Political Analysis (2024), **32**, 445–462
doi:10.1017/pan.2024.3

PA

ARTICLE

Audit Experiments of Racial Discrimination and the Importance of Symmetry in Exposure to Cues

Thomas Leavitt[1] and Viviana Rivera-Burgos[2]

[1]Marxe School of Public and International Affairs, Baruch College, City University of New York (CUNY), New York, NY, USA; [2]Department of Political Science, Baruch College, City University of New York (CUNY), New York, NY, USA

Corresponding author: Thomas Leavitt; Email: thomas.leavitt@baruch.cuny.edu

(Received 2 June 2023; revised 19 December 2023; accepted 14 January 2024; published online 6 May 2024)

Abstract

Researchers are often interested in whether discrimination on the basis of racial cues persists above and beyond discrimination on the basis of nonracial attributes that decision makers—e.g., employers and legislators—infer from such cues. We show that existing audit experiments may be unable to parse these mechanisms because of an asymmetry in when decision makers are exposed to cues of race and additional signals intended to rule out discrimination due to other attributes. For example, email audit experiments typically cue race via the name in the email address, at which point legislators can choose to open the email, but cue other attributes in the body of the email, which decision makers can be exposed to only after opening the email. We derive the bias resulting from this asymmetry and then propose two distinct solutions for email audit experiments. The first exposes decision makers to all cues *before* the decision to open. The second crafts the email to ensure no discrimination in opening and then exposes decision makers to all cues in the body of the email *after* opening. This second solution works without measures of opening, but can be improved when researchers do measure opening, even if with error.

Keywords: experimental design; audit study; discrimination

Edited by: Jeff Gill

1. Introduction

Randomized audit experiments have emerged as an indispensable tool for detecting racial and ethnic discrimination (Bertrand and Mullainathan 2004; Daniel 1968; Gaddis 2018; Heckman 1998; Wienk *et al.* 1979). Although audit experiments can be conducted through a variety of technologies, today these experiments are implemented primarily via emails from (either real or fictitious) individuals who signal their race via the names in their email addresses (Crabtree 2018). While these experiments have shown the extent of racial discrimination in a variety of domains, researchers have increasingly endeavored to "move beyond measurement" (Butler and Crabtree 2017) by uncovering not only *whether*, but also *why* racial discrimination exists (Gaddis 2019; Pedulla 2018).

Researchers are often interested in whether discrimination on the basis of race—signaled in the names of email addresses—exists above and beyond discrimination on the basis of specific nonracial attributes that decision makers might infer from racial cues. For example, consider a seminal audit experiment in which Butler and Broockman (2011) emailed constituency service requests to state legislators from putatively white or Black constituents. Given the stark racial differences in party membership and legislators' strategic incentives to favor co-partisan constituents (Bartels 2008; Fenno 1978), differences in responsiveness to white and Black aliases could be due to legislators' inferring party membership from racial cues. A comparison of legislators' responsiveness to Black and white aliases

would not be able to isolate racial discrimination due to inferred party and racial discrimination that persists above and beyond inferred party.

A common way to rule out discrimination on the basis of nonracial attributes that decision makers infer from racial cues is to hold constant across treatment conditions additional information about such nonracial attributes. Consider again the example from Butler and Broockman (2011) in which state legislators may infer whether a constituent is a copartisan on the basis of a constituent's race. As Butler and Broockman (2011, 466) write, "[b]y holding constant the partisan preference of the letter's sender, we can see if the discrimination we observed was due to strategic partisan considerations and also determine if any residual discrimination remains that is not attributable to these considerations." This general template also typifies audit experiments that aim to parse the effects of multiple attributes (e.g., race and socioeconomic status) that racial-sounding names may directly cue (Elder and Hayes 2023; Landgrave and Weller 2022).

In this paper, we show the methodological difficulties of this common approach. These difficulties stem from an asymmetry in when decision makers are exposed to different cues. For example, in email audit experiments, the putative race of the email sender is signaled by the name in the email address (at which point decision makers can choose whether to open the email), but additional information—e.g., the political party of the sender—is signaled to decision makers only after they open the email. Therefore, in principle all decision makers can be exposed to the racial cue; however, only decision makers who open the emails can be exposed to cues of additional information, such as political party.

Although our argument is more general, we tailor it to a design that holds copartisanship constant across emails from randomly assigned Black or white aliases to legislators in order to detect legislators' racial discrimination above and beyond that which is due to inferred copartisanship. In this setting, we show that the estimand aligned with the causal contrast of interest is defined only among legislators who can be exposed to both race and copartisanship cues. We derive the bias of existing designs for this estimand and then propose two solutions that resolve the asymmetry in exposure to cues. Both solutions have trade-offs, but taken together help researchers discern the mechanisms behind—and hence solutions to—racial discrimination.

The first solution ensures that all legislators can be exposed to both cues by signaling them prior to the decision to open—e.g., in the name of the email address and in the email's subject line. However, in some cases, signaling copartisanship or other attributes before the decision to open may be unnatural. Thus, our second solution crafts emails to justify the assumption of no racial discrimination in opening. This design then exposes legislators to cues in the body of the email.

We show that without measures of opening, this second solution enables unbiased estimation of an informative lower bound (in magnitude) of the estimand aligned with the causal contrast of interest. Researchers, however, can employ standard technology to measure the opening of emails. When they do, this design enables unbiased estimation of more informative bounds under measurement error in opening and of the target estimand when there is no measurement error. We also derive a formal test that enables researchers to detect whether measurement error exists.

The immediately succeeding section provides the formal setup for the argument, including an explanation of the appropriate estimand aligned with the causal contrast of interest. The next section derives the bias of existing designs for this estimand. Section 4 lays out the first solution, and Section 5 lays out the second. The second solution begins under the setting in which researchers do not measure the opening of emails and then considers the setting when researchers do measure opening, both with and without error. Section 6 covers variance estimation and inference in the settings of each of the two proposed solutions. The ensuing discussion in Section 7 compares the two solutions in terms of statistical, substantive, and ethical considerations. The final section concludes.

2. Setup

To make matters concrete, consider an email audit experiment that consists of constituency service requests to legislators from one of two aliases, which cue either a Black (treatment) or white (control)

constituent. The treatment variable of interest is the racial cue signaled by the name of the sender. Existing research states that different names have differing levels of racial "soundingness" (Bertrand and Mullainathan 2004; Butler and Homola 2017; Fryer and Levitt 2004). For example, Bertrand and Mullainathan (2004) use the names Lakisha and Jamal to signal Black identity and use the names Emily and Greg to signal white identity.

The email to each legislator contains a randomly assigned racial cue (either a Black or white alias). All emails from both Black and white aliases include a signal of copartisanship with the legislator. The cue of race varies across emails, but the copartisanship cue is fixed. The primary outcome of interest is whether the legislator replies to the email and an intermediate outcome is whether the legislator opens the email.

More formally, let this experiment consists of a finite study population with $N \geq 4$ units and let the index $i = 1, \ldots, N$ run over these N units. In our running example, $i = 1, \ldots, N$ indexes the N legislators (or, more precisely, their email addresses), *not* the email senders. The indicator variable $z_i = 1$ or $z_i = 0$ denotes whether individual unit i is assigned to treatment $(z_i = 1)$ or control $(z_i = 0)$. We let treatment, $z_i = 1$, be assignment to a Black alias and control, $z_i = 0$, be assignment to a white alias. The vector $\boldsymbol{z} = \begin{bmatrix} z_1 & z_2 & \ldots & z_N \end{bmatrix}^\top$, where the superscript $\top$ denotes matrix transposition, is the collection of N individual treatment indicator variables. The set of treatment assignment vectors is denoted by $\{0,1\}^N$, which consists of 2^N possible assignments.

We ground causal effects in the potential outcomes conception of causality (Holland 1986; Neyman 1923; Rubin 1974), in which a potential outcomes schedule is a vector-valued function that maps the set of possible assignments to an N-dimensional vector of real numbers. For the primary outcome of interest, email replies, denote the vectors of potential outcomes by $\boldsymbol{y}(\boldsymbol{z})$ for $\boldsymbol{z} \in \{0,1\}^N$, which are the elements in the range of the potential outcomes schedule. The individual potential outcomes for unit i are the ith entries of each of the N-dimensional vectors of potential outcomes, denoted by $y_i(\boldsymbol{z})$ for $\boldsymbol{z} \in \{0,1\}^N$. That is, $y_i(\boldsymbol{z})$ is whether the ith legislator would reply to an email if the assignment of all legislators were equal to some $\boldsymbol{z} \in \{0,1\}^N$. In our setting, individual potential outcomes are binary, that is, $y_i(\boldsymbol{z}) \in \{0,1\}$ for all $\boldsymbol{z} \in \{0,1\}^N$, where $y_i(\boldsymbol{z}) = 1$ indicates a reply to the email and $y_i(\boldsymbol{z}) = 0$ indicates a lack thereof.

