Political Science Research and Methods (2026), 1–14
doi:10.1017/psrm.2025.10082

ORIGINAL ARTICLE

Navigating the mismeasurement of intermediary variables in message-based experiments

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

[1]Baruch College, Austin W. Marxe School of Public and International Affairs, New York, NY, USA and [2]Baruch College, Political Science, New York, NY, USA
Corresponding author: Thomas Leavitt; Email: thomas.leavitt@baruch.cuny.edu

(Received 15 February 2025; revised 26 August 2025; accepted 3 December 2025)

Abstract:
Researchers frequently deliver treatments through messages, as in many audit and get-out-the-vote (GOTV) experiments. These message-based experiments often hinge on intermediary variables—actions subjects must take to actually receive the treatment or control embedded in a message. Whether subjects open the message is a crucial intermediary step, which can serve as a condition for estimating downstream treatment effects or as an outcome of interest in its own right. Yet opens are often measured with error, most notably when some openers are misclassified as non-openers in email-based studies. We characterize the resulting bias, derive interpretable bounds on effects for well-defined subgroups, and provide sensitivity analyses for mismeasurement, thereby offering practical guidance for message-based experiments conducted through email and other communication technologies.

Keywords: causal inference; measurement error; principal stratification; partial identification; sensitivity analysis; audit experiments; get-out-the-vote experiments

1. Introduction

Many experiments randomly assign individuals to message-based treatments to study their effects on outcomes ranging from voter mobilization to bureaucratic responsiveness. These experiments have been carried out using a range of technologies. In get-out-the-vote (GOTV) studies, for example, researchers may use telephones to deliver messages encouraging voters to turn out (e.g., Adams and Smith, 1980). In audit studies, by contrast, researchers have used fax machines to unobtrusively send messages to decision-makers in order to reveal their behavior in real-world contexts (e.g., Bertrand and Mullainathan, 2004). Regardless of the technology, a critical intermediate step is whether subjects actually receive the message, such as by answering the phone, reading the fax, or opening the door or envelope.

Today, one of the most common technologies for delivering message-based treatments is email, which is predominant in audit experiments (Crabtree, 2018) and common in GOTV studies (e.g., Nickerson, 2008a; Bennion and Nickerson, 2011; Malhotra *et al.*, 2012; Rivera *et al.*, 2023). Many experiments also use social media applications, such as public tweets or direct messages on Twitter (now X) (Coppock *et al.*, 2016; Bail *et al.*, 2018) and private messages on Facebook (Van Remoortere *et al.*, 2024). Whether delivered via email, social media or other channels, measuring whether a message is opened can be important in its own right (e.g., Calfano, 2019; Hughes *et al.*, 2020; Gaynor and Gimpel, 2024) or as a way to estimate effects among recipients who would actually view the treatment message (e.g., Moy, 2021; Schiff and Schiff, 2023; Incerti, 2024; Lee, 2024).

Researchers increasingly incorporate the measurement of intermediary variables (like the opening of messages) into the design of message-based experiments, which is a valuable advancement. However, an overlooked issue is the error inherent to measures of opening. (See McClendon 2014; Bergner *et al.*, 2019 and Persian *et al.*, 2023 for three exceptions.) In experiments that deliver messages via email, researchers typically measure opens via tracking pixels—tiny, invisible images embedded in the message. When a recipient opens the email, that recipient's email client downloads the image from the sender's server, which logs details such as the time of opening and the recipient's device. However, given the wide availability of software that blocks open tracking, researchers may incorrectly classify some recipients who opened the email as non-openers.

Similar challenges presumably exist in the more nascent practice of conducting message-based experiments on social media platforms. For example, since 2016, direct messages on Twitter (X) have included read receipts by default, though users can disable this feature in their privacy settings (see Woollaston-Webber, 2016). Researchers can, in principle, use these receipts to determine whether a recipient opened the message. However, if a recipient disables the feature, researchers cannot tell whether that person genuinely did not open the message or opened it without triggering a read receipt. Thus, although researchers rarely measure opening on social media platforms, the problems that beset email-based experiments likely extend to message-based experiments conducted through social media and other technologies.

In what follows, we address this problem of measurement error in two settings: when opening is itself the outcome of interest and when opening is used to estimate effects on downstream outcomes (e.g., voter turnout or message replies) among a specific stratum of subjects. In both settings, we explicate how measurement error in opening can bias effect estimates. Nevertheless, we formally show that researchers can still use measures of opening to estimate informative bounds of effects among a meaningful subset of experimental subjects—namely, the individuals who do not block open tracking. We also show how researchers can incorporate sensitivity analyses for the estimation of bounds on causal targets of interest. The methodological framework advanced in this paper helps clarify what conclusions are actually justified by existing studies, and also points to new methods researchers can implement in future studies.

