor	0.35	0.32
Associate Professor	0.29	0.28
Business	0.04	0.03
Engineering and Technology	0.15	0.14
Humanities	0.14	0.16
Life Sciences	0.23	0.21
Physical Sciences	0.07	0.08
Professional Schools	0.12	0.11
Social Sciences	0.25	0.27
Black	0.00	0.00
White	0.79	0.78
East Asian	0.08	0.09
Hispanic	0.04	0.05
South Asian	0.07	0.07
Other Race/Ethnicity	0.01	0.01
Years On Twitter	8.60	8.71
Number of Tweets	1,022.01	1,206.86
Number of Twitter Followers	2,781.09	3,425.69
Number of Accounts Following	754.98	775.73
Vocal	0.62	0.65
Percentage of Racial Justice Tweets	1.41	1.50
Fraction of Vocal Academics Following	0.76	0.77
Blindspotter % Left-Wing Score	58.46	59.01
Republican Accounts Following	0.23	0.23
Democratic Accounts Following	1.76	1.98
Any Republican Contributions	0.01	0.01
Any Democratic Contributions	0.30	0.31
Total Republican Contributions (USD)	1.55	2.72
Total Democratic Contributions (USD)	244.59	332.95

Notes: Column 1 shows variable means for the audited academics and Column 2 covers the non-audited academics (all non-Black). The following are dummy variables: Female, Full Professor, Assistant Professor, Associate Professor, Business, Engineering and Technology, Humanities, Life Sciences, Physical Sciences, Professional Schools, Social Sciences, White, East Asian, South Asian, Hispanic, Other Race/Ethnicity, Vocal, and Any Republican/Democratic Contributions. Rank of University ranges from 1 to 150. Years On Twitter is the number of years since the academic joined Twitter, as of May 10, 2022. Number of Tweets is the number of tweets of any type made from January 1, 2020 to March 27, 2022. Number of Twitter Followers and Accounts Following is as of May 10, 2022. Vocal is equal to one for academics that tweeted at least once about racial justice from January 1, 2020 to March 27, 2022. Percentage of Racial Justice Tweets is the number of racial justice-related tweets as a percentage of the total number of tweets, over the same time period. Fraction of Vocal Academics Following is the fraction of academics (in our sample) followed that are Vocal. The Blindspotter score is a measure of the left-wing slant of the news the academic engages with on Twitter. Republican Accounts Followed is the number of Republican Senators and House Representatives followed as of July 2022 (similar for Democrats). Any Republican Contributions is equal to one if the academic is linked to at least one FEC-reported political contribution to a Republican FEC Committee from January 1, 2020 to March 27, 2022 (similar for Democrats). Total Contributions are for the same period. The Fraction of Vocal Academics Following is missing for 3% of the full sample, the Blindspotter score is missing for 4% of the full sample, and Republican and Democratic Accounts Followed are missing for 0.5% of the full sample. The table shows the means for the academics with non-missing data.

Table A2: Donation Study Summary Statistics

	N	Min	Mean	Max
Age (Years)	1,704	18.00	37.42	85.00
Female/Woman	1,704	0.00	0.49	1.00
Male/Man	1,704	0.00	0.47	1.00
Non-binary/Prefer Not to Answer	1,704	0.00	0.03	1.00
Race not White or European	1,704	0.00	0.12	1.00
Hispanic, Latino, or Spanish Origin	1,704	0.00	0.12	1.00
Political Views (1 = v. liberal, 5 = v. conservative)	1,704	1.00	2.38	5.00
Income Category (1 = less than 25k USD, 6 = 150k USD or more)	1,704	1.00	2.97	6.00
Number of Posts Per Month Since Joined Twitter	1,704	0.06	48.45	1,975.32
Amount Donated to NAACP (0 to 5 USD)	1,704	0.00	1.56	5.00
Tweeted in Support of Racial Justice May to October 2020	854	0.00	0.37	1.00
Retweeted NAACP Tweet (Manually Checked)	1,704	0.00	0.13	1.00
Raw Support for NAACP (0 to 100)	1,704	0.00	52.06	100.00
Binary Support for NAACP	1,704	0.00	0.13	1.00

Notes: The table shows summary statistics for the participants that completed the donation study. Political views are self-reported as either 1 = Very Liberal, 2 = Liberal, 3 = Moderate, 4 = Conservative, and 5 = Very Conservative. The income categories are 1 = less than $25k, 2 = $25k-49,999, 3 = $50k-74,999, 4 = $75k-99,999, 5 = $100k-149,999, 6 = $150k or more.

Table A3: Audit Experiment Balance Check

	Female	Assistant Professor	Associate Professor	White	Number Of Tweets	Vocal	Number Of Followers	Any Democrat Contributions
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
					(a) Balance for Full Sample			
Black	-0.01*	0.00	-0.01*	-0.00	34.75	-0.00	-119.21	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(53.80)	(0.01)	(610.15)	(0.01)
					(b) Balance by Vocality			
Black × Vocal	-0.01	0.01	-0.01	0.00	8.79		-954.58	-0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(82.09)		(751.60)	(0.01)
Black × Silent	-0.01	-0.00	-0.02	-0.02	86.60*		1260.24	-0.02
	(0.01)	(0.01)	(0.01)	(0.01)	(45.85)		(1321.36)	(0.01)
Observations	11,393	11,393	11,393	11,393	11,393	11,393	11,393	11,392
Full Sample Outcome Mean	.42	.35	.29	.79	1,023	.62	2,789	.3
p-value (Black × Vocal = Black × Silent)	.97	.63	.64	.25	.42		.18	.26
Vocal Dummy (Panel (b) only)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Email Type Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Standard errors are clustered at university-department-sender name-level. Black is a dummy variable equal to one for receiving an email from a distinctively Black name. Outcome variables are: (1) dummy variable for female academic, (2) dummy variable for Assistant Professor, (3) dummy variable for Associate Professor, (4) dummy variable for White academic, (5) number of tweets (Jan 1, 2020 to Mar 27, 2022), (6) dummy variable for whether tweeted about racial justice (Jan 1, 2020 to Mar 27, 2022), (7) number of Twitter followers as of May 10, 2022, and (8) dummy variable for any FEC-linked contributions to Democrat-related committees (Jan 1, 2020 to Mar 27, 2022). *** p<0.01, ** p<0.05, * p<0.1.

Table A4: Racial Justice Tweets Predict High Stakes Behaviors: Robustness

	Would Take Prof's Class Again (%) (1)	Work on Racial Topic (AI-coded) (2)	Black Coauthor (0-1) (3)
	(a) Without Controls		
Vocal	2.75***	0.19***	0.02**
	(0.71)	(0.01)	(0.01)
	(b) With Preferred Controls		
Vocal	1.61**	0.06***	0.03***
	(0.77)	(0.01)	(0.01)
Observations	8,676	16,753	15,801
Silent Mean	67.82	0.21	0.50
Silent SD	31.68	0.40	0.50

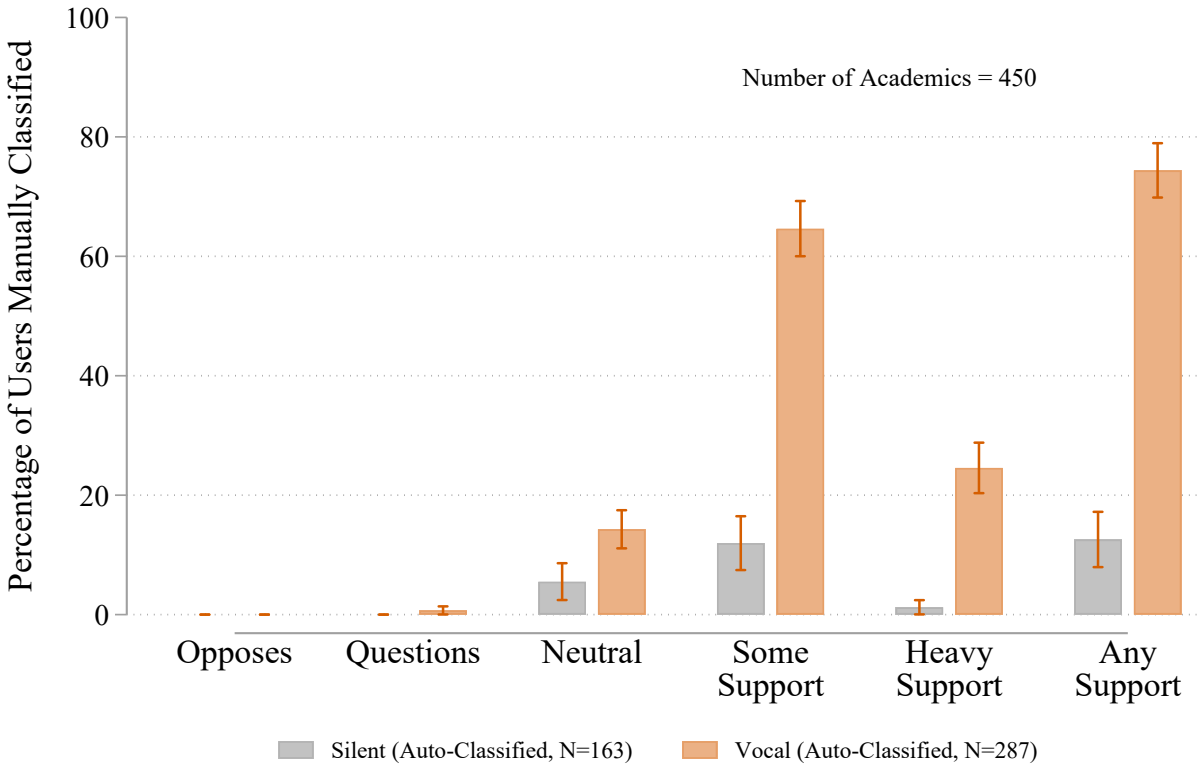
Note: Sample is on-Twitter non-Black academics. Outcomes: (1) Percentage of raters that answered "Yes" to " Would you take this professor again?", (2) dummy variable indicating whether the academic has any work on a race-related topic since 2000, with the topic coding done with gpt-4o, (3) dummy variable indicating that the academic has any work with a Black coauthor, based on the coauthor's name and for works between 2020 and 2022. Vocal is a dummy variable indicating whether the academic tweeted at least once about racial justice from January 2020 to March 2022. Silent Mean and Silent SD are the mean and standard deviation of the outcome among those with Vocal = 0. Controls in panel (b) are as in Table 1. Standard errors are robust.