While legislators' replies to emails are of primary interest, another intermediate outcome of interest is whether legislators open emails. We denote these potential outcomes by $\boldsymbol{m}(\boldsymbol{z})$ for $\boldsymbol{z} \in \{0,1\}^N$. A legislator's potential opening of an email is also binary, where $m_i(\boldsymbol{z}) = 1$ indicates that the ith legislator would open the email under assignment $\boldsymbol{z} \in \{0,1\}^N$ and $m_i(\boldsymbol{z}) = 0$ indicates that the ith legislator would not open the email under assignment $\boldsymbol{z} \in \{0,1\}^N$.

With 2^N assignments, there are in principle 2^N potential outcomes for each individual unit. However, we make the stable unit treatment value assumption (SUTVA) for both primary and intermediate potential outcomes.

Assumption 1 (Stable unit treatment value assumption). *For all $i = 1, \ldots, N$ units, $y_i(\boldsymbol{z})$ and $m_i(\boldsymbol{z})$ take on fixed values, $y_i(1)$ and $m_i(1)$, for all $\boldsymbol{z} : z_i = 1$ and take on fixed values, $y_i(0)$ and $m_i(0)$, for all $\boldsymbol{z} : z_i = 0$.*

Under SUTVA, we write a potential reply to an email for unit i as $y_i(z)$, which is either $y_i(1)$ or $y_i(0)$ depending on whether $\boldsymbol{z}$ is with $z_i = 1$ or $z_i = 0$, and write a potential opening of an email for unit i as $m_i(z)$, which is either $m_i(1)$ or $m_i(0)$ depending on whether $\boldsymbol{z}$ is with $z_i = 1$ or $z_i = 0$. The same is true for intermediate variables measured post-treatment, such as opening.

Under SUTVA in Assumption 1, we can partition legislators into principal strata (Frangakis and Rubin 2002) on the basis of the intermediate outcome, whether a legislator opens the email. We define the following principal strata for an arbitrary legislator, i.

Let $s_i = s \in \{\text{AO}, \text{OWO}, \text{OBO}, \text{NO}\}$ denote the principal stratum of the ith legislator. The total number of legislators in each stratum is

$$N_s \coloneqq \sum_{i=1}^{N} \mathbb{1}\{s_i = s\},$$

Table 1. Principal strata by potential opening of emails.

$z_i = 0$	$z_i = 1$	Stratum
$m_i(0) = 1$	$m_i(1) = 1$	*Always-Open* (AO)
$m_i(0) = 1$	$m_i(1) = 0$	*Only-White-Open* (OWO)
$m_i(0) = 0$	$m_i(1) = 1$	*Only-Black-Open* (OBO)
$m_i(0) = 0$	$m_i(1) = 0$	*Never-Open* (NO)

where $\mathbb{1}\{\cdot\}$ is the indicator function. Then write the average of treated ($z = 1$) or control ($z = 0$) potential outcomes in principal stratum s as

$$\bar{y}_s(z) \coloneqq \left(\frac{1}{N_s}\right) \sum_{i=1}^{N} \mathbb{1}\{s_i = s\}\, y_i(z), \tag{1}$$

and the average treatment effect, ATE, in stratum s as $\tau_s \coloneqq \bar{y}_s(1) - \bar{y}_s(0)$. The overall ATE across all strata is

$$\begin{aligned} \tau \coloneqq\ & \pi_{\text{AO}}\left(\bar{y}_{\text{AO}}(1) - \bar{y}_{\text{AO}}(0)\right) + \pi_{\text{OWO}}\left(\bar{y}_{\text{OWO}}(1) - \bar{y}_{\text{OWO}}(0)\right) \\ & + \pi_{\text{OBO}}\left(\bar{y}_{\text{OBO}}(1) - \bar{y}_{\text{OBO}}(0)\right) + \pi_{\text{NO}}\left(\bar{y}_{\text{NO}}(1) - \bar{y}_{\text{NO}}(0)\right), \end{aligned} \tag{2}$$

where $\pi_s \coloneqq N_s/N$ with $N = N_{\text{AO}} + N_{\text{OWO}} + N_{\text{OBO}} + N_{\text{NO}}$.

Given this setup thus far, we make an additional (trivially true) assumption that legislators cannot reply to an email without first opening it.

Assumption 2 (No replying without opening). *For all* $i = 1, \ldots, N$ *units,* $y_i(1) \leq m_i(1)$ *and* $y_i(0) \leq m_i(0)$.

We follow the existing literature by taking opening and replying to mean the conscious decision of an email user to open the email and compose a response. Hence, we do not regard software that automatically opens or replies to emails as an email user's actual opening or replying to an email.

2.1. Aligning the Statistical Estimand with the Causal Contrast of Interest

In order to isolate racial discrimination that remains after accounting for inferred party, the implied contrast of interest is a cue of Black identity and copartisanship versus a cue of white identity and copartisanship. The alias of the sender is contained in the email address, which is available to legislators before opening the email; however, in existing experiments, legislators are exposed to the signal of copartisanship only after opening the email. Hence, individual effects in some principal strata in Table 1 are poorly aligned with this causal contrast of interest.

Only Always-Openers can be exposed to a cue of Black identity plus copartisanship and a cue of white identity plus copartisanship. Never-Openers cannot be exposed to the copartisanship signal; in principle, they can be exposed to only the Black or white cues, both without any cue of copartisanship. Only-Black-Openers can be exposed to either the Black and copartisanship cues or the white cue without copartisanship. Analogously, Only-White-Openers can be exposed to either the Black cue or the white and copartisanship cues. Overall, if a legislator does not open an email, then that legislator can be exposed to only the race condition (Black or white), *not* the race and party conditions.

This problem is likely to be especially acute in light of research suggesting that racial cues affect the decision to open emails (Hughes *et al.* 2020). Opening emails is what Hughes *et al.* (2020, 184) refer to as a "high volume, low-attention task"—one that is exactly of the sort for which we would expect implicit racial bias to operate (Bertrand, Chugh, and Mullainathan 2005; Devine 1989). This is especially so among legislators (or their staff) with a high workload (Andersen and Guul 2019) and potentially less-professionalized legislative offices (Landgrave and Weller 2020).

Insofar as race affects a legislator's decision to open the email, then it is difficult to know whether a reply to the email is driven by discrimination on the basis of inferred party or by racial discrimination above and beyond inferred party. For example, consider an Only-White-Opener who would reply to an email from a white alias and, as is implied by Assumption 2, would not reply to an email from a Black alias. Since this legislator can be exposed to only the Black alias, *not* the Black alias plus copartisanship cue, we cannot know if the decision not to open the email from a Black alias (and hence not to reply to that email) is due to inferred party or anti-Black discrimination above and beyond inferred party. This is an example of what Bueno de Mesquita and Tyson (2020) refer to as the "commensurability problem," that is, when the evidence produced by the empirical design does not correspond to the theoretical quantity of interest. This problem is also closely related to the notion of "phantom counterfactuals" in Slough (2023).

Given the particular design described thus far, its target estimand is the ATE among Always-Openers, formally given by

$$\tau_{\text{AO}} \coloneqq \bar{y}_{\text{AO}}(1) - \bar{y}_{\text{AO}}(0). \tag{3}$$

A focus on τ_{AO} in this experimental design does *not* imply that the overall ATE in (2) is unimportant. The overall ATE remains important insofar as one is interested in the effect that a Black versus white alias has on legislators' responsiveness. However, insofar as one explicitly aims to parse the aforementioned mechanisms of racial discrimination, the ATE among Always-Openers is the appropriate target.

2.2. The Assignment Process and Estimator of the Average Effect

The assignment process selects a single $\boldsymbol{z} \in \{0,1\}^N$ with probability $p(\boldsymbol{z})$. Hence, the treatment assignment vector is a random quantity, $\boldsymbol{Z}$, which takes on the value $\boldsymbol{z} \in \{0,1\}^N$ with probability $\Pr(\boldsymbol{Z} = \boldsymbol{z}) = p(\boldsymbol{z})$. We assume complete random assignment (CRA) in which, of the $N \geq 4$ units, $n_1 \geq 2$ are assigned to treatment and the remaining $n_0 = N - n_1 \geq 2$ are assigned to control.

Assumption 3 (Complete random assignment). *The set of allowable assignments is* $\Omega \coloneqq \{\boldsymbol{z} : p(\boldsymbol{z}) > 0\} = \{\boldsymbol{z} : \sum_{i=1}^{N} z_i = n_1\}$ *with* $n_1 \geq 2$, $n_0 \geq 2$, *and* $p(\boldsymbol{z}) = 1/\binom{N}{n_1}$ *for all* $\boldsymbol{z} \in \Omega$.

The canonical estimator in randomized audit experiments is the Difference-in-Means. The random Difference-in-Means under CRA is

$$\hat{\tau}(\boldsymbol{Z}, \boldsymbol{y}(\boldsymbol{Z})) = \left(\frac{1}{n_1}\right)\boldsymbol{Z}^{\top}\boldsymbol{y}(\boldsymbol{Z}) - \left(\frac{1}{n_0}\right)(\boldsymbol{1} - \boldsymbol{Z})^{\top}\boldsymbol{y}(\boldsymbol{Z}), \tag{4}$$

which, under Assumptions 1 and 3, is unbiased for the overall ATE. The Difference-in-Means' randomness is inherited solely from the random variable $\boldsymbol{Z}$, which has a known probability distribution, namely, that which is implied by Assumption 3 of CRA. From this source of randomness, we can analyze the properties (e.g., the bias) of the Difference-in-Means with respect to the causal target of interest.

3. Bias in Existing Designs

Proposition 1 below derives the bias of the Difference-in-Means for the estimand aligned with the causal contrast of interest, the ATE among Always-Openers in (3). All proofs are in the Appendix.

Proposition 1. *Under Assumptions 1–3, the bias of the Difference-in-Means for* τ_{AO} *is equal to* $(\pi_{AO} - 1)\tau_{AO} + \pi_{OWO}\tau_{OWO} + \pi_{OBO}\tau_{OBO}$, *that is,*

$$\mathrm{E}[\hat{\tau}(\boldsymbol{Z}, \boldsymbol{y}(\boldsymbol{Z}))] - \tau_{AO} = (\pi_{AO} - 1)\tau_{AO} + \pi_{OWO}\tau_{OWO} + \pi_{OBO}\tau_{OBO}. \tag{5}$$

As Proposition 1 shows, the bias of the Difference-in-Means depends on the proportions of units in each stratum and the average effects among Only-White-Openers and Only-Black-Openers. Among Never-Openers, the average effect is 0 under Assumption 2. A sufficient condition for the Difference-in-Means to be unbiased for τ_{AO} is that $\pi_{\text{AO}} = 1$, that is, all legislators are Always-Openers. The weaker assumption

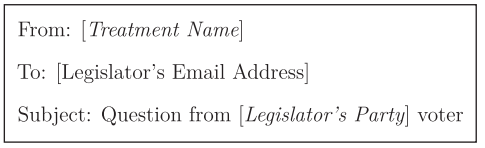

Figure 1. Sample email heading that exposes legislators to both race and co-partisan cues before opening.

that legislators do not racially discriminate in which email they open (i.e., there are no Only-White- or Only-Black-Openers) is not sufficient for unbiasedness.

4. First Solution: Exposure to Cues before Opening

Our first solution proposes cueing both race and copartisanship before opening. Properly doing so requires a notion of what email users see upon receiving an email. For example, users of `gmail` or `outlook` see the subject line of the email and a name corresponding to the email address. Some email users may also see a preview of the email's contents. One way to cue information other than race is via the email's subject line.

As an example for how one might do so, we can imagine an email with the following heading. The sample heading in Figure 1 randomly varies the treatment name to cue race and always cues the same party as the legislator's.

This example represents a relatively natural way to signal copartisanship in the subject line of the email. There are other ways to do so, and the choice will likely depend on the particular audit experiment. For example, in Butler and Broockman (2011), the email to state legislators was about how to register to vote. Hence, a subject line that conveys copartisanship for such an email might be "How to vote in [*Legislator's Party*] primary?" Alternatively, an email soliciting advice about how to get involved in political campaigns might contain the subject line "How to get involved in [*Legislator's Party*] campaigns?"

Our first solution resolves the issues described in Section 3 *not* by identifying which units are Always-Openers, but instead by realigning the ATE among all legislators with the causal contrast of interest. To reiterate, in existing designs, the ATE among Always-Openers is of interest because they are the only legislators who can be exposed to cues of both race and copartisanship. However, with a design that exposes legislators to cues *before* the decision to open, in principle all legislators (not only Always-Openers) can be exposed to cues of race and copartisanship. Thus, the overall ATE in (2) is aligned with the causal contrast of Black and copartisanship versus white and copartisanship. Since the Difference-in-Means in (4) is unbiased for the overall ATE, it follows that exposure to cues before opening restores unbiasedness of the Difference-in-Means for the causal quantity of interest.

One concern with this solution is that it may be difficult in some cases to naturally convey information other than race prior to opening. Pretesting can be a valuable way of assessing which ways of cueing attributes in an email's subject line are most natural. However, in some cases, researchers may have no reasonable means of naturally conveying attributes prior to opening. For these cases, we turn to our second proposed solution.

5. Second Solution: Exposure to Cues after Opening

Another way to achieve symmetry in exposure to cues is by exposing decision makers to both cues in the body of the email. Hence, a legislator can see both cues only after the decision to open the email. In this design, the target remains the ATE among Always-Openers, not the ATE among all legislators.

In setting up this second solution, we show how researchers can reliably draw conclusions about the ATE among Always-Openers under the assumption of no racial discrimination in legislators' decisions to open the email.

Assumption 4 (No racial discrimination in opening). *For all* $i = 1, \ldots, N$ *units,* $m_i(1) = m_i(0)$.

Because $m_i(1) = m_i(0)$ under Assumption 4, the opening of an email does not depend on treatment and, hence, is equivalent to a fixed baseline covariate. Therefore, going forward under Assumption 4, we denote the opening of the email for an individual legislator by m_i and the collection of these values for all $i = 1, \ldots, N$ legislators by $\boldsymbol{m}$, where now the dependence of opening on treatment assignment is removed.

Researchers can satisfy Assumption 4 by using email addresses and subject lines that are independent of the randomly assigned racial cue in the body of the email. Perhaps the simplest way to do so is by holding the name of the sender fixed across a legislator's assignment to treatment and control conditions. In this case, the opening of the email must be independent of treatment assignment because the information available to a legislator is identical regardless of whether a legislator is assigned to treatment or control.

One potential concern, though, is that any name fixed across racial cues may signal other attributes, e.g., gender. These other attributes cannot impact whether no racial discrimination in opening holds. However, they can impact the proportions of Always-Openers and Never-Openers if decision makers discriminate on the basis of these other attributes. To alleviate the concern that some names may lead to few Always-Openers, researchers can carefully think through (and use pretests to infer) which email addresses and subject lines are likely to yield large proportions of Always-Openers.

We now consider estimation of the ATE among Always-Openers in designs that satisfy no racial discrimination in opening. We consider three different settings that applied researchers may confront. The first setting is when researchers do not measure the opening of emails, in which case researchers can unbiasedly estimate informative bounds of the ATE among Always-Openers. The second setting is when researchers measure opening, but with error, in which case researchers can unbiasedly estimate more informative bounds of the ATE among Always-Openers. The third setting is when researchers measure opening without error, in which case it is straightforward to unbiasedly estimate the ATE among Always-Openers.

5.1. No Measures of Opening

In a setting without measures of opening, Assumption 4 of no racial discrimination in opening is important. This assumption enables us to recast the ATE among Always-Openers in terms of the overall ATE, τ, and the proportion of Always-Openers, π_{AO}.

Lemma 1. *Under Assumptions 1, 2, and 4, and supposing that the proportion of Always-Openers, π_{AO}, is greater than* 0*, the ATE among Always-Openers is*

$$\tau_{AO} = \left(\frac{1}{\pi_{AO}}\right)\tau = \left(\frac{N}{N_{AO}}\right)\left(\frac{1}{N}\right)\left[\sum_{i=1}^{N} y_i(1) - \sum_{i=1}^{N} y_i(0)\right]. \tag{6}$$

The intuition for Lemma 1 is that no racial discrimination in opening (Assumption 4) implies that all legislators are either Always-Openers or Never-Openers. Since opening an email is a necessary condition for replying to an email (Assumption 2), the sum of all legislators' potential replies to putatively Black constituents is equivalent to the same sum among Always-Openers. The same is true for the sum of all legislators' potential replies to putatively white constituents. Dividing these sums by the number of Always-Openers then yields the average effect among Always-Openers.