Table A5: Graduate Students Summary Statistics

	Black Students				Non-Black Students			
	N	Min	Mean	Max	N	Min	Mean	Max
Age (Years)	515	23.50	29.52	54.50	1,237	19.50	29.24	54.50
Female/Woman	515	0.00	0.60	1.00	1,237	0.00	0.58	1.00
Male/Man	515	0.00	0.34	1.00	1,237	0.00	0.36	1.00
Non-binary/Genderfluid/Genderqueer	515	0.00	0.05	1.00	1,237	0.00	0.05	1.00
Black or African American	515	1.00	1.00	1.00	1,237	0.00	0.00	0.00
White or European	515	0.00	0.07	1.00	1,237	0.00	0.58	1.00
Asian	515	0.00	0.03	1.00	1,237	0.00	0.34	1.00
First Nations or Indigenous	515	0.00	0.01	1.00	1,237	0.00	0.01	1.00
Native Hawaiian or Other Pacific Islander	515	0.00	0.00	0.00	1,237	0.00	0.00	1.00
Hispanic, Latino, or Spanish origin	515	0.00	0.07	1.00	1,237	0.00	0.11	1.00
Born in the USA	515	0.00	0.70	1.00	1,237	0.00	0.57	1.00
Studying for a PhD	515	0.00	0.98	1.00	1,237	0.00	0.97	1.00
Rank of University (1 to 80)	515	1.00	35.42	78.00	1,237	1.00	35.55	80.00
Year in Graduate Program	515	1.00	3.67	7.00	1,237	1.00	3.86	7.00
Political Views (1 = v. liberal, 5 = v. conservative)	515	1.00	1.86	5.00	1,237	1.00	1.90	5.00
Has Twitter	515	0.00	0.75	1.00	1,237	0.00	0.60	1.00
How Often Tweets About Racial Justice (1 to 4)	384	1.00	2.45	4.00	737	1.00	1.94	4.00

Notes: Columns 1 to 4 show summary statistics for the surveyed students that self-identify as Black or African American, while columns 5 to 8 show summary statistics for the remaining surveyed students. Age (Years) is the mid-point of three-year categorical answers, with the exception of the category "Over 53" where we code Age as 54.5. Female/Woman, Male/Man, and Non-binary/Genderfluid/Genderqueer are binary measures of gender identity. The handful of respondents who are not studying for a PhD are either Master's students or Postdoctoral Fellows. Political views are self-reported as either 1 = Very Liberal, 2 = Liberal, 3 = Moderate, 4 = Conservative, and 5 = Very Conservative. How Often Tweets About Racial Justice was only asked to students that have Twitter, with 1 = Never, 2 = Rarely, 3 = Sometimes, and 4 = Often.

Figure A1: Validating the User-level Measure of Vocality



Notes: This figure validates our automated signaling algorithm. The team scrolled through the post-May 2020 Twitter feeds of a random subset of our experimental sample ($N = 450$), 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 287 academics we automatically classify as Vocal. The grey bars include data for the 163 academics we automatically classify as Silent. As an example, the orange "Neutral" bar shows the percentage of users that the team *manually* found to have ever tweeted neutrally about racial justice, only among the academics that we *automatically* classify as Vocal. The "Any Support" category shows the percentage of users that ever tweeted some or heavy support. 95% confidence intervals are shown. [Note: The figure differs slightly from the corresponding one in our pre-analysis plan because we updated our automated measure of vocality to (i) incorporate racial justice-related words and phrases in tweeted hyperlinks (as promised in our pre-analysis plan), and (ii) include the full text of all retweets, correcting an error in our earlier measure. See Appendix C for more details.]

Figure A2: Donation Study: Measuring Donations to the NAACP

We are supporting the NAACP. The NAACP's vision is to have a world without racism where Black people enjoy equitable opportunities in thriving communities.

The NAACP advocates for Black individuals killed by police, and also calls lawmakers to pass police reform legislation at the state and national levels. As one example, Tyre Nichols (picture below) was fatally beaten by five police officers in January 2023. The NAACP advocated for those involved in his death to be prosecuted. A childhood friend of Mr. Nichols told the BBC his legacy would be preserved through legal reform and that "he always wanted to change the world."

As compensation for answering our survey, you will receive an additional $5. You can choose to give any or none of that $5 to the NAACP. As with the other questions, your answer will not be shared with anyone.

Please use the scale below to indicate how much you want to donate (if any).

Mr Nichols died three days after an encounter with police at a traffic stop

Donate nothing Donate everything

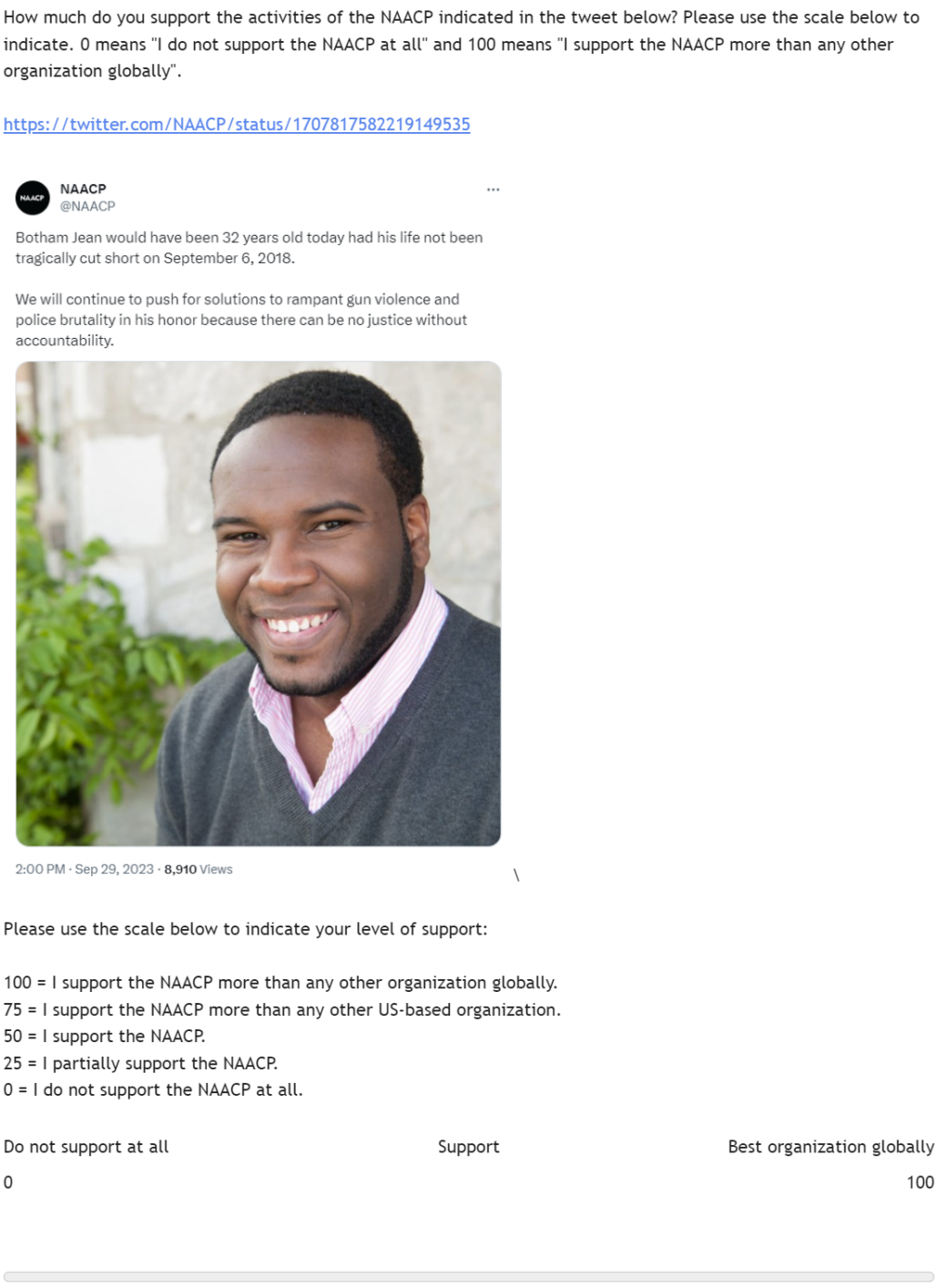
0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5

Notes: The figure shows the Qualtrics screen in the donation study, where participants decide how much of their $5 survey incentive to donate to the NAACP.

Figure A3: Donation Study: Opportunity to Retweet an NAACP Tweet

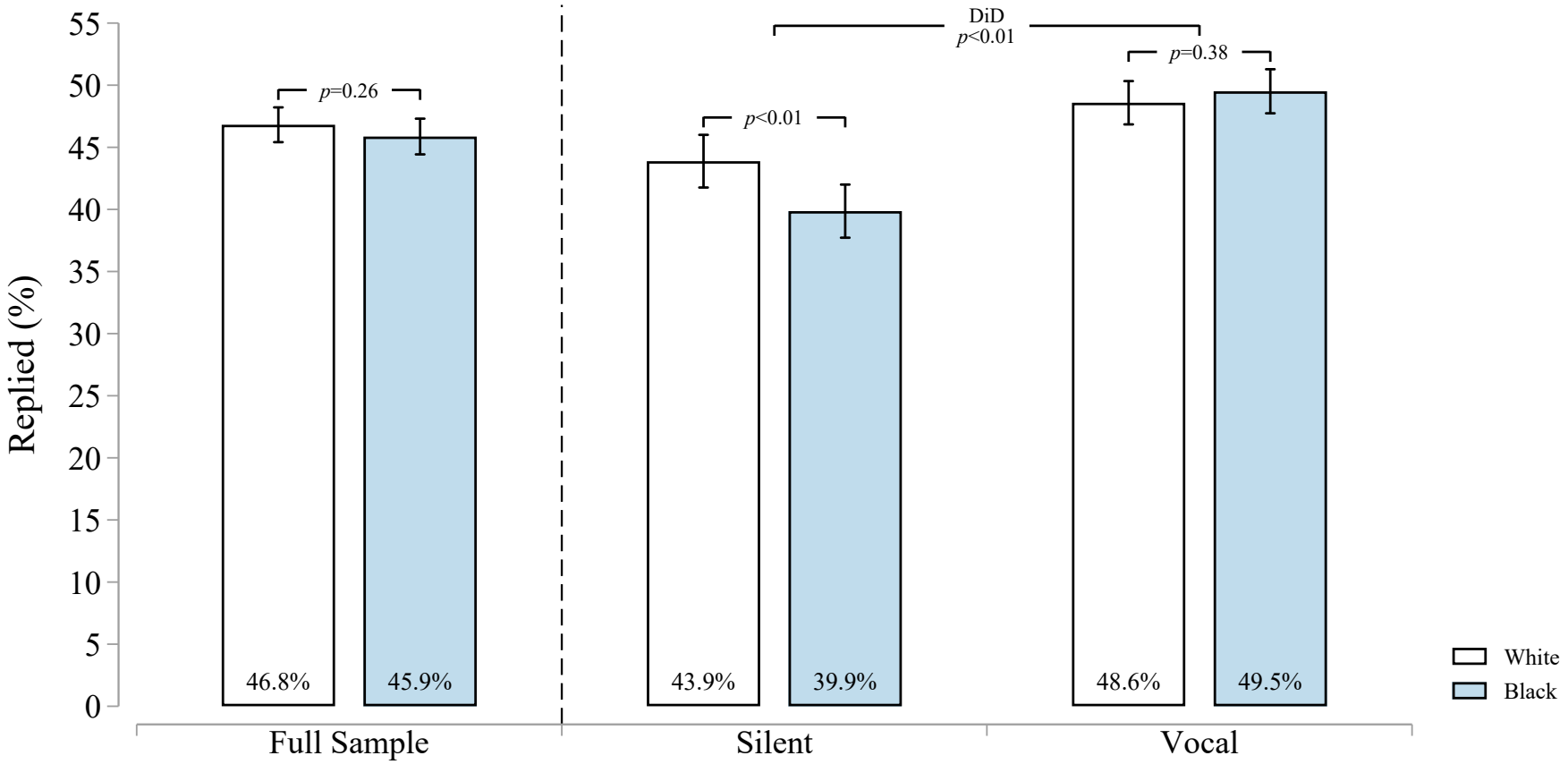
Notes: The figure shows the Qualtrics screen in the donation study, where participants decide whether to retweet an NAACP tweet. Participants cannot advance to the next screen until a 30-second timer ends. This question and the support question are asked in random order.