In practice, when researchers do not measure opening, the proportion of Always-Openers is unknown. However, before measuring replies, the lower and upper bounds of the proportion of Always-Openers, denoted by $\underline{\pi}_{\text{AO}}$ and $\overline{\pi}_{\text{AO}}$, are known. That is, supposing that the number of Always-Openers is greater than 0, the proportion of Always-Openers can take on values in the space given by $\{1/N, 2/N, \ldots, 1\}$ with lower and upper bounds given by

$$\underline{\pi}_{\text{AO}} = \frac{1}{N}, \tag{7}$$

$$\overline{\pi}_{\text{AO}} = 1. \tag{8}$$

Since Assumption 2 implies that every reply to an email must be from an Always-Opener, the proportion of Always-Openers' lower bound will typically be greater upon measuring replies. The lowest proportion of Always-Openers consistent with the observed data is the proportion of all legislators (either treated or untreated) who reply to the email. We therefore recast the lower bound of the proportion of Always-Openers given observed data as

$$\underline{\pi}_{\text{AO}} = \frac{1}{N}\sum_{i=1}^{N} z_i y_i(1) + (1-z_i) y_i(0). \tag{9}$$

Inspection of (6) reveals that, holding the overall ATE fixed, the magnitude of the ATE among Always-Openers is decreasing in the proportion of Always-Openers. When the ATE among Always-Openers is negative, the upper bound of the proportion of Always-Openers implies an upper bound of the ATE among Always-Openers. Conversely, when the ATE among Always-Openers is positive, the upper bound of the proportion of Always-Openers implies a lower-bound of the ATE among Always-Openers. That is, the magnitude of the ATE among Always-Openers is always greater than the magnitude of the ATE among all legislators. Proposition 2 formally establishes this point.

Proposition 2. *Suppose Assumptions 1, 2, and 4, and that the proportion of Always-Openers, π_{AO}, is greater than* 0. *It follows that:*

- *If $\tau_{AO} < 0$, then $\overline{\tau}_{AO} = \tau$.*
- *If $\tau_{AO} > 0$, then $\underline{\tau}_{AO} = \tau$.*
- *If $\tau_{AO} = 0$, then $\tau_{AO} = \underline{\tau}_{AO} = \overline{\tau}_{AO} = \tau$.*

Proposition 2 carries an important implication. The lower bound in magnitude of the ATE among Always-Openers is always equal to the ATE among all legislators. Since this ATE among all legislators can be unbiasedly estimated via the Difference-in-Means in (4), researchers can draw informative conclusions about the ATE among Always-Openers without knowledge of the proportion of Always-Openers. For example, suppose that a researcher rejects the null hypothesis of no racial discrimination, on average, among all legislators in favor of the alternative of anti-Black discrimination, on average, among all legislators. This implies the rejection of the same null (in favor of the same alternative) among specifically Always-Openers.

One potential concern is that, if there are few Always-Openers, then the power to detect the lower bound in magnitude of the ATE among Always-Openers may be small. Put differently, crafting an audit experiment to satisfy no racial discrimination in opening could make the overall ATE smaller by turning Only-Black- or Only-White-Openers into Never-Openers, among whom the individual effects are all equal to 0 under Assumption 2. Hence, power to detect the overall ATE (equivalently, the lower bound in magnitude of the ATE among Always-Openers) may decrease.

To alleviate this concern, researchers can craft emails not only to satisfy no racial discrimination in opening but also to maximize the opening rate among all legislators. The intuition for doing so is to maximize the proportion of Always-Openers among legislators who would have been Only-Black- or Only-White-Openers in existing designs that fail to satisfy no racial discrimination in opening. Ensuring that Always-Openers in existing designs remain Always-Openers in designs satisfying no racial discrimination in opening is important, too.

Thus far, Proposition 2 has implied that the canonical Difference-in-Means, which effectively supposes that the proportion of Always-Openers is equal to 1, is unbiased for a lower bound (in magnitude) of the ATE among Always-Openers. However, if researchers were to have access to the proportion of Always-Openers, then an unbiased estimator of τ_{AO} would be

$$\left(\frac{1}{\pi_{\text{AO}}}\right)\hat{\tau}(\mathbf{Z}, \boldsymbol{y}(\mathbf{Z})) = \left(\frac{1}{\pi_{\text{AO}}}\right)\left[\left(\frac{1}{n_1}\right)\mathbf{Z}^{\top}\boldsymbol{y}(\mathbf{Z}) - \left(\frac{1}{n_0}\right)(\mathbf{1}-\mathbf{Z})^{\top}\boldsymbol{y}(\mathbf{Z})\right]. \tag{10}$$

Proposition 3 formally establishes this unbiasedness.

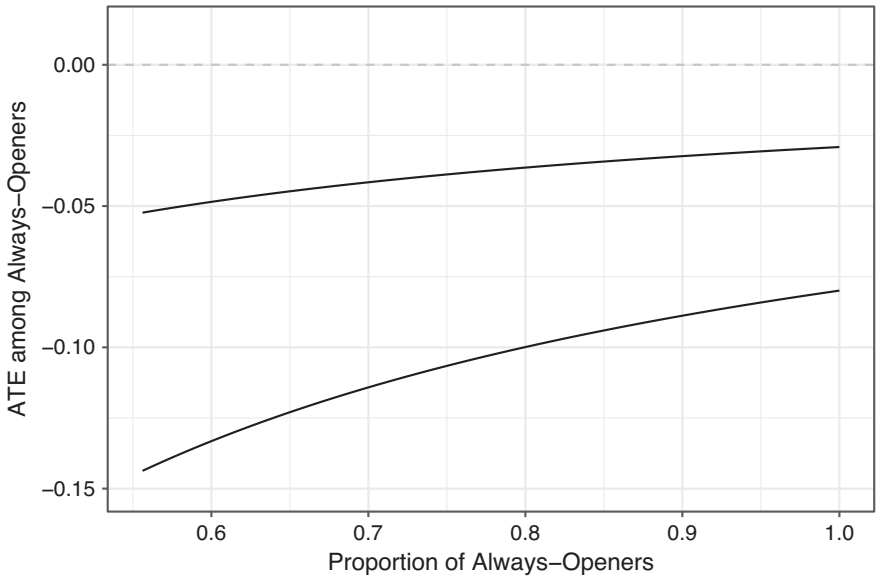

Figure 2. Ninety-five percent confidence intervals over decreasing values of the proportion of Always-Openers.

Proposition 3. *Under Assumptions 1–4 and a known value of π_{AO}, the expectation of the estimator in (10) is equal to the ATE among Always-Openers in (3), that is,*

$$\mathrm{E}\left[\left(\frac{1}{\pi_{AO}}\right)\hat{\tau}\left(\boldsymbol{Z},\boldsymbol{y}(\boldsymbol{Z})\right)\right] = \tau_{AO}.$$

In practice, with no measures of opening, the proportion of Always-Openers is unknown. Nevertheless, Proposition 3 points to a useful strategy. Researchers can begin by estimating and testing hypotheses about the unknown value of the ATE among Always-Openers under the assumption that all legislators are Always-Openers. Then researchers can assess how inferences about the ATE among Always-Openers would change under increasingly smaller proportions of Always-Openers.

This sensitivity analysis has several important features. As Proposition 2 foreshadows, the "best guess" (point estimate) about the ATE among Always-Openers will change in magnitude over decreasing proportions of Always-Openers, but its sign will not. How much the "best guess" changes in magnitude depends on the width of the bounds on the proportion of Always-Openers. Tighter bounds on the proportion of Always-Openers yield tighter bounds on the ATE among Always-Openers and estimates thereof.

In addition, suppose that a researcher constructs a 95 percent confidence interval (CI) by taking the point estimate plus or minus 1.96 times the estimated standard error. If this CI excludes 0 under the assumption that all legislators are Always-Openers, then the CIs under decreasing proportions of Always-Openers will also exclude 0. This property holds even though the overall width of the CIs will increase as the proportion of Always-Openers decreases.

To give an example, roughly patterned on existing audit experiments, suppose there are $N = 6{,}000$ legislators who are randomly assigned an email from a Black alias (treatment) or a white alias (control) (Leavitt and Rivera-Burgos 2024). The alias is cued in the body of the email, and there is no racial discrimination in opening. Suppose that the estimated ATE among all legislators is -0.05 with an estimated standard error of 0.01, implying a 95% CI of $[-0.08, -0.03]$. Suppose further that the proportion of legislators who replied to emails is 0.55, implying that the proportion of Always-Openers is bounded by 0.55 and 1. Figure 2 shows that the narrowest CI is the initial CI under the assumption that all legislators are Always-Openers. As the assumed proportion of Always-Openers decreases, the

width of the CI increases, but both the lower and upperbounds of the CI remain negative, never bracketing 0.