Figure A4: Donation Study: Measuring Private Support for the NAACP



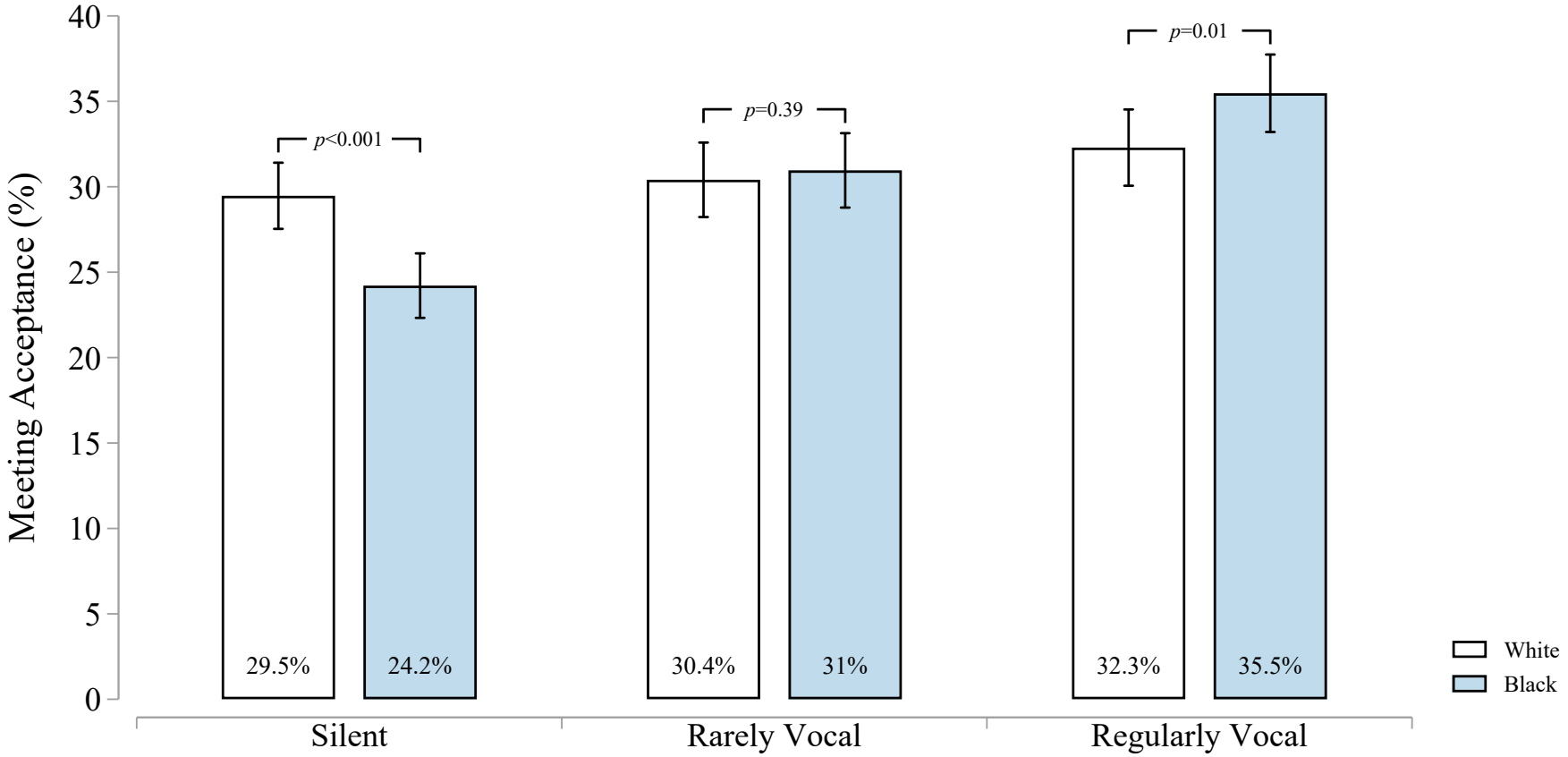
Notes: The figure shows the Qualtrics screen in the donation study, where participants indicate their support for the NAACP on a scale from 0 to 100. Participants cannot advance to the next screen until a 30-second timer ends. This question and the retweet question are asked in random order.

Figure A5: Replied: Vocal Professors Discriminate Against Black Students 4.9 Percentage Points Less



Notes: The bars show what percentage of audited academics replied to meeting requests from distinctively White and distinctively Black names. The Full Sample includes the 11,450 audited academics (4,318 Silent and 7,132 Vocal). Vocal academics are those that tweeted at least once about racial justice from January 2020 to March 2022. Silent academics are those that did not. The raw means and 95% confidence intervals come from a regression of Replied_i on dummy variables for White and Black email sender (to the left of the vertical dashed line), and a regression on dummy variables for White email sender to Silent academic, White email sender to Vocal academic, and the same for Black email sender (to the right of the vertical dashed line). The p-values come from the specification that also includes strata and email type fixed effects. The DiD (diff-in-diff) p-value is from a test for equal discrimination rates across Vocal and Silent academics ($\hat{\gamma}_2$ in specification 2). Standard errors are clustered at the university-by-department-by-sender name-level.

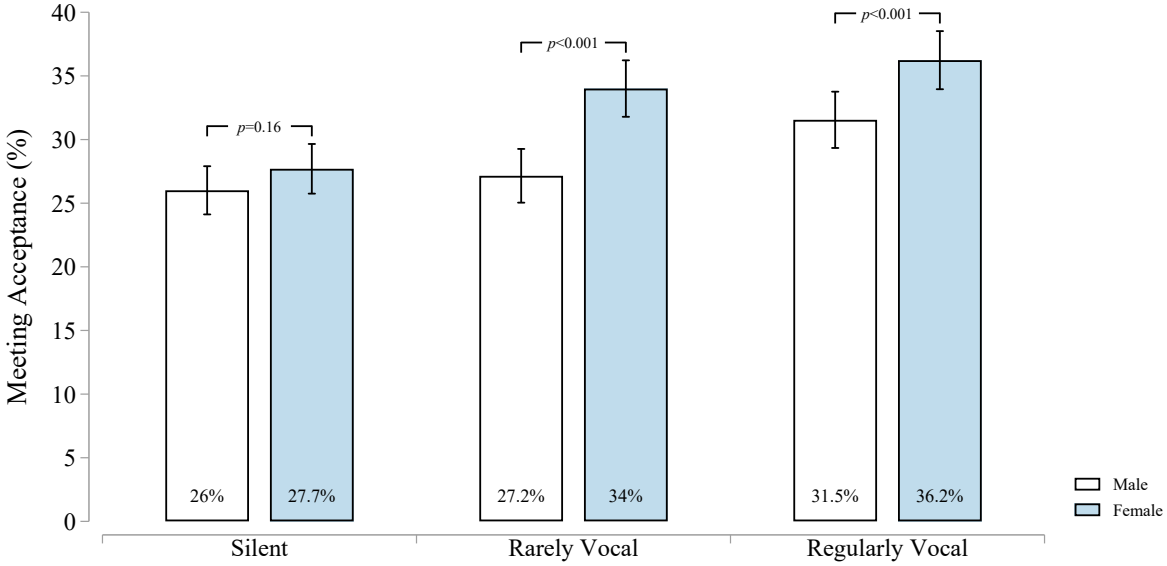
Figure A6: The Rarely Vocal Are Unbiased, the Regularly Vocal Favor Black Students



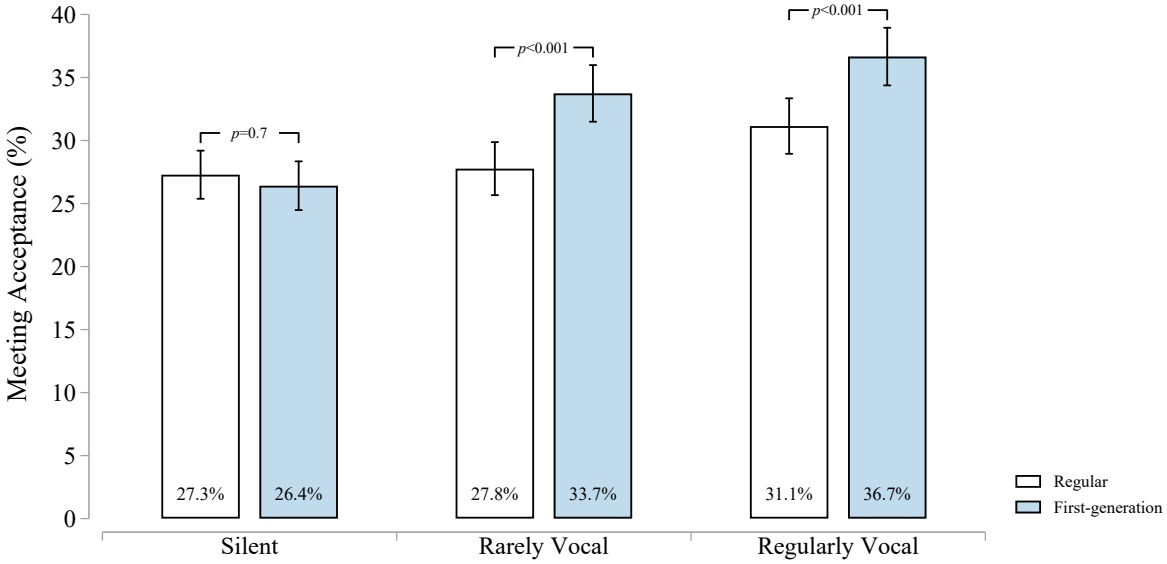
Notes: The bars show what percentage of audited academics accepted meeting requests from distinctively White and distinctively Black names. Silent academics are those that did not tweet about racial justice from January 2020 to March 2022. The bars for the Silent academics replicate the findings in Figure 3. Among the Vocal, the Rarely Vocal are those with below-median percentage of tweets from January 2020 to March 2020 that are about racial justice (0.6% on average), while the Regularly Vocal academics are above-median (3.9% on average). The raw means and 95% confidence intervals come from a regression of Accepted_i on dummy variables for Black email sender to Silent academic, White email sender to Silent academic, and the same for emailed to the Rarely Vocal and to the Regularly Vocal. The p-values come from the specification that also includes strata and email type fixed effects. Standard errors are clustered at the university-by-department-by-sender name-level.

Figure A7: The Rarely and Regularly Vocal Favor Female and First-Generation Students Similarly

(a) Female vs. Male Names



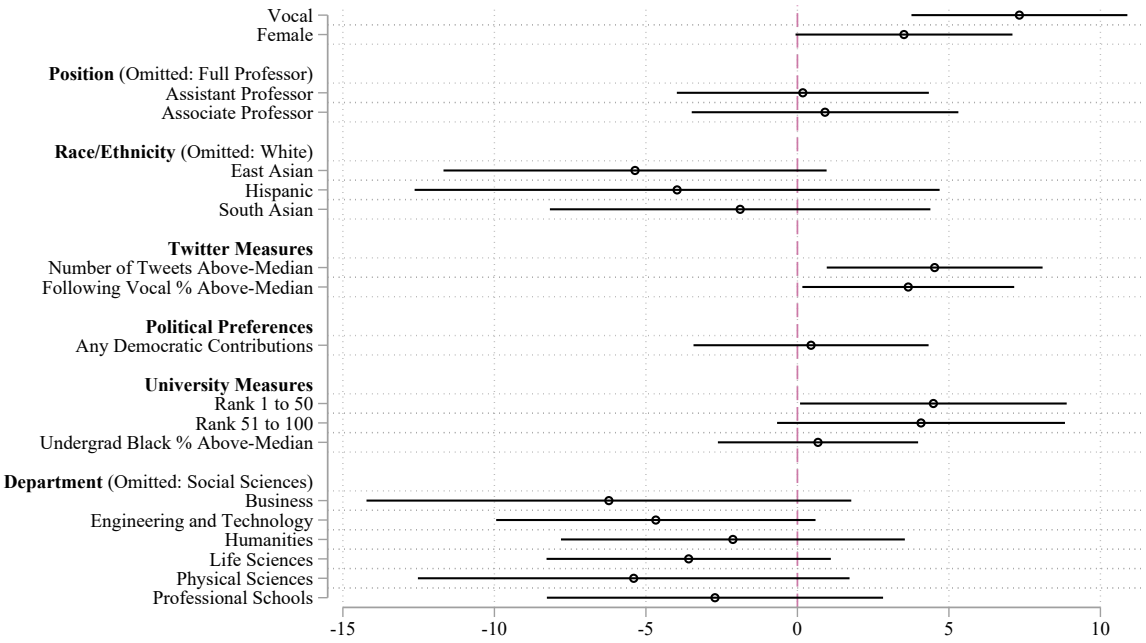
(b) First-Generation vs. Regular Students



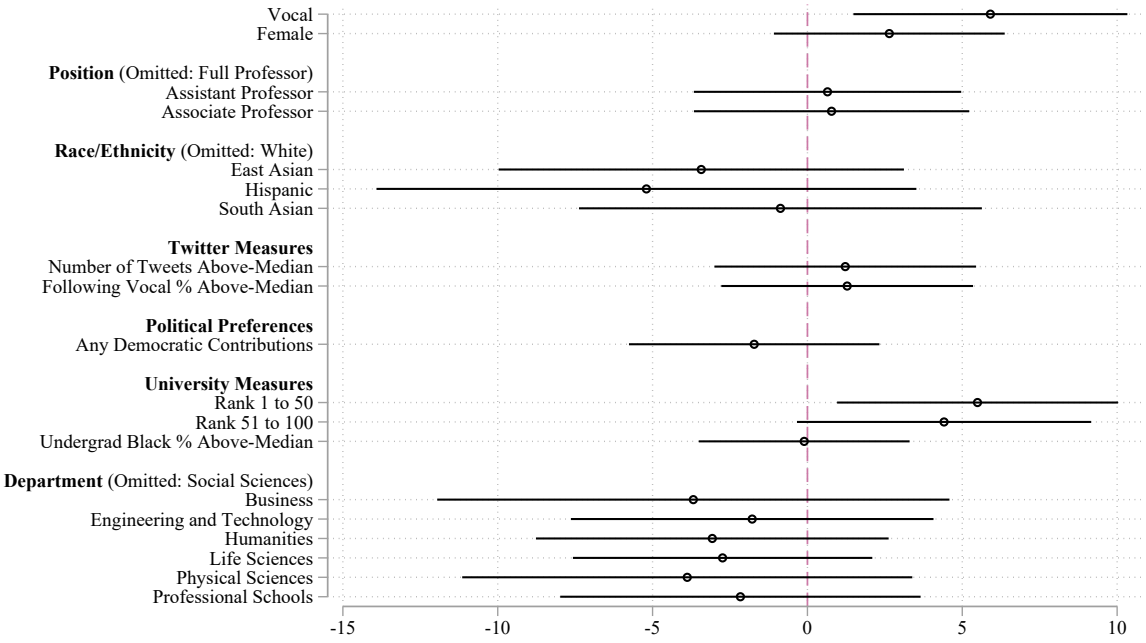
Notes: Silent academics are those that have not tweeted about racial justice. The bars for the Silent academics replicate the findings in Figure 4. Among the Vocal, the Rarely Vocal are those with below-median percentage of tweets from January 1, 2020 to March 27, 2020 that are about racial justice (0.6% on average), while the Regularly Vocal academics are above-median (3.9% on average). In panel (a), the raw means and 95% confidence intervals come from a regression of Accepted_i on dummy variables for female email sender to Silent academic, male email sender to Silent academic, and the same for emailed to the Rarely Vocal and to the Regularly Vocal. Panel (b) is similar, but for first-generation versus regular email senders. The p-values come from the specification that also includes strata and email type fixed effects. Standard errors are clustered at the university-by-department-by-sender name-level.

Figure A8: What Predicts Racial Discrimination?

(a) Factors One-By-One

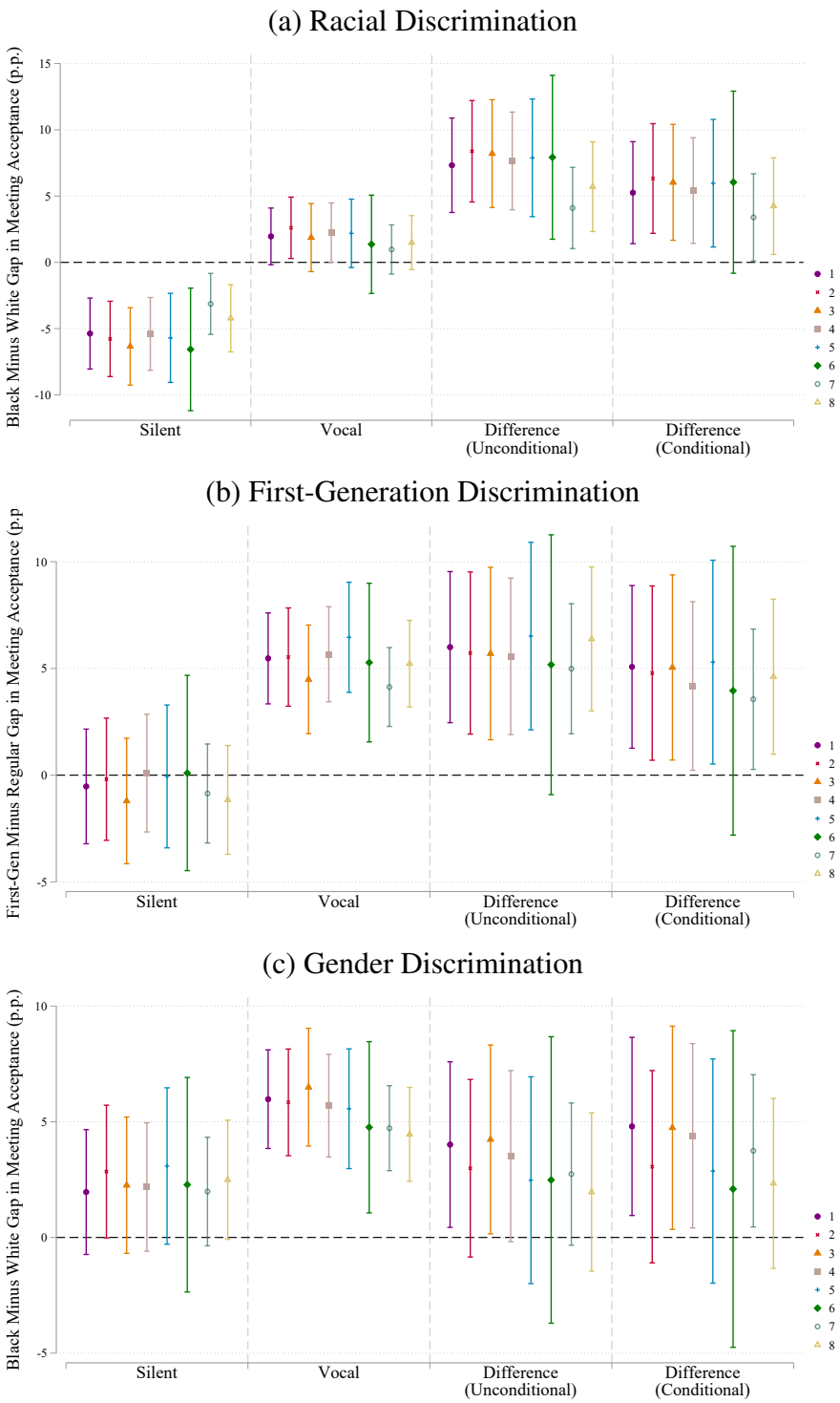


(b) All Factors Together



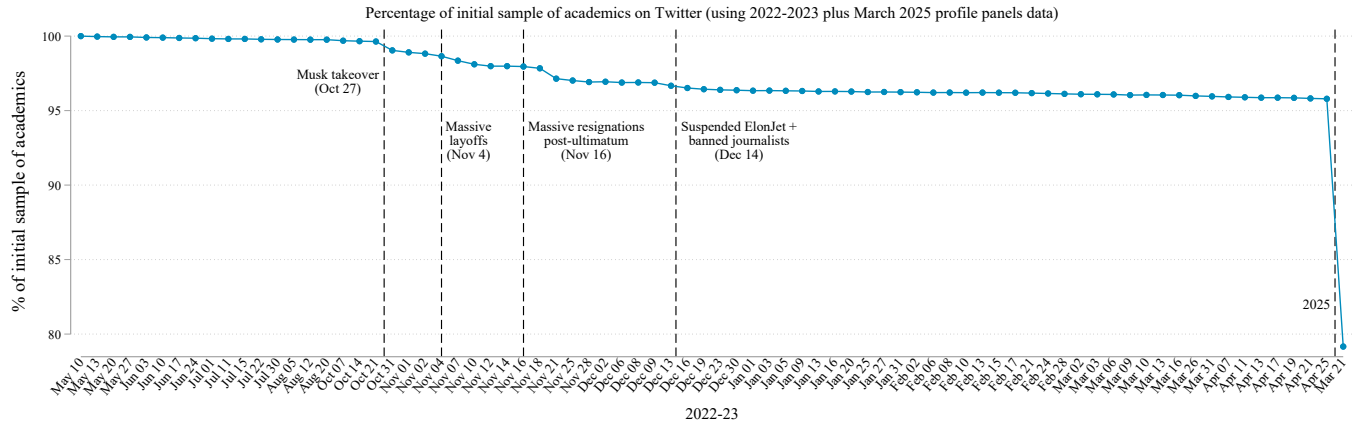
Notes: The figure shows which factors predict racial discrimination. Specifically, the figure shows the $\hat{\gamma}_2$ and $\hat{\theta}_j$s (along with 95% confidence intervals) estimated from specification 2, where the X^js are the same as those used in Figure 2. In panel (a), each coefficient is from a separate regression, with only one factor included at a time. Panel (b) reports the coefficients from one regression in which all factors are included at once. The coefficients come from specification 2, with the following variables interacted with Black: (1) Vocal, (2) Female, (3) Position dummies, (4) Race/Ethnicity dummies, (5) Number of Tweets Above-Median, (6) Any Democratic Contributions, (7) Rank dummies, (8) Undergrad Black % Above-Median, and (9) Department dummies. As an example, the first coefficient in panel (a) tells us that Vocal academics discriminate against Black students 7.3 percentage points less than Silent academics. Whereas the first coefficient in panel (b) tells us that Vocal academics discriminate against Black students over five percentage points less than Silent academics, holding the other variables in the figure constant.