In short, this proposed sensitivity analysis is invaluable for assessing how the magnitude of estimated effects changes under different assumptions about the proportion of Always-Openers. A crucial feature, though, is that detection of racial discrimination among all legislators suffices for the detection of racial discrimination among Always-Openers. Valid inferences can still be had in the absence of precise knowledge about the proportion of Always-Openers.

5.2. Measures of Opening

Technology to measure the opening of emails is standard in email automation platforms, such as `Overloop`, and has been employed in existing audit experiments (e.g., Hughes *et al.* 2020). This technology measures opening by including in emails a tracking pixel—a tiny, invisible image in the body of an email. Email automation platforms then track when this image is downloaded, which is how one knows whether an email is opened.

Given this technology to measure opening and under Assumption 4, there are two crucial ways in which measurement error may exist.

1. False positive (hidden Never-Opener): If an email address has a security tool that scans an incoming email (and therefore downloads the tracking pixel), then an email can be recorded as opened when the user did not in fact open the email.
2. False negative (hidden Always-Opener): If an email address has a security tool that blocks open-tracking, then an email can be recorded as not opened when the user did in fact open the email.

Fortunately, under a mild assumption, email automation platforms can detect whether an email marked as opened is a false positive. If an email is opened due to a security tool that scans the incoming email, then that email will be marked as opened at the exact same minute in which the email was sent. Hence, for emails opened at the same minute in which they were sent, it is safe to regard these emails as not yet opened and potentially never opened.

If a user who has this security tool ends up genuinely opening the email (and downloading the tracking pixel again), researchers will be able to observe this event from software platforms' contact logs. Therefore, under Assumption 4 of no racial discrimination in opening, it is straightforward to discern which legislators are measured as opening the email but are in fact Never-Openers. These legislators are those who are recorded as opening the email at the same time in which the email was sent and as only opening the email once.

In light of this discussion, we now introduce the assumption of no false positives. In doing so, we let $\tilde{m}_i = 1$ or $\tilde{m}_i = 0$ denote whether an individual user of an email account is measured as opening the email (i.e., whether the user downloads the tracking pixel). Analogously, $\tilde{\boldsymbol{m}}$ denotes the collection of all $i = 1, \ldots, N$ values of $\tilde{m}_i$. Whether a legislator blocks open tracking is presumably determined prior to receiving an email from a Black or white alias; hence, $\tilde{\boldsymbol{m}}$ is essentially a baseline covariate, fixed over different possible assignments.

Assumption 5 (No false positives). *For all* $i = 1, \ldots, N$ *units,* $\tilde{m}_i \leq m_i$.

Assumption 5 implies that the concern with measurement error boils down to the possible existence of hidden Always-Openers, that is, legislators who are not measured as opening their emails even though they in fact did. In what follows, we propose a formal test that is able to reliably detect whether hidden Always-Openers exist. Depending on the results of this test, a researcher can conduct analyses either with the knowledge that measurement error exists (see Section 5.2.2) or under the assumption of no measurement error (see Section 5.2.3).

5.2.1. Test for Measurement Error in Opening

Given Assumption 2 of no replying without opening, the most straightforward way to test for the presence of false negatives (i.e., hidden Always-Openers) is by assessing whether there are any emails with replies that were also marked as unopened. More precisely, let $\varphi(\boldsymbol{y}(\boldsymbol{Z}), \tilde{\boldsymbol{m}})$ take the value 1 if the test rejects the hypothesis of no measurement error in opening and 0 otherwise, where

$$\varphi(\boldsymbol{y}(\boldsymbol{Z}), \tilde{\boldsymbol{m}}) = \mathbb{1}\left\{\sum_{i=1}^{N} \mathbb{1}\left\{Z_i y_i(1) + (1 - Z_i) y_i(0) = 1, \tilde{m}_i = 0\right\} > 0\right\}. \tag{11}$$

In principle, this test in (11) is susceptible to two types of errors, rejecting the null hypothesis of no measurement error when it is true (Type I error) and failing to reject no measurement error when it is false. Assumptions 2 and 5 imply that the probability of making a Type I error is 0. Anytime one rejects the null of no false negatives, this null must be false.

It is theoretically possible, however, to commit a Type II error, but only under implausible configurations of potential outcomes. If there is at least one legislator who blocks open tracking, but would reply to both a Black and white alias, then the probability of committing a Type II error is 0. Otherwise, if there are legislators who block open tracking, but would respond to only a Black or only a white alias, then the Type II error will be exactly equal to the assignment that happens to put all Only-Black-Repliers who block open tracking in the control condition and all Only-White-Repliers who block open tracking in the treatment condition. It is difficult to envision situations in which this probability would not be minuscule.

This plausibly low probability of a Type II error (or, equivalently, the high power) of the test has important implications. The high power of the test (equal to 1 under a plausible condition) implies that a failure to reject the null of no measurement error provides strong evidence in favor of that null. Hence, it is reasonable to proceed with an analysis under the assumption of no measurement error or not depending on the results of this test. We first consider the case in which the test rejects no measurement error in opening and then the case in which the test does not.

5.2.2. Measurement Error in Opening

Because the Type I error probability of the test of no measurement error is 0, rejecting the null of no measurement error implies (without concomitant uncertainty) that measurement error exists. The degree of measurement error (i.e., the number of measured non-openers who are in fact Always-Openers) is unknown, which implies that the proportion of Always-Openers is also unknown. Nevertheless, measures of opening with error can be used to bound the proportion of Always-Openers. While the upper bound on the proportion of Always-Openers remains the same as in (8), the lower bound can be much greater. Consequently, researchers can estimate tighter bounds on the ATE among Always-Openers.

Under Assumption 5 and before measuring replies, the lower and upper bounds of the proportion of Always-Openers are

$$\underline{\pi}_{\text{AO}} = \left(\frac{1}{N}\right)\sum_{i=1}^{N} \tilde{m}_i, \tag{12}$$

$$\overline{\pi}_{\text{AO}} = 1. \tag{13}$$

The lower bound is when the number of Always-Openers is equal to the number of measured openers and all legislators measured as not opening the email are Never-Openers. The upper-bound corresponds to the case of maximum measurement error under Assumption 5, whereby all legislators measured as not opening the email are Always-Openers.

After observing replies, Assumption 2 implies that the lower bound of the proportion of Always-Openers can be increased further. Since replying to an email implies having opened it, we can express $\underline{\pi}_{\text{AO}}$ as

$$\underline{\pi}_{\text{AO}} = \left(\frac{1}{N}\right)\left[\sum_{i=1}^{N}\tilde{m}_i + \sum_{i=1}^{N}(1-\tilde{m}_i)\left(z_i y_i(1) + (1-z_i)y_i(0)\right)\right]. \tag{14}$$

These bounds on the proportion of Always-Openers then imply bounds on the ATE among Always-Openers by plugging in $\underline{\pi}_{\text{AO}}$ and $\overline{\pi}_{\text{AO}}$ for π_{AO} in the expression for τ_{AO} in (6). Analogously, researchers can assess sensitivity to measurement error by conducting their analyses with the estimator in (10) under assumptions about the proportion of Always-Openers ranging from the lower bound in (14) to 1.

5.2.3. No Measurement Error in Opening

Thus far, we have only made the assumption of no false positives in measures of opening. We now make the assumption of no measurement error (no false negatives and no false positives).

Assumption 6 (No measurement error of opening). *For all* $i = 1, \ldots, N$ *units,* $\tilde{m}_i = m_i$.

Under Assumption 4, the absence of measurement error in opening implies that the proportion of Always-Openers can be deduced directly from data. That is, the proportion of Always-Openers is

$$\pi_{\text{AO}} = \left(\frac{1}{N}\right)\sum_{i=1}^{N}\tilde{m}_i. \tag{15}$$

With this known value of π_{AO}, the estimator in (10) can be used to unbiasedly estimate the ATE among Always-Openers.

However, with no measurement error in opening, we know not only how many legislators are Always-Openers, but also exactly which legislators are Always-Openers. This greater information allows for the use of a post-stratified Difference-in-Means among legislators who are measured as opening. We write this Difference-in-Means conditional on opening as

$$\hat{\tau}^{\text{Open}}(\boldsymbol{Z}, \tilde{\boldsymbol{m}}, \boldsymbol{y}(\boldsymbol{Z})) = \left(\frac{1}{\boldsymbol{Z}^\top \tilde{\boldsymbol{m}}}\right)\boldsymbol{Z}^\top(\tilde{\boldsymbol{m}} \odot \boldsymbol{y}(\boldsymbol{Z})) - \left(\frac{1}{(\boldsymbol{1}-\boldsymbol{Z})^\top \tilde{\boldsymbol{m}}}\right)(\boldsymbol{1}-\boldsymbol{Z})^\top(\tilde{\boldsymbol{m}} \odot \boldsymbol{y}(\boldsymbol{Z})), \tag{16}$$

where $\odot$ is the element-wise product of two matrices of the same dimension that returns another matrix with the same dimension. Like the scaled Difference-in-Means in (10), the post-stratified Difference-in-Means in (16) is unbiased for the ATE among Always-Openers.