Figure A9: Detection Is Unlikely to Explain the Audit Results



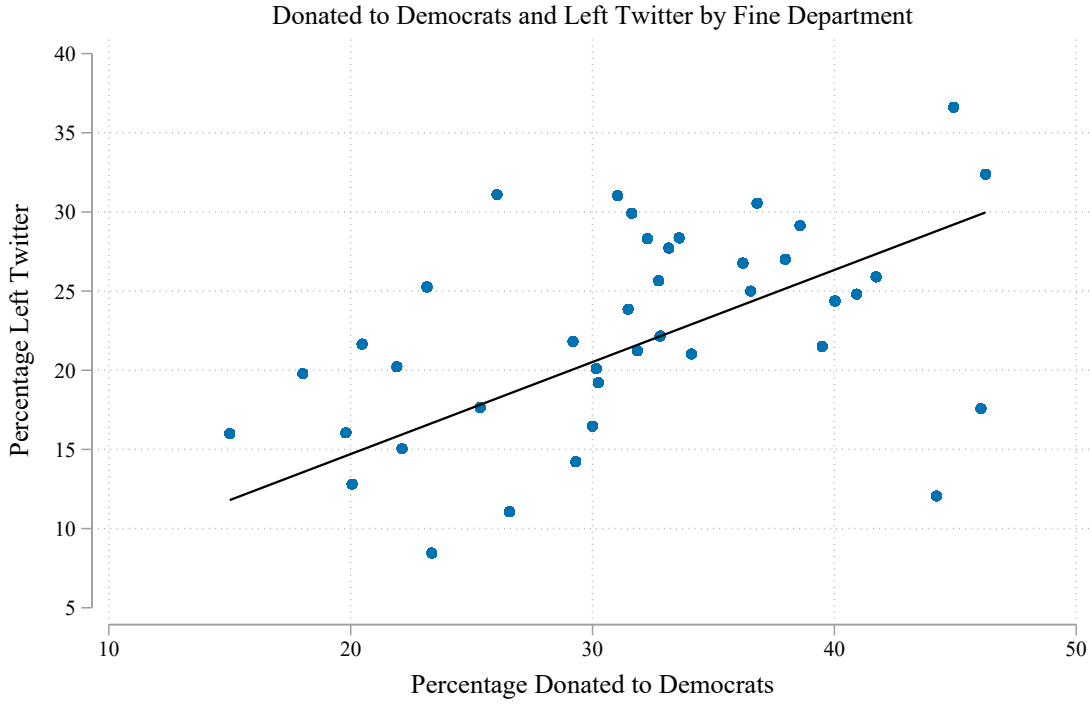
Notes: The figure shows the robustness of our main results (coefficient set 1) to seven alternative samples or outcomes where we expect audit detection to be less likely (coefficients sets 2 to 8). Specifically, for each of the eight we show from left to right: the discriminatory gap in meeting acceptance for Silent academics (negative for discrimination against Black students in panel (a)), the gap for Vocal academics, the unconditional difference in the gap ($\hat{\gamma}_2$ from specification 2 without any X_i^j covariates, positive in panel (a) if Vocal academics discriminate against Black students less than Silent academics), and the conditional difference using the fifth-from-the-right specification from Figure 6. The specification and sample variants are (percentage of the sample dropped in parentheses): (2) drop academics in Economics, Political Science, Sociology, and Business (12.4%), (3) drop academics in the Social Sciences (25%), (4) drop academics to whom we sent more generic emails (7%), (5) drop university-departments to which we sent more than ten emails (27%), (6) drop university-departments to which we sent more than five emails (61%), (7) outcome is meeting accepted within one day, and (8) outcome is meeting accepted within three days. Standard errors are clustered at the university-by-department-by-sender name-level. 95% confidence intervals are shown.

Figure A10: Academics Began Leaving Twitter After the Elon Musk Takeover



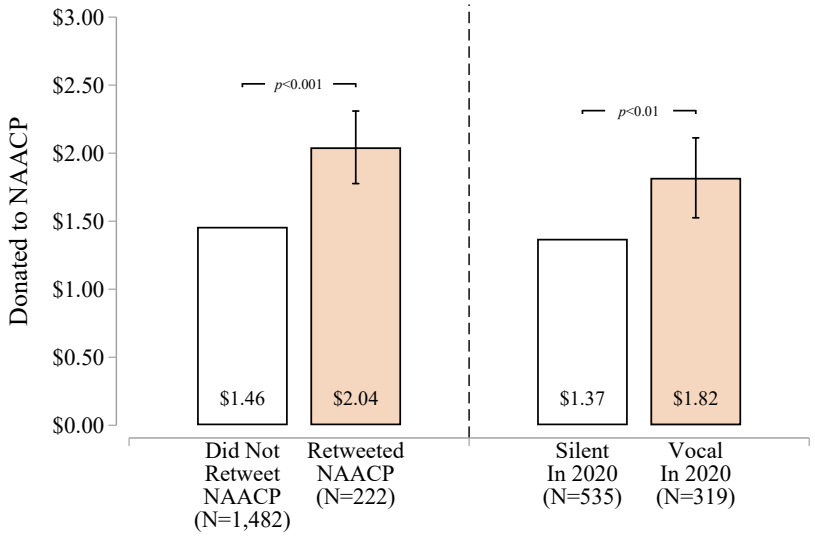
Notes: The figure shows what percentage of our initial sample of 18,514 tweeting academics (collected May 2022) remained on Twitter in the subsequent months. We have snapshots every few days from May 2022 to April 2023, and then one additional snapshot in March 2025.

Figure A11: Academics in Democrat-Supporting Departments Were More Likely to Leave Twitter



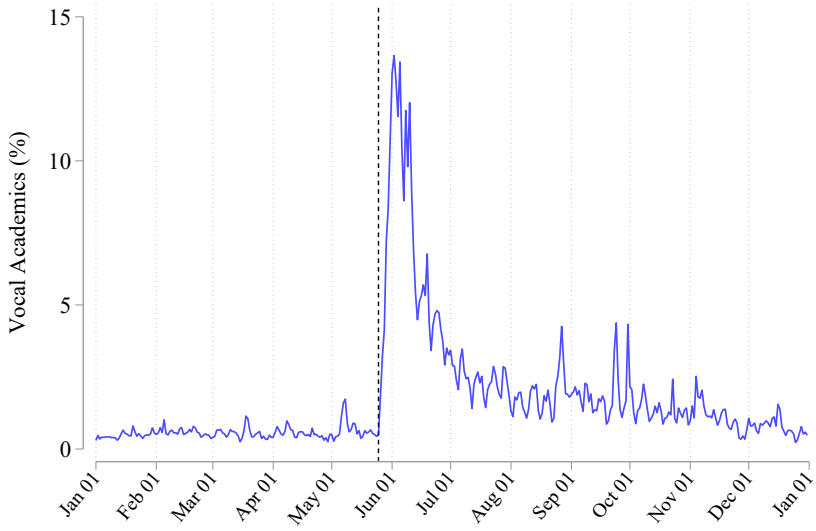
Notes: The figure plots department-level data for the 40 academic departments (e.g. Sociology, History) with at least 100 tweeting academics each in our data. The y-axis is the percentage of tweeting academics in a given department that had left Twitter by March 2025. The x-axis is the percentage of tweeting academics in a given department that contributed to Democratic federal political campaigns during 2020 to 2022 (using FEC data). The line is a linear fit.

Figure A12: Vocal Participants Donate More to the NAACP, Conditional on Controls



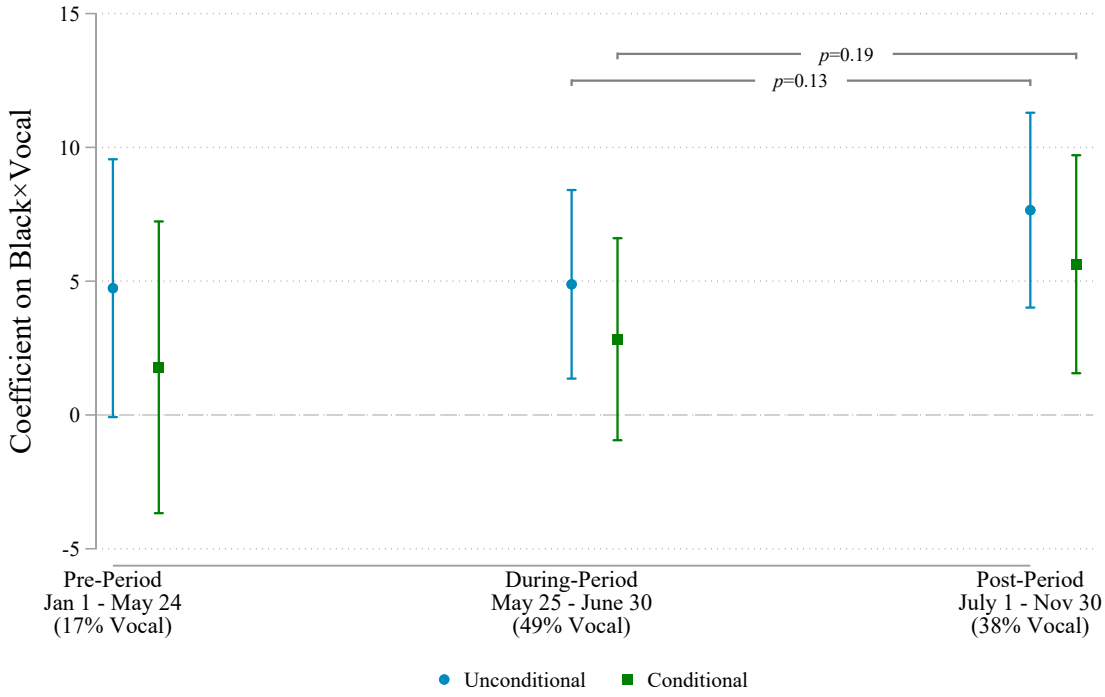
Notes: The figure shows the mean NAACP donation (ranging from $0 to $5) for participants in the donation study that (i) did versus did not retweet the NAACP tweet (with the gap conditional on controls), and (ii) tweeted in support of racial justice during May to October 2020 versus did not (with the gap conditional on controls). The latter is shown only for the subsample for which we could manually code tweets made from May to October 2020. The control variables included are: dummy variable for female, dummy variables for income categories, dummy variables for political views, dummy variables for non-White and Hispanic, and age. We also control for a measure of Twitter activity: number of posts per month since joining Twitter for the NAACP-retweet regression, and the number of posts made from May to October 2020 (coded as 51 for 51 plus), plus a dummy variable for 51 or more posts from May to October 2020, for the Vocal-in-2020 regression. 95% confidence intervals are shown.

Figure A13: Racial Justice Tweeting Spiked After the Murder of George Floyd



Notes: The figure shows the percentage of audited academics ($N = 11,450$) that tweeted about racial justice each day in 2020. See Section 2.1 for details on how we identify racial justice tweets. The vertical dashed line denotes the murder of George Floyd on May 25th.

Figure A14: Informativeness Is Higher When Fewer People Are Tweeting About Racial Justice (Binary Measure)



Notes: The figure shows how tweet informativeness ($\hat{\gamma}_2$ from specification 2) changes before, during, and after the murder of George Floyd on May 25, 2020. 17% of the 11,450 audited academics tweeted about racial justice at least once during January 1 to May 24, rising to 49% during May 25 to June 30, and falling to 38% during July 1 to November 30. The unconditional estimate denotes the unconditional difference in audit-measured racial discrimination between the Vocal and Silent during each period. The conditional estimate denotes the conditional difference in discrimination, using the fifth-from-the-right specification from Figure 6. These estimates are positive when Vocal academics discriminate against Black students less than Silent academics. Standard errors are clustered at the university-by-department-by-sender name-level. 95% confidence intervals are shown.

A The Research Assistant Team

The audit experiment was only possible because of the dedication of the following research assistants (most of whom are UBC undergraduates):

G.O.A.T.: Akash Uppal, Albena Vassileva, Aurellia Sunarja, Carla Colina, Carlos Perez Cavero, Chihiro Tanigawa, Colby Chambers, Conor McCaffrey, Daniella Rolle, Esha Vaze, Eugene Kwok, Jiayu Li, Jordan Hutchings, Kevin Yu, Laura Truong, Louise Cheng, Maria Ines Moran, Noor Kumar, Saloni Sharma, Shardha Nayar, Tierra Habedus-Sorensen, Vinayak Kalra, Yash Ahlawat.

Tier 1: Angela Fan, Bryanna Li, Jaida Smith, Nicholas Latimer, Olivia Klaassen, Angela Lee, Ty Stevenson, Alex Dyky, Amir Ala'a, Avreeta Sandhu, Billy Lam, Cynthia Cui, Gabriel Odeyemi, Jennifer Tran, Jenny Qin, Jeremy Singer, Julianne Nina Marie Uy, Karman Phuong, Kaye Thinh-To, Keshikaa Suthaaharan, Kevin Li, Kevin Tan, Mahrukh Khan, Marianne Sigouin, Maxwell Martel, Nela Radecki, Nikita Gautham, Paulino Tan, Robin Jhatu, Sabine Villaroman, Shahed Salahi, Sophia Huang, Sophia Samilski, Trang Truong, Vanessa Cheung, Yu Fei, Harpreet Khattar, Lulu Wang, Minh Anh Pham, Rayan Aich, Uddhav Kalra

Tier 2: Odmaa Bayartsogt, Ahana Thakur, Alanna Man, Alejandra Mercadillo, Alejandro Solano Romero, Amit Biswas, Ana Beatriz Pereira, Anahat Kaur Chahal, Anastasia Mishina, Andrea Bartolome, Angela Sequeira, Angela Villavicencio, Becky Zhu, Bhakthie Senanayake, Chris Cheuk Yin Lam, Chris Haun, Connor Wiesner, Deniz Sagnak, Dhairya Chaudhri, Dhruv Bhatia, Emi Oyakawa, Emma Borhi, Erik Bruendl, Fariha Sultana, Florent Dusenge, Francesca Tamberi, Genevieve Moody, Huiqi Liu, Jacqueline Lee, James Brewster, Jane Platt, Jiayi Zhu, Jingyi (Olivia) Cao, Jinmeng Xu, Joaquin Glinoga, Joey Dolayba, Judith Sofiana Haryanto, Julian Kwan, Julianne Louie, Junye Xu, Kaitlin Khu, Karam Kanwar, Karen Liu, Kunwar Modi, Lauren Snow, Liri Zou, Liugu Tan, Lucas Mehling, Marco Lanfranchi, Matteo Tan Zheng Hao, Mridul Manas, Naoki Sakura, Navkiran Takhar, Nicholas Harterre, Percy Chen, Ravi Rinarco, Rohan Prasad, Ruby Taylor, Ruolin Mo, Sanchita Sannigrahi, Sarah Mejia, Sarir Parvizi, Shady Abo El Kasim, Shreya Iyer, Siddhant Kumar, Sirui Bi, Smriti Sukhani, Surotama (Suri) Banerjee, Syra Dhaliwal, Tim Qiao, Tosya Khodarkovsky, Valeria Zolla, Vivian Liu, Vivian Wei, Vyomesh Daga, Xiaoyan Wu, Xixi Xu, Yasin A Zahir, Yawen Zhang, Yuxuan Deng

B Model

We write a simple model to provide a framework for the empirics. More general models yield similar comparative statics (e.g. Frankel and Kartik 2019).

There are two types of agents, and these types are private. A fraction π of agents have high support for racial justice efforts ($\eta = \bar{\eta} > 0$), while the remaining fraction $1 - \pi$ have low support ($\eta = 0$). In our empirical exercises, we measure η as: (i) racial discrimination in accepting meeting requests from students (more pro-Black discrimination meaning a higher η), and (ii) private financial donations to the NAACP (larger donations meaning a higher η). η is then the type that determines private racial justice-related behaviors.[25]

Each agent chooses a binary action a, equal to one if they choose to tweet in favor of racial justice efforts (Vocal), and zero otherwise (Silent).[26] The psychic cost of tweeting in support of racial justice is zero for high-support types ($\eta = \bar{\eta}$) and $c > 0$ for low-support types ($\eta = 0$).[27] The cost can be thought of as an internal cost of lying or misrepresentation (Kartik 2009), or the cost of composing a tweet outside one's domain of expertise.

The audience of tweets is composed of two types: a fraction s are sophisticated and the remaining onlookers are cynical. The sophisticated onlookers form rational expectations about η based on whether an agent tweets or not. These onlookers accurately recognize the correlation between tweeting and support for racial justice in equilibrium. The cynical onlookers believe that tweets are uninformative about η. We model cynical onlookers, rather than overly-optimistic onlookers, given our empirical evidence in Section 3.5 that onlookers are much more likely to underestimate than overestimate informativeness.

All agents have social image concerns. They would like to be perceived as pro-racial justice by their audience – a natural assumption given that Twitter tends to be left-leaning, even after the Elon Musk takeover (Economist 2023), and given that academics are especially left-leaning, as we show using data on political contributions in Figure 1. Agents then receive a net benefit of tweeting relative to not tweeting equal to

$$b(v, s) = vs\left(\mathbb{E}[\eta \mid a = 1] - \mathbb{E}[\eta \mid a = 0]\right)$$

where v reflects the signaling stakes of tweets, s is the fraction of sophisticated onlookers (tweeting cannot contribute to social image gains from cynical onlookers), and $\mathbb{E}[\eta \mid a]$ reflects the rational expectations formed by these sophisticated onlookers. Several factors influence signaling stakes – for example, stakes are larger when tweets are more visible to others, more highly scrutinized conditional on being visible,

[25] While we also look at informativeness of racial justice tweets for higher-stakes behaviors (teacher ratings, topic choice, race of co-authors, whether left Twitter), we do not consider these behaviors to map directly to the private type η given that all of these behaviors are clearly visible to others.

[26] Frankel and Kartik (2019) instead consider a continuous action space and heterogeneity along two dimensions: racial justice support (the "natural action" in their language) and also social image concerns ("gaming ability"). They reach similar conclusions about information loss. The mechanism for information loss however differs in our model, where the loss is a consequence of the bounded action space leading to "bunching at the top" (see Appendix C in Frankel and Kartik (2022) for a model with the same mechanism).

[27] In our model, racial justice support perfectly correlates with tweet cost, ruling out types of people who support racial justice but are unwilling to tweet. Frankel and Kartik (2019) allow for this possibility and find similar comparative statics.

or when audiences increase the importance they place on others being high racial justice support types.

In a Bayesian Nash Equilibrium, (i) both types of agents choose whether to tweet, or not, or to mix over both strategies, and no agent would prefer to deviate to a different strategy, given audience beliefs, and (ii) sophisticated onlookers form accurate beliefs about the average η among agents that do and do not tweet. In a pooling equilibrium in which both types tweet about racial justice, we assume that the off-equilibrium path $\mathbb{E}[\eta \mid a = 0] = 0$, i.e. agents that deviate, by choosing not to tweet, are presumed to be low-support types.[28]

Our primary interest is in comparative statics for equilibrium informativeness, which we define as the mean difference $I = \mathbb{E}[\eta \mid a = 1] - \mathbb{E}[\eta \mid a = 0]$. Our empirical parallel in the audit experiment is the mean difference in racial discrimination between Vocal and Silent academics, while in the donation study it is the mean difference in NAACP donations between Vocal and Silent participants.

Proposition 1.

1. There is a threshold v^* such that a separating equilibrium exists if and only if $v \leq v^*$. By definition, there is full information revelation in this separating equilibrium, i.e. $I = \bar{\eta}$.

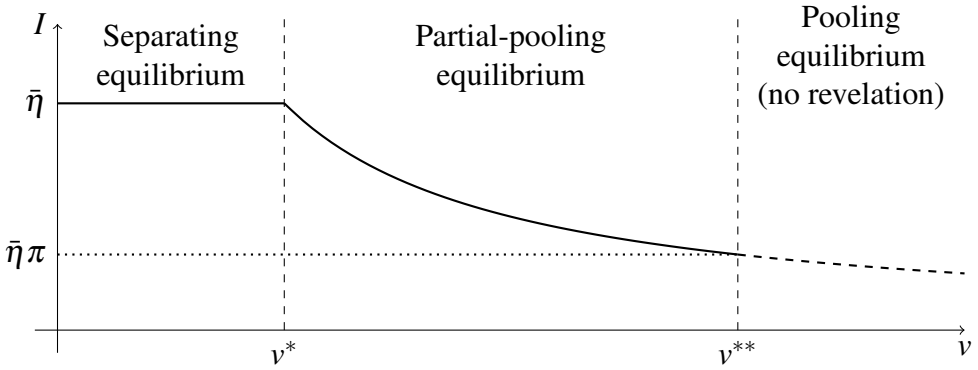
2. There is a threshold $v^{**} > v^*$ such that for $v \in (v^*, v^{**})$, a partial pooling equilibrium exists, with $I = \frac{c}{vs}$. It follows that informativeness is falling in visibility v and in audience sophistication s for intermediate levels of visibility v.

3. For $v \geq v^{**}$, there is a pooling equilibrium in which both types tweet about racial justice. There is no information revelation. See Appendix B.1 for proofs.

At low levels of signaling stakes ($v \leq v^*$) types are fully revealed: low-support types do not tweet because they do not gain enough social image from tweeting to cover their cost, c. In contrast, high-support types always tweet because for them it is costless, and they receive positive social image gains from signaling their type. In this low-stakes region, informativeness is unaffected by visibility.

Once stakes are above v^*, low-support types now have an incentive to deviate – the social image benefit has risen above the cost of tweeting. The low-support types shift to tweeting until the informativeness is sufficiently diluted as to make low-support types indifferent between tweeting and not tweeting. Here we have a partial pooling equilibrium, with low-support types mixing between the two strategies. Now informativeness is falling in signaling stakes, as higher stakes lead more low-support types to pool with the high-support types, reducing the signal value of tweets. Eventually we reach a threshold v^{**} in which there are no low-support types left to pool. At and above v^{**} – as long as observers assign a minimum level of perceived informativeness to tweets, assured by the assumption that $\mathbb{E}[\eta \mid a = 0] = 0$ – everyone tweets, and thus tweets do not reveal types at all.

[28] This assumption is sufficient but not necessary. The necessary assumption is that $\mathbb{E}[\eta \mid a = 0] \leq \pi\bar{\eta} - \frac{c}{vs}$ for all v such that there is a pooling equilibrium.



Notes: Figure shows signaling informativeness I as a function of the signaling stakes v. In the pooling region, the dashed line shows the minimum I that observers must perceive to sustain the equilibrium.

Summarizing, for low signaling stakes, informativeness is not affected by stakes, while for intermediate levels, informativeness is falling in stakes. For high levels, there is no information revelation.[29] The comparative statics with respect to audience sophistication are identical to those with respect to stakes. Intuitively, as the audience becomes more sophisticated, social image gains from tweeting increase, leading to more low-support types tweeting (in the intermediate region), and a dilution of the signal provided by tweets.

B.1 Model Proofs

Proofs for Proposition 1. [Prop. 1.1] Suppose there exists a fully separating equilibrium. In this equilibrium, $a(\eta = 0) = 0$ and $a(\eta = \bar{\eta}) = 1$, and so $I = \bar{\eta}$ (there is full revelation of types). The high-support type does not want to deviate by not tweeting about racial justice. If they deviated, their cost is the same, but their $b(v, s)$ falls because now people think that they are the low type. The low-support type doesn't want to deviate as long as the cost of tweeting outweighs the social image benefits: $c \geq vs\bar{\eta}$. It follows that a separating equilibrium exists iff. $v \leq v^* = \frac{c}{s\bar{\eta}}$.

[Prop 1.2] Suppose there is a partial pooling equilibrium. Low-support types mix between tweeting and not tweeting (they are indifferent). High-support types all tweet. Suppose that the fraction of tweeters that are low-support is r. For indifference of low-support types, we need: $vs(r(0) + (1-r)\bar{\eta} - 0) = c \Rightarrow$ $vs((1-r)\bar{\eta}) = c$

$$\Rightarrow 1 - r = \frac{c}{vs\bar{\eta}}$$

This ensures that informativeness is $I = (1-r)\bar{\eta} = \frac{c}{vs}$. Again, high-support types do not want to deviate. The mixing for low-support types requires that $r \in (0, \pi)$, which requires that $1 - r \in (\pi, 1)$ which using

[29] In more general models, informativeness falls strictly in signaling stakes (Frankel and Kartik 2019). Otherwise, plausible model extensions can flip the comparative static. For example, using our model, suppose instead that the net benefit of tweeting is $vsI - \alpha v^2$ for the low types, where αv^2 is the expected cost of one's hypocrisy being discovered. The convexity in v can be justified by the fact that a higher v increases both the probability of being caught the intensity of the punishment. Solving this model, there is now a region in which informativeness is *increasing* in stakes. This ambiguity across models further motivates our empirical tests.

the indifference condition requires that $\frac{c}{vs\bar{\eta}} \in (\pi, 1)$

$$\Rightarrow v^* = \frac{c}{s\bar{\eta}} < v < \frac{c}{s\bar{\eta}\pi} = v^{**}$$

So for $v \leq v^*$, informativeness does not vary with v, but for $v \in (v^*, v^{**})$, informativeness is falling in v and s.