Proposition 4. *Under Assumptions 1–4 and 6, the expected value of the estimator in (16) is equal to the ATE among Always-Openers in (3), that is,*

$$\mathrm{E}\left[\hat{\tau}^{Open}(\boldsymbol{Z}, \tilde{\boldsymbol{m}}, \boldsymbol{y}(\boldsymbol{Z}))\right] = \tau_{AO}.$$

Proposition 4 depends crucially on Assumptions 4 and 6. The former assumption implies that opening does not vary over assignments and the latter implies that we can observe exactly which legislators opened their emails.

To be sure, researchers can use either the scaled Difference-in-Means in (10) or the post-stratified Difference-in-Means in (16), both of which are unbiased for the ATE among Always-Openers, although their variances (and estimates thereof) may differ. One benefit, though, of the estimator in (16) is that it will only generate estimates of the ATE among Always-Openers between −1 and 1, that is, values that are within the natural bounds of the parameter space. By contrast, because the estimator in (10) divides the Difference-in-Means (also producing estimates between −1 and 1) by the proportion of Always-Openers, some estimates may be outside the parameter space's natural bounds.

6. Variance Estimation and Inference

Thus far, our arguments have focused on estimation in different scenarios. The first is when a researcher cues attributes prior to the decision to open. The others are when a researcher cues attributes after

opening, either without or with measures of opening. In all of these scenarios, variance estimation and inference is straightforward.

We first consider the scenario in which a researcher cues attributes before opening. In this case, the canonical Difference-in-Means is unbiased for the overall ATE, which, because of the design, is the relevant causal target. Neyman (1923) derived the variance of this estimator (see also Imbens and Rubin 2015, 87–92), denoted by $\text{Var}\left[\hat{\tau}\left(\boldsymbol{Z},\boldsymbol{y}(\boldsymbol{Z})\right)\right]$, and a canonical conservative estimator for this variance, denoted by $\widehat{\text{Var}}\left[\hat{\tau}\left(\boldsymbol{Z},\boldsymbol{y}(\boldsymbol{Z})\right)\right]$.

The Difference-in-Means in (4) and this conservative variance estimator can be used to construct asymptotically valid tests. Suppose standard regularity conditions that suffice for a central limit theorem (Li and Ding 2017) and the convergence of the suitably scaled variance estimator to a constant greater than the Difference-in-Means' true limiting variance. It then follows that we can conduct asymptotically valid hypothesis tests by referring the following test statistic to a standard Normal distribution:

$$\frac{\hat{\tau}\left(\boldsymbol{Z},\boldsymbol{y}(\boldsymbol{Z})\right)-\tau_0}{\sqrt{\widehat{\text{Var}}\left[\hat{\tau}\left(\boldsymbol{Z},\boldsymbol{y}(\boldsymbol{Z})\right)\right]}}, \tag{17}$$

where τ_0 is a null hypothesis about the average effect relative to a well-defined alternative hypothesis.

Analogous logic applies when researchers cue attributes in the body of emails. When researchers either do not measure opening or do so with error, the relevant estimator is the scaled Difference-in-Means in (10). For any assumed value of the proportion of Always-Openers, the variance of this estimator is

$$\text{Var}\left[\left(\frac{1}{\pi_{\text{AO}}}\right)\hat{\tau}\left(\boldsymbol{Z},\boldsymbol{y}(\boldsymbol{Z})\right)\right]=\left(\frac{N^2}{N_{\text{AO}}^2}\right)\text{Var}\left[\hat{\tau}\left(\boldsymbol{Z},\boldsymbol{y}(\boldsymbol{Z})\right)\right], \tag{18}$$

which can be conservatively estimated by $\left(N^2/N_{\text{AO}}^2\right)\widehat{\text{Var}}\left[\hat{\tau}\left(\boldsymbol{Z},\boldsymbol{y}(\boldsymbol{Z})\right)\right]$. Under suitable regularity conditions (now pertaining to potential outcomes for both replies and opens), referring the analogue of the standardized test statistic in (17) to a standard Normal distribution yields asymptotically valid tests of hypotheses about the ATE among Always-Openers.

Without measurement error in opening, researchers have access to not only the number of Always-Openers, but also exactly which legislators are Always-Openers. Hence, researchers can use the post-stratified Difference-in-Means in (16), which conditions on the legislators who open their emails. Because Assumption 4 implies that the opening of emails is equivalent to a baseline covariate, the variance of this Difference-in-Means is the same as that of the Difference-in-Means post-stratified on the value of a categorical covariate (Miratrix, Sekhon, and Yu 2013). With a conservative estimator of this variance and under suitable regularity conditions, asymptotically valid tests of hypotheses about the ATE among Always-Openers can be constructed in the same manner as the aforementioned tests using the Difference-in-Means and scaled Difference-in-Means.

The Supplementary Material provides expressions for all of the aforementioned variances and their conservative estimators. The Supplementary Material also provides two simple worked examples for the settings in which researchers measure opening with and without error. Both of these examples include corresponding `R` code available in Leavitt and Rivera-Burgos (2024).

7. Discussion: Statistical, Substantive, and Ethical Considerations

The solutions we have proposed thus far provide applied researchers with different design choices. These choices present trade-offs in terms of statistics, substance, and ethics. We argue that researchers ought to weigh all three of these considerations when deciding which of these choices to implement, if any.

7.1. Statistical Considerations

One of the main statistical considerations is that of power. All else equal, if one does not suppose that all legislators are Always-Openers, then the variance of the scaled Difference-in-Means in (10)

(applicable when researchers cue attributes after opening) will be greater than that of the canonical Difference-in-Means in (4) (applicable when researchers cue attributes before opening). Likewise, the estimator in (16), which conditions on Always-Openers in the absence of measurement error, effectively reduces the size of an experiment. This reduced size may hurt power relative to a design that exposes legislators to cues before opening.

Nevertheless, researchers do have leeway in increasing power under additional assumptions. One possibility is to vary the names and texts of emails, as well as to space out the timing of emails. As Rivera-Burgos and Rubio (2024) demonstrate, researchers may be able to do so without inducing SUTVA violations or arousing legislators' suspicions that their behavior is being observed by researchers.

7.2. Substantive Considerations

Among the many substantive trade-offs that researchers may face, one that stands out is the differing nature of cues (either implicit or explicit) that researchers can convey in different designs. Insofar as audit experiments signal race only via the name in the email address, then the cue of race is implicit. The implicit nature of this cue is important: Kirgios *et al.* (2022) show that when women and ethnoracial minorities explicitly state their identities—beginning an inquiry with, e.g., "as a Black woman . . ." (Kirgios *et al.* 2022, 383)—politicians are more likely to respond to these inquiries.

Insofar as researchers want to implicitly cue race, one potential benefit of cueing attributes before opening is that researchers can continue to cue race via the name in the email address. An analogous, implicit cue may be more difficult to convey after the opening of the email. One way of doing so might be, e.g., varying the vernacular English in the email messages. A large body of literature in sociolinguistics and related disciplines shows that different speech and writing styles are often racially marked (e.g., Labov 1972; Perry and Delpit 1998). The viability of this possibility and others will depend on the substance of a given audit experiment.

Alternatively, researchers who want to implicitly cue race might include, e.g., only a last name in the email address (or an email address with no name whatsoever) and then cue race via the first name in the body of the email. This strategy would enable researchers to continue implicitly cueing race in the body of the email via a putative constituent's first name. However, the ability to shield a recipient from the sender's first name may depend on the email providers of both the sender and the receiver.

In some audit experiments, researchers may be interested in the effects of only explicit racial cues. Explicitly cueing race in a natural way before legislators open the email may be difficult to achieve. Hence, researchers are likely to prefer a design that explicitly cues race after the opening of emails. If so, researchers should take care to avoid the use of any implicit racial cues before opening, which could lead to a violation of Assumption 4—no racial discrimination in opening. Alternatively, if researchers are interested in effects of a bundle of implicit and explicit racial cues, then cueing race in multiple ways both before and after opening may be a useful design choice so long as this design also maintains symmetry in cues of other attributes, such as party.

7.3. Ethical Considerations

The measuring of email opening as described in Section 5.2 raises new ethical issues related to informed consent in audit experiments among political elites. Existing scholarship has addressed this issue in general (Bischof *et al.* 2022; Crabtree and Dhima 2022; Desposato 2022; McClendon 2012; Riach and Rich 2004). Therefore, we focus specifically on the new wrinkle introduced by the possible measuring of opening.