[Prop 1.3] For $v > v^{**}$ we have $v > \frac{c}{s\bar{\eta}\pi}$. There is no separating equilibrium. In a separating equilibrium low types would want to deviate given that $b(v,s) = vs\bar{\eta} > \frac{c}{\pi} > c$. There is no partial pooling equilibrium, as shown above. Assume a pooling equilibrium in which both types tweet. If the off-equilibrium path belief $E[\eta \mid a = 0]$ is low enough to guarantee that the low-support types are still willing to tweet (i.e., $E[\eta \mid a = 1] - E[\eta \mid a = 0] > \frac{c}{vs}$), then both types tweet. A pooling equilibrium exists.

C Deviations from the Pre-Analysis Plan

We posted our pre-analysis plan to the AEA registry on May 12, 2022, prior to the launch of the audit experiment. We updated the pre-analysis plan on May 24th to explain the detection-related logic for stopping the experiment before sending emails to the full set of 18,514 academics. We also explained a minor issue where we addressed a handful of emails to the wrong professors, leading us to drop 23 academics from the audit sample. We updated the pre-registration on February 1, 2024 to add a description of the donation study. We list deviations from the pre-analysis plan here:

- In the pre-analysis plan we briefly described the graduate student survey, signaling that we hoped to collect a "third-party report of the behavior of academics at their institution." Given low response rates to our student survey in piloting, we decided not to ask for third-party reports on professors, helping us to keep the survey to roughly 10 minutes.

- In the pre-analysis plan we signalled that we would omit the analysis of gender discrimination to use instead in a companion paper. We have followed the advice of a previous editor in adding the gender analysis to the current paper.

- Also following editorial advice, we have added analysis in which we compare the emails sent from any minority group (Black, female, or first-generation) to those sent by White males with no mention of first-generation status. This analysis has the advantage of giving a clean measure of minority vs. non-minority treatment, as opposed to our other comparisons (e.g. male vs. female), where by design, 3/4 of the 'majority' group (i.e. male) belongs to one of the other two minority groups (i.e. Black or first-generation).

- While we calculated our key measure Vocal_i prior to running the audit experiment, we since realised that we were unintentionally using text from truncated retweets for roughly half of the academics.

For the current paper, we have updated the measures of vocality to include mentions of racial justice-related words and phrases in the full text of the retweets.

D Procedure for FEC Contributions

Summary. Bouton et al. (2022) show that the vast majority of contributions are now made through conduits (particularly ActBlue and WinRed), and that reporting requirements for these online platforms ensure that all contributions are reported, along with the full name of the donor. For linking, we make use of the fact that 98.9% of FEC-reported individual contributions also report the occupation and employer of the contributor. We first kept only the contributions that list an employer that could be one of the top-150 universities in our sample, and only those that list an occupation that could be consistent with being a research-active academic. After these steps, we carried out an exact-match on full name (allowing for nicknames) and university.

Details. To download and link FEC-reported political contributions, we follow the detailed data appendix of Bouton et al. (2022), with some adaptations to fit our context. In particular, while Bouton et al. (2022) aim to describe the donation patterns of *all* donors in the US (requiring them to assign each and every donation to a given donor), we only need to identify the donations of the academics in our dataset. This simplifies the matching process, since the occupation and employer variables in the FEC data are particularly useful for cases in which the target population all have a similar occupation.

Targeting Academics. First, we web-scraped all FEC-reported individual contributions from January 1, 2020 to March 27, 2022 from here. We carried out basic cleaning checks as in Bouton et al. (2022) – dropping duplicates and dropping those from "lines" other than 11A(i) and 17A(i) (these lines denote contributions from individuals).

Second, we used the employer variable to keep only contributions from individuals employed by the top-150 universities, allowing for abbreviations (e.g. UCSB instead of University of California, Santa Barbara), other major name variants, and common misspellings.

Third, we used the occupation variable to keep only contributions from individuals that might be research-active academics (e.g. Professor, Scientist, Historian, etc.). We then carried out basic cleaning checks of the first and last names of contributors in this smaller dataset of contributions.

Matching. To link with our dataset of academics, we looked for perfect matches on first name, last name, and university. We allowed for nickname variants of each first name using the American English Nickname Collection from the Linguistic Data Consortium at UPenn.

Contribution Characteristics. Each contribution in the data has a committee ID, though many contributions go to conduits (especially ActBlue and WinRed) that then channel the donation to a final commit-

tee. As much as possible, we identified the final committee of a donation using the memo_text and the receipt_type_full variables. We then used FEC-provided crosswalks from here to merge on the characteristics of each committee – including any linked candidates and their political parties. All of our measures of contribution type then denote features of the final committee to which the contribution is directed.[30]

To determine the political party of each contribution, we first used the political party of the connected candidates, if they exist. For the committees without connected candidates, we used online sources to establish which party the committee is primarily raising funding for.

E Email Example

Subject: Some questions about «**recipientDegree**»s

Main Body:
Dear Professor «**recipientFullName**»,

I came across your academic work since I am considering applying this fall to «**recipientDegree**»s «**recipientField**».

While I have done a fair amount of research online, I still feel quite unsure about what «**recipientDegree**» life is like exactly, and whether I would be well-suited for it. «**This is probably partly due to me being a first generation college student.**»

Though I am sure you are very busy, would you have any spare time in the next week or so to answer some of my questions over a call?

I would be very grateful for the help, though I understand if it is not possible.

Thanks, «**senderName**»

F Further Audit Experiment Details

Given space constraints, we omitted minor details on the audit experiment from the main paper. We cover those details in this appendix section.

Sampling. We omitted one extremely minor sampling detail in the main text: we also dropped a handful of academics that we had discussed the project with from the audit sample.

[30] As explained in Bouton et al. (2022), this also requires us to drop duplicate contributions in cases where the conduit and the final committee both reported the same contribution.

University and Department. For some of the analysis, we use features of the university and department of each academic. For each of the universities, we linked the Black or African American share of undergraduates in Fall 2020 from the National Center for Education Statistics. The team standardized the coded department of each academic, assigning each academic to one of seven broad categories (e.g. Social Sciences), and to one of 75 narrow categories (e.g. Economics).

Identifying Twitter accounts. We used a search engine with automated searches to create a shortlist of possible Twitter handles for each academic. Research assistants manually picked the correct handle from the shortlist. If an academic's handle was not shortlisted, the research assistant conducted a manual search for the correct handle.

Selecting distinctive names. For first names, we used data on baby names from New York City and Massachusetts. For the data for New York City, see here. We received the Massachusetts data from the Massachusetts Registry of Vital Records and Statistics. We kept only those with birth years from 1995 to 2004, making the individuals around late-college age today. We dropped distinctively Jewish and Italian names, any first names used in Bertrand and Mullainathan (2004) (since these names may be distinctively fictitious-sounding to some academics), and the eight first names used in our pilot experiment.[31] We imposed 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 kept the top-36 most distinctive for each gender-race cell. For example, for White men, we kept 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 used the 2010 US Census, and a within-race popularity cutoff of 0.1% (exactly as in Kessler et al. 2019). We again dropped the eight last names used in our pilot.[32] We kept the 72 most racially distinctive last names – 36 for Black last names, 36 for White.

In the final step, we randomly matched each first name to a last name, with each last name used twice – once each for a distinctively male and female name. This leaves us with 144 full names. To select the 120 most racially distinctive names from among these, we paid MTurkers to guess the race of the names. We dropped 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.

Email addresses. Stratifying by race and gender, 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 X we used a protocol that ensures that the number of digits in X is balanced between distinctively Black and White names. In particular, we first randomly paired each full name with a different full name from the same gender but different race. We then found the lowest X such that a gmail account with that

[31] Iyanna, Tyra, Latrell, Tyreek, Jaclyn, Molly, Graham, and Jonah.
[32] Washington, Glover, Ware, Clay, Collins, Peterson, Ward, and Phillips.

X was available for both the Black and White full name in a given pair. We then randomly picked two numbers above that number, with the same number of digits, and assigned one to the Black name and one to the White name.

Stratification. For the randomization to Black vs. White name of sender, we stratified on university-by-department. For the randomization to male vs. female name of sender, we stratified on university-by-department-by race of sender. For the first generation status sentence, we stratified on university-by-department.

Minor Randomization Details. For randomization stratified on university-by-department, we made 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).[33]

We split the universities into nine groups according to the final exam dates for the last semester. We emailed the academics according to this order, with the email order within each of the nine groups randomized.

Since most of our email types mention the undergraduate institution of the fictitious sender, we assigned this institution randomly at the level of the sender-by-university-by-department. For the set of possible institutions, we started with the same top-150 US News ranked institutions as for our sample of academics. We then used NCES data from Fall 2020 to keep the 90 institutions that satisfy these eligibility criteria: (i) at 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 covered by our audit sample, we kept the eight of the 90 institutions that are closest in rank to be considered as the institution of the fictitious student.

Coding E-mails. We assigned each email coder to the same number of White and Black email accounts, ensuring that email coder fixed effects are orthogonal to the race of the fictitious student.

Ethics. We received full ethics approval for the audit experiment from UBC's Behavioural Research Ethics Board. The experiment involves deception. We opted against the non-deceptive incentivized resume rating of Kessler et al. (2019) for sampling-related reasons. While Kessler et al. (2019) partnered with the career services of two universities, recruiting 158 employers, our interest is in understanding the informativeness of social media for academics as a whole. To answer this question using incentivized resume rating, we would need to (i) recruit academics across fields and schools to review CVs, without strong selection into participation, (ii) recruit sufficiently many academics interested in screening CVs

[33] Since we ultimately only emailed 11,450 of the 18,514 academics, we have some singleton strata in the analysis sample. Whenever our analysis includes strata fixed effects, these singleton observations are dropped, leaving us with 11,393 observations.

to have statistical power to detect *differences* in discrimination, and not only levels, and (iii) credibly promise to use the elicited preferences to match students to academics (which would require data on a large number of real students interested in all types of graduate school). We were unable to design a feasible strategy to achieve all three of these goals. In addition, even with incentivized resume rating, a remaining element of deception would likely still be required: one would not want to reveal that each academic's choices would be linked with their public tweets.

We took several steps to minimize ethical concerns. First, to reduce the burden on academics, we sent the emails during May when most research academics are not teaching. Second, since their participation is not crucial for answering our research questions, we excluded Black academics from the experiment entirely to avoid imposing unnecessary hassle costs.[34] Third, whenever an academic accepted a meeting invite, we sent emails manually to cancel Zoom meetings promptly and politely. Fourth, we did not debrief academics on the fictitious nature of the email after the experiment ended, to reduce the possibility that academics become more suspicious of future emails from genuine students.

To the concern of poisoning the well, we note that correspondence studies with US-based professors are rare – in particular, a recent meta-analysis of correspondence studies measuring racial discrimination in the US since 2000 found only one study with professors – Milkman et al. (2012) – carried out over ten years ago (Gaddis et al. 2021). Even with these efforts, the moral case for our audit experiment rests on the benefits of the study's results outweighing the costs. Here, our view is that the incremental knowledge from our paper is substantial. In particular, when we ran the audit experiment, we were not aware of any well-identified large-scale measures of discrimination across academia since Milkman et al. (2012), or of any measures of first-generation student discrimination. More importantly, we are aware of no evidence of the usefulness of social media for predicting who discriminates. Each of these findings is decision-relevant for students.

G High Stakes Behaviors: Data Description

Sample. For the exercise using high stakes behavioral outcomes of academics, we restrict our sample of academics to On-Twitter Non-Black academics. We describe here the relevant data sources, and the main data cleaning steps.

Rate My Professor. For each of the US News top-150 universities in our sample, we manually search the school identifier associated with the university. To do so, we search for each university by name in `www.ratemyprofessors.com` (RMP) and select the correct match from the results. The university page link contains the school identifier (`www.ratemyprofessors.com/school/[schoolidentifier]`). We then use RateMyProfessorAPI's command `get_professors_by_school_and_name` using the school identifier and the academic's name to search for each academic. From this step, we obtain an RMP page

[34] With only 1,094 Black academics satisfying the eligibility criteria, 88% of which we classify as Vocal, we would anyway have had little statistical power to estimate tweet informativeness separately for Black academics.

for each academic in our sample.

Data cleaning. We identify an RMP page for 18,488 of our 18,514 academics. Next, we check whether the RMP page truly matches the academic in our sample.

First, we consider the page to match our academic if the name in the RMP page exactly matches the name of the academic. Data inspection shows that these are reasonable matches, as the academic's field in RMP and in our records is the same or within the same area.

Second, we manually select cases that have the same university and same field in RMP and our records, but were not initially matched due to small differences in name spelling (e.g. John P. Smith versus John Patrick Smith).

Third, we manually select the good matches from among the cases that have the same university and high similitude scores in name (i.e. greater than 0.75 using STATA's `matchit` command). We use Google search and closer inspection of the RMP page information to determine matches if necessary.

Finally, we drop academics that we could not match to a correct RMP page or those that had no ratings. This leaves us with 9,197 academics in our analysis sample.

Outcome measures. Through the data collection process above, we obtain the following two outcome measures for each academic:

1. Teacher rating: Average answer to question "Rate your professor" on a scale from 1-5.
2. Percent that would take again: Percentage of raters that answered "Yes" to "Would you take this professor again?"

We also use the number of student ratings underlying these two outcomes as a control variable.

Left Twitter. In late-March 2025, we used `nitter.net` to check the current status of the Twitter accounts of the 18,514 tweeting academics. This results in the profile being (i) found and public, (ii) found and protected, (iii) found but having no posts, (iv) found but suspended, or (v) not found. For (i)-(iii) we also obtain the month and year at which they joined Twitter.

Outcome measures. We define the outcome Left Twitter as a dummy variable equal to one if the academic's handle was not found ((v) above), or found but with a join date that differs from what we had in our records (n = 262), as this reflects that the account was closed and then a new account was created (the difference in between the two dates tends to be over one year). We have checked some of those cases and the new accounts tend not to belong to academics. Of the 18,514 academics, we code 3,858 academics (21%) as having left Twitter.

OpenAlex Data (Topics and Co-authors). OpenAlex is open access, the successor to Microsoft Academic Graph,[35] and has comparable coverage to Web of Science (as used in Lerner et al. (2024)) and Scopus (Culbert et al. 2025). The dataset includes over 240 million works, including working papers (e.g. those on SSRN).

We start by attempting to match each individual in our sample of academics with an OpenAlex author id (N=28k; we start with the complete sample of academics merely to improve match quality in the 18.5k tweeting-academic sample of interest). We first made requests to OpenAlex's API searching for our tweeting academics by name and university. The API returned at least one match for 26k of the tweeting academics. For 20k academics, the search returned a single potential match. To pick the best match for each academic in our sample and weed out bad matches, we constructed a match score. For each possible match, we calculate a match score based on:

1. String distance between the name of the academics in each dataset.
2. Similarity between the academic's department name and the research topics that OpenAlex associates with that academic. The similarity measure includes both
    - An indicator for whether the topics or subfields associated with an author in OpenAlex contain that academic's department name.
    - Cosine similarity (using the all-MiniLM-L6-v2 embedding) between topics plus subfields (concatenated) and department name.
3. OpenAlex's own "relevance score" for the search result, i.e., an indicator suggesting how likely it is that the search result corresponds to the person searched. More prolific authors tend to have a higher relevance score. If an academic has two profiles in OpenAlex, this will ensure that we select the profile that is better-populated.
4. Whether the first and last year for which a profile has a scientific work in OpenAlex profile is at odds with the academic's position (e.g., we penalize a potential match if its first work is in 2018 but the academic was supposed to be a full professor by 2020).
5. Whether the number of works for a profile in OpenAlex is inconsistent with the academic's position (e.g., we penalize a potential match if the OpenAlex profile has publications below the 10th percentile of the number of publications of authors in that field when that author is supposed to be a full professor by 2020).

We impose that an OpenAlex profile can only match with a single academic in our sample, and then drop academics that either (i) did not meet the inclusion criteria for our audit experiment or (ii) had a match score below the 4th percentile across all matches (this threshold defined based on a manual inspection of

[35] Microsoft Academic Graph was a project by Microsoft Research to catalog all scholarly works on the internet.

matches around the cut-off). We are then left with 16,767 academics, meaning that we could find good matches on OpenAlex for 90.4% of the relevant set of tweeting academics.

Encouragingly, when we manually inspect 50 random matches, we find that 100% correspond to the intended academic.

Measuring Black Co-authors. The list of co-authors for each academic is based on all works that the academic was linked to on OpenAlex from 2020 to 2022. We dropped 966 academics from our sample who had 500 or more co-authors over those three years, since a large share of those co-authorship relationships should be very shallow. So the final sample size is $16,767 - 966 = 15,801$.

Once we have the list of co-authors for each academic, we obtain all the names across the lists and clean the names. First, we remove single letters from names (e.g. "John P Smith") and keep names with at least two parts (e.g. we remove names that are a first name only, or a last name only), since initials or single names don't provide as much information to racially code names. Second we remove strings over 43 characters as these tend not to be person names. Third, we remove names with foreign characters and separately determine their most-likely race (e.g. names with Chinese characters coded as East Asian). Finally, we keep unique names to avoid racially coding names more than once.

For each name in our list of clean names, we ask gpt-4o to determine the most likely race associated with each co-author with the following prompt:

> What is the most likely race of the name NAME? Please use ONLY the following seven categories: **White, Black, East Asian, South Asian, Hispanic, Other, Uncertain**. If the name is associated with multiple races, list all likely races in order of likelihood (from most likely to less likely), separated by commas. Again, use ONLY the seven races provided in the list. Consider cultural origins and likelihood to guide which races are most likely associated with the name. In your response, list only the race(s), in order of likelihood, separated by commas.

The output has the name and the race or races assigned by gpt-4o. For each name, we keep the first race assigned. Then, for each eligible academic we construct the list of races associated with their co-authors' names. We then construct the following outcomes:

1. Black Co-author (%): Percentage of co-author names that were coded as "Black".

2. Black Co-author: Dummy variable indicating that the academic has at least one co-author whose name was coded as Black.

Measuring Research on Race-Related Topics. The list of topics for each academic is based on works listed on OpenAlex from January 2000 to March 2025. 95% of all works in that period have been assigned a primary topic by OpenAlex. 14 academics have no works or no works with an assigned primary topic in the period, so we exclude them from the analysis (leaving N=16,753).

We identified OpenAlex topics related to racial issues using a two-step process. First, we use gpt-4o with the following prompt:

> OpenAlex describes papers in the topic of {topic} as: {description}. Would you say that this cluster of papers is in most part related to the study of race, ethnicity, or race-related social justice? Answer with "Yes" if you believe that the topic is mostly about one or more of those topics, and "No" otherwise.

With that prompt, gpt-4o classifies 112 of 4,516 topics as being about racial or ethnic issues. Then we manually review those topics to identify those that are likely related to racial justice issues in the US. For instance, we classified a topic related to Christian-Jewish relations in post-war East Germany as not closely related to the current racial justice debate in the US. We also include two topics related to racial discrimination that gpt-4o missed (related to non-discrimination law in Europe and diversity-related research). We then construct the following outcomes:

1. Work on Racial Topic: Academic has at least one work on a racial topic, based on the authors' classification of topics.
2. Work on Racial Topic - AI: Academic has at least one work on a racial topic, based on gpt-4o's classification of topics.

H Graduate Student Survey: Sampling and Details

We used graduate program websites to collect the email addresses of doctoral students at the top-80 universities as per the US News Rankings of 2019. We opted for the top-80 universities rather than the full top-150 because of research assistance capacity constraints. To oversample Black students, we collected all email addresses of graduate students with photographs where the team judged the student to be likely to self-identify as Black ($N = 3,502$). Though for our analysis we use the self-identification of each student respondent to measure race and ethnicity. Otherwise, we randomly sampled three students per doctoral program, provided email addresses were available ($N = 7,337$). This number is not divisible by three because some doctoral programs had only one or two students.

Incentives. We offered each student a $5 Amazon gift voucher and a chance to win one of ten $100 cash prizes for taking the survey. In addition, we randomly assigned half of the students to receive prediction incentives. After each prediction question these students would read "You will get one additional lottery ticket for a $250 cash prize if your answer is within 3 percentage points of the number we found." This approach is incentive-compatible for eliciting the mode of each respondent's subjective belief distribution (Haaland et al. forthcoming). A drawback of monetary incentives is that respondents may bake bias into their reports – reporting not what they believe to be the truth, but what they believe a ivory-tower academic to find. Given this, we opted to randomly assign prediction incentives, rather than to incentivize all predictions.

Unconditional vs. Conditional Differences. In the paper, we present only the graduate students' predictions of unconditional differences in behavior between Vocal and Silent academics. In the survey, we also asked students to report their prediction of the *conditional* difference, after they had reported their prediction of the unconditional difference. We used the following text: "...suppose you know of two professors of the same rank in the same department and university. They also share the same gender and race/ethnicity and tweet the same amount. But one of the professors tweeted about racial justice in the past two years and the other did not. What would you expect the difference in racial discrimination to be between these two professors?" Given that this question is less straightforward, particularly for non-quantitative graduate students, we are more confident in the predictions of unconditional differences in discrimination, and thus focus on those in the paper. The results with conditional predictions are similar, and available on request.

I Graduate Student Survey: Prediction Questions

Overall Discrimination Prediction:

We ran an experiment in May to measure racial discrimination in academia. **We sent emails from fictitious students to roughly 11,000 non-Black academics** at top-150 US universities. As we wanted to see how Twitter activity predicts email responses, **we included only academics with Twitter accounts**. Each academic received one email, and **each email requested a Zoom meeting to discuss the possibility of graduate studies**.

Half of the emails were from typically White-sounding names like Owen Wood and Helena Bennett. The remaining emails were from typically Black-sounding names like Lamar Jenkins and Taliyah Williams. Emails from Black-sounding names had similar content to those from White-sounding names. This means that we can measure racial discrimination by comparing the meeting acceptance rate for the two types of names.

We found that 30.6% of meeting requests sent from White-sounding names were accepted.

What percentage of meeting requests sent from Black-sounding names would you guess were accepted?

Predicting Unconditional Differences in Discrimination:

Among the academics we emailed, 62% posted at least one racial justice-related tweet in the two years prior to the experiment. These tweets were almost always **in support of racial justice-related efforts**. We would now like you to guess the email response rates separately for those that tweeted about racial justice and those that did not.

31.3% of meeting requests sent from White-sounding names were accepted by **academics that tweeted about racial justice**.

What percentage of meeting requests sent from Black-sounding names would you guess were accepted by these academics?

29.5% of meeting requests sent from White-sounding names were accepted by **academics that did NOT tweet about racial justice.**

What percentage of meeting requests sent from Black-sounding names would you guess were accepted by these academics?