Email open tracking may represent a greater breach of informed consent relative to existing audit experiments without open tracking. When deciding whether to reply, an email user presumably knows that a reply or not to an email will be observable to the user who sent the email and potentially others (e.g., anyone to whom the sender forwards the email). However, an email user may have a greater expectation of privacy in the decision to open an email. Assuming that researchers do not use a technology asking for consent before measuring opening, the measurement of an act for which

participants have a greater expectation of privacy raises the ethical costs of an audit experiment without informed consent.

Following the cost–benefit framework of Crabtree and Dhima (2022), researchers ought to take into account this additional ethical cost to participants when deciding whether to conduct an audit experiment and, conditional on doing so, whether and how to measure the opening of emails. Researchers may choose not to measure opening and to instead estimate and test hypotheses about the lower bound (in magnitude) of the ATE among Always-Openers. Alternatively, researchers may choose to measure only the proportion of openers, not exactly which legislators open emails. With this partial information, as the discussion in Section 5.1 mentions, researchers can use the scaled Difference-in-Means in (10) instead of the post-stratified Difference-in-Means in (16) in order to unbiasedly estimate the ATE among Always-Openers. Using this estimator in (10) enables researchers to preserve anonymity in which legislators open their emails, thereby mitigating the ethical costs of measuring opening in audit experiments without informed consent.

In addition to these ethical concerns when researchers measure opening, Crabtree and Dhima (2022) (and also Desposato 2022) emphasize consideration of the broader social costs and benefits of audit experiments. For example, insofar as the measuring of opening can help discern solutions to racial discrimination, this possible benefit ought to be weighed against the aforementioned ethical costs to experimental participants. While it is unclear how the cost–benefit comparison of measuring opening or not will shake out, the additional ethical costs to participants of open tracking should figure into this calculus.

8. Conclusion

Parsing mechanisms of discrimination is important for both policy and normative commitments. In this paper, we have shown the difficulties that audit experiments encounter in attempting to do so. Although the insights of our paper are broader, we have focused on parsing racial discrimination due to inferences about other attributes signaled by race—e.g., political party—and racial discrimination that exists above and beyond these other attributes. We show how to align statistical estimands with causal contrasts of interest, the bias in existing designs for these aligned estimands, and possible solutions to this bias that researchers can readily implement in their designs.

Nevertheless, at least two open questions remain. The first pertains to the settings and normative commitments under which researchers should aim to estimate one estimand (e.g., the overall ATE) relative to another (e.g., the ATE among Always-Openers). This paper has followed the literature in supposing that disentangling the two aforementioned mechanisms is important. Future work can help motivate and justify the conditions under which researchers should seek to infer different statistical estimands. The arguments in this paper should not be interpreted as supposing that racial discrimination due to decisionmakers' inferring other, nonracial attributes is somehow a more innocuous form of racial discrimination compared to other mechanisms.

Second, a solution proposed in this paper is designing audit experiments to justify the assumption of no racial discrimination in opening. Under this assumption, researchers can draw meaningful inferences about the ATE among Always-Openers. One concern, though, is that these inferences may suffer from lower statistical power, especially when few legislators open the emails. Therefore, future research can experimentally assess how to increase open rates to mitigate this concern, as well as other ways to increase statistical power.

Appendix: Proofs

A.1. Proof of Proposition 1

Proof. Under Assumptions 1 and 3, the Difference-in-Means is unbiased for τ in (2), that is, $\mathrm{E}[\tau(\boldsymbol{Z},\boldsymbol{y}(\boldsymbol{Z}))] = \tau$. Assumptions 1 and 2 imply that the ATE among Never-Openers is 0, that is, $\tau_{\mathrm{NO}} = 0$; hence, τ in (2) is

$$\tau = \pi_{\mathrm{AO}}\tau_{\mathrm{AO}} + \pi_{\mathrm{OWO}}\tau_{\mathrm{OWO}} + \pi_{\mathrm{OBO}}\tau_{\mathrm{OBO}}. \tag{A.1}$$

Then, taking the difference between τ in (A.1) and τ_{AO} yields

$$(\pi_{\text{AO}} - 1)\,\tau_{\text{AO}} + \pi_{\text{OWO}}\tau_{\text{OWO}} + \pi_{\text{OBO}}\tau_{\text{OBO}}.$$ □

A.2. Proof of Lemma 1

Proof. Recall that the average effect among Always-Openers is defined as

$$\tau_{\text{AO}} \coloneqq \bar{y}_{\text{AO}}(1) - \bar{y}_{\text{AO}}(0) = \left(\frac{1}{N_{\text{AO}}}\right)\sum_{i=1}^{N} \mathbb{1}\{s_i = \text{AO}\}\,(y_i(1) - y_i(0)). \tag{A.2}$$

Assumption 4 implies that all units are either Always-Openers or Never-Openers and Assumption 2 implies that the total number of potential replies among Never-Openers is 0. Therefore, $\sum_{i=1}^{N} y_i(1)$ is the sum of treated potential replies among Always-Openers and $\sum_{i=1}^{N} y_i(0)$ is the sum of control potential replies among Always-Openers. Therefore, (A.2) can be expressed as

$$\tau_{\text{AO}} = \left(\frac{1}{N_{\text{AO}}}\right)\sum_{i=1}^{N} (y_i(1) - y_i(0)). \tag{A.3}$$

Then expressing $\left(\frac{1}{N_{\text{AO}}}\right)$ as

$$\left(\frac{N}{N_{\text{AO}}}\right)\left(\frac{1}{N}\right)$$

and noting that $\pi_{\text{AO}} \coloneqq \left(\frac{N_{\text{AO}}}{N}\right)$ yields

$$\tau_{\text{AO}} = \left(\frac{N}{N_{\text{AO}}}\right)\left(\frac{1}{N}\right)\left[\sum_{i=1}^{N} y_i(1) - \sum_{i=1}^{N} y_i(0)\right] = \left(\frac{1}{\pi_{\text{AO}}}\right)\tau,$$

thereby completing the proof. □

A.3. Proof of Proposition 2

Proof. Lemma 1 under Assumptions 1, 2, and 4 implies that the ATE among Always-Openers is

$$\tau_{\text{AO}} = \left(\frac{N}{N_{\text{AO}}}\right)\left(\frac{1}{N}\right)\left[\sum_{i=1}^{N} y_i(1) - \sum_{i=1}^{N} y_i(0)\right] = \left(\frac{1}{\pi_{\text{AO}}}\right)\tau. \tag{A.4}$$

Since π_{AO} lies in $[0,1]$, it follows from the expression in (A.4) that whenever τ is negative, the ATE among Always-Openers is increasing in π_{AO}. Hence, with $\tau < 0$, the upper bound of τ_{AO} obtains when $\pi_{\text{AO}} = 1$, which, after plugging in 1 for π_{AO} in (A.4), yields τ.

By contrast, whenever τ is positive, the ATE among Always-Openers is decreasing in π_{AO}. Hence, with $\tau > 0$, the lower bound of τ_{AO} obtains when $\pi_{\text{AO}} = 1$, which, after plugging in 1 for π_{AO} in (A.4), also yields τ.

Finally, when $\tau = 0$, it follows from the expression in (A.4) that the ATE among Always-Openers, τ_{AO}, is equal to τ for all values of π_{AO}, which completes the proof. □

A.4. Proof of Proposition 3

Proof. First, note that

$$\left(\frac{1}{\pi_{\text{AO}}}\right) = \frac{1}{(N_{\text{AO}}/N)} = \frac{N}{N_{\text{AO}}}$$

and that the linearity of expectations implies that

$$\mathrm{E}\left[\left(\frac{\hat{\tau}}{\pi_{\mathrm{AO}}}\right)\right] = \left(\frac{1}{\pi_{\mathrm{AO}}}\right)\mathrm{E}\left[\hat{\tau}\right] = \left(\frac{N}{N_{\mathrm{AO}}}\right)\mathrm{E}\left[\hat{\tau}\right].$$

Under SUTVA in Assumption 1 and CRA in Assumption 3, it follows that

$$\begin{aligned}\left(\frac{N}{N_{\mathrm{AO}}}\right)\mathrm{E}\left[\hat{\tau}\right] &= \left(\frac{N}{N_{\mathrm{AO}}}\right)\left[\left(\frac{1}{N}\right)\sum_{i=1}^{N} y_i(1) - \left(\frac{1}{N}\right)\sum_{i=1}^{N} y_i(0)\right] \\ &= \left(\frac{1}{N_{\mathrm{AO}}}\right)\sum_{i=1}^{N} y_i(1) - \left(\frac{1}{N_{\mathrm{AO}}}\right)\sum_{i=1}^{N} y_i(0), \end{aligned} \tag{A.5}$$

which completes the proof. □

A.5. Proof of Proposition 4

Proof. Under Assumptions 4 and 6, $\tilde{m}_i$ is fixed over Z_i for all $i = 1, \ldots, N$. Hence, $\tilde{m}_i$ is a fixed baseline covariate for all $i = 1, \ldots, N$. Consequently, the Difference-in-Means conditional on opening is the post-stratified estimator described in Miratrix *et al.* (2013) and, thus, the proof follows immediately from Theorem 2.1 of Miratrix *et al.* (2013). □

Acknowledgements. For their valuable feedback, we thank Anna Wilke, Georgiy Syunyaev, Don Green, Tara Slough, Naoki Egami, Sandy Korenman, and Dahlia Remler. We also thank audiences at Baruch College's Marxe School Faculty Seminar, Baruch College's Data Science Cluster Meeting, the 2022 American Causal Inference Conference and New York University's Data Science Lunch Seminar Series, as well as students in Washington University's Field Experiments in Political Science course during the Fall of 2023.

Funding Statement. The authors state no funding involved.

Competing interest. The authors have no competing interest to declare.

Data Availability Statement. Code Ocean was not used. The data can also be accessed via Dataverse at https://doi.org/10.7910/DVN/R3JGWS. The replication materials and reproduced output can be accessed at https://drive.google.com/drive/u/1/folders/1r56SvUDSeCvswjVNUeAwnH5ApeOV5J5y.

Supplementary Material. For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2024.3.

References

Andersen, S. C., and T. S. Guul. 2019. "Reducing Minority Discrimination at the Front Line—Combined Survey and Field Experimental Evidence." *Journal of Public Administration Research and Theory* 29 (3): 429–444.

Bartels, L. M. 2008. *Unequal Democracy: The Political Economy of the New Gilded Age*. Princeton: Princeton University Press.

Bertrand, M., D. Chugh, and S. Mullainathan. 2005. "Implicit Discrimination." *American Economic Review* 95 (2): 94–98.

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

Bischof, D., et al. 2022. "Advantages, Challenges and Limitations of Audit Experiments with Constituents." *Political Studies Review* 20 (2): 192–200.

Bueno de Mesquita, E., and S. A. Tyson. 2020. "The Commensurability Problem: Conceptual Difficulties in Estimating the Effect of Behavior on Behavior." *The American Political Science Review* 114 (2): 375–391.

Butler, D. M., and D. E. Broockman. 2011. "Do Politicians Racially Discriminate against Constituents? A Field Experiment on State Legislators." *American Journal of Political Science* 55 (3): 463–477.

Butler, D. M., and C. Crabtree. 2017. "Moving beyond Measurement: Adapting Audit Studies to Test Bias-Reducing Interventions." *Journal of Experimental Political Science* 4 (1): 57–67.

Butler, D. M., and J. Homola. 2017. "An Empirical Justification for the Use of Racially Distinctive Names to Signal Race in Experiments." *Political Analysis* 25 (1): 122–130.

Crabtree, C. 2018. "An Introduction to Conducting Email Audit Studies." In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, edited by S. M. Gaddis, Volume 14 of Methodos Series, Chapter 5, 103–117. Cham: Springer.

Crabtree, C., and K. Dhima. 2022. "Auditing Ethics: A Cost–Benefit Framework for Audit Studies." *Political Studies Review* 20 (2): 209–216.
Daniel, W. W. 1968. *Racial Discrimination in England: Based on the P.E.P. Report, Volume 1084*. Harmondsworth: Penguin.
Desposato, S. 2022. "Public Impacts from Elite Audit Experiments: Aggregate and Line-Cutting Harms." *Political Studies Review* 20 (2): 217–227.
Devine, P. G. 1989. "Stereotypes and Prejudice: Their Automatic and Controlled Components." *Journal of Personality and Social Psychology* 56 (1): 5–18.
Elder, E. M., and M. Hayes. 2023. "Signaling Race, Ethnicity, and Gender with Names: Challenges and Recommendations." *The Journal of Politics* 85 (2): 764–770.
Fenno, R. F. 1978. *Home Style: House Members in Their Districts*. Boston: Little, Brown & Co.
Frangakis, C. E., and D. B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58 (1): 21–29.
Fryer, R. G., and S. D. Levitt. 2004. "The Causes and Consequences of Distinctively Black Names." *The Quarterly Journal of Economics* 119 (3): 767–805.
Gaddis, S. M. 2018. "An Introduction to Audit Studies in the Social Sciences." In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, edited by S. M. Gaddis, Volume 14 of Methodos Series, Chapter 1, 3–44. Cham: Springer.
Gaddis, S. M. 2019. "Understanding the 'How' and 'Why' Aspects of Racial–Ethnic Discrimination: A Multimethod Approach to Audit Studies." *Sociology of Race and Ethnicity* 5 (4): 443–455.
Heckman, J. J. 1998. "Detecting Discrimination." *Journal of Economic Perspectives* 12 (2): 101–116.
Holland, P. W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–960.
Hughes, D. A., M. Gell-Redman, C. Crabtree, N. Krishnaswami, D. Rodenberger, and G. Monge. 2020. "Persistent Bias among Local Election Officials." *Journal of Experimental Political Science* 7 (3): 179–187.
Imbens, G. W., and D. B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press.
Kirgios, E. L., A. Rai, E. H. Chang, and K. L. Milkman. 2022. "When Seeking Help, Women and Racial/Ethnic Minorities Benefit from Explicitly Stating Their Identity." *Nature Human Behaviour* 6 (3): 383–391.
Labov, W. 1972. *Language in the Inner City: Studies in the Black English Vernacular*. Philadelphia: University of Pennsylvania Press.
Landgrave, M., and N. Weller. 2020. "Do more Professionalized Legislatures Discriminate Less? The Role of Staffers in Constituency Service." *American Politics Research* 48 (5): 571–578.
Landgrave, M., and N. Weller. 2022. "Do Name-Based Treatments Violate Information Equivalence? Evidence from a Correspondence Audit Experiment." *Political Analysis* 30 (1): 142–148.
Leavitt, T., and V. Rivera-Burgos. 2024. "Replication Data for: Audit Experiments of Racial Discrimination and the Importance of Symmetry in Exposure to Cues." Harvard Dataverse V1. https://doi.org/10.7910/DVN/R3JGWS
Li, X., and P. Ding. 2017. "General Forms of Finite Population Central Limit Theorems with Applications to Causal Inference." *Journal of the American Statistical Association* 112 (520): 1759–1769.
McClendon, G. H. 2012. "Ethics of Using Public Officials as Field Experimental Subjects." *The Experimental Political Scientist* 3 (1): 13–20.
Miratrix, L. W., J. S. Sekhon, and B. Yu. 2013. "Adjusting Treatment Effect Estimates by Post-Stratification in Randomized Experiments." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75 (2): 369–396.
Neyman, J. 1923. "Sur les Applications de la théorie Des probabilités Aux Experiences Agricoles: Essai Des Principes." *Roczniki Nauk Rolniczych* 10: 1–51.
Pedulla, D. S. (2018). Emerging Frontiers in Audit Study Research: Mechanisms, Variation, and Representativeness. In S. M. Gaddis (Ed.), *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, Volume 14 of Methodos Series, Chapter 9, 179–195. Cham: Springer.
Perry, T., and L. Delpit. 1998. *The Real Ebonics Debate: Power, Language, and the Education of African-American Children*. Boston: Beacon Press.
Riach, P. A., and J. Rich. 2004. "Deceptive Field Experiments of Discrimination: Are They Ethical?" *Kyklos* 57 (3): 457–470.
Rivera-Burgos, V., and J. M. Rubio. 2024. "Responsiveness to Coethnics and Cominorities: Evidence from an Audit Experiment of State Legislators." *Journal of Race, Ethnicity, and Politics* 9 (1): 55–79.
Rubin, D. B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688–701.
Slough, T. 2023. "Phantom Counterfactuals." *American Journal of Political Science* 67 (1): 137–153.
Wienk, R. E., C. E. Reid, J. C. Simonson, and F. J. Eggers. 1979. *Measuring Racial Discrimination in American Housing Markets: The Housing Market Practices Survey, Volume 444*. Washington, DC: Division of Evaluation, US Department of Housing and Urban Development, Office of Policy Development and Research.

Cite this article: Leavitt, T. and Rivera-Burgos, V. (2024). Audit Experiments of Racial Discrimination and the Importance of Symmetry in Exposure to Cues. *Political Analysis*, 445–462. https://doi.org/10.1017/pan.2024.3