We begin the remainder of this paper with a formal setup for subsequent arguments. The following section describes the issue of error in the measurement of opening before the subsequent two sections lay out the implications of this measurement error in the two aforementioned settings. The final, concluding section discusses the paper, with an emphasis on its implications for applied practice, and points to open questions for message-based experiments conducted via myriad technologies.

2. Formal setup

2.1. Assignment process and potential outcomes

Consider an experiment that 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 message-based experiments, $i = 1, \ldots, N$ often indexes the N subjects' message-receiving accounts (such as email addresses, phone numbers, or social media profiles). The indicator variable $z_i = 1$ or $z_i = 0$ denotes whether individual i is assigned to treatment ($z_i = 1$) or control ($z_i = 0$). 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 framework of causality (Neyman, 1923; Rubin, 1974; Holland, 1986), where a potential outcomes schedule is defined as a vector-valued function that maps each possible treatment assignment to an N-dimensional vector of real numbers. The vectors of potential outcomes, denoted by $\boldsymbol{y}(\boldsymbol{z})$ for $\boldsymbol{z} \in \{0, 1\}^N$, are the elements in the range of the potential

Table 1. Principal strata of subjects

	$m_i(1) = 1$	$m_i(1) = 0$
$m_i(0) = 1$	*Always-Opener*	*Only-Control-Opener*
$m_i(0) = 0$	*Only-Treatment-Opener*	*Never-Opener*

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 $\boldsymbol{z} \in \{0,1\}^N$. These outcomes may depend on opening the message—for example, clicking an embedded link or replying—or they may not, as with offline behaviors such as voting in an election.

We will refer to $\boldsymbol{y}(\boldsymbol{z})$ for $\boldsymbol{z} \in \{0,1\}^N$ as the *final outcome*, in contrast to the *intermediate outcome* of opening. We denote the intermediate potential outcomes of whether the subjects would open the messages under assignment $\boldsymbol{z} \in \{0,1\}^N$ by $\boldsymbol{m}(\boldsymbol{z})$, where $\boldsymbol{m}(\boldsymbol{z}) \in \{0,1\}^N$. The individual outcome, $m_i(\boldsymbol{z})$, denotes whether individual i would open the message under assignment $\boldsymbol{z} \in \{0,1\}^N$.

With 2^N assignments, there are in principle 2^N potential outcomes for each individual subject. However, we make the stable unit treatment value assumption (SUTVA) for both final 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 Assumption 1, we write a final potential outcome 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 $z_i = 0$. The same is true for intermediate variables measured post-treatment.

Under SUTVA, we can partition individuals into principal strata (Frangakis and Rubin, 2002) based on the intermediate variable of opening. We define the principal strata for an arbitrary subject, i, in Table 1.

The proportions of units in the respective principal strata are defined as

$$\pi_{11} := \left(\frac{1}{N}\right)\sum_{i=1}^{N} \mathbb{1}\left\{m_i(1) = 1,\, m_i(0) = 1\right\}, \quad \pi_{10} := \left(\frac{1}{N}\right)\sum_{i=1}^{N} \mathbb{1}\left\{m_i(1) = 1,\, m_i(0) = 0\right\},$$

$$\pi_{01} := \left(\frac{1}{N}\right)\sum_{i=1}^{N} \mathbb{1}\left\{m_i(1) = 0,\, m_i(0) = 1\right\}, \quad \pi_{00} := \left(\frac{1}{N}\right)\sum_{i=1}^{N} \mathbb{1}\left\{m_i(1) = 0,\, m_i(0) = 0\right\}.$$

We also let π_1 denote the proportion of subjects who belong to either the Always-Opener or Only-Treatment-Opener strata, i.e., $\pi_1 := (1/N)\sum_{i=1}^{N} \mathbb{1}\{m_i(1) = 1\}$. Consequently, $1 - \pi_1$ is the proportion of subjects belonging to the Only-Control-Opener or Never-Opener strata.

We write an individual treatment effect for the final outcome as $\tau_i := y_i(1) - y_i(0)$ and for the intermediate outcome of opens as $\theta_i := m_i(1) - m_i(0)$. For each outcome, the average treatment effect (ATE) is simply the average of the individual effects over all units. That is, these two ATEs are

$$\tau := \left(\frac{1}{N}\right)\sum_{i=1}^{N} \tau_i = \left(\frac{1}{N}\right)\sum_{i=1}^{N} y_i(1) - y_i(0) \tag{1}$$

for the final outcome and

$$\theta := \left(\frac{1}{N}\right)\sum_{i=1}^{N} \theta_i = \left(\frac{1}{N}\right)\sum_{i=1}^{N} m_i(1) - m_i(0) \tag{2}$$

for the intermediate outcome of opens. Using the principal strata in Table 1, we can equivalently express θ as $\pi_{10} - \pi_{01}$.

Going forward, we suppose complete random assignment (CRA) of the $N \geq 2$ units, $n_1 \geq 1$ to treatment and the remaining $n_0 := N - n_1 \geq 1$ to control. CRA describes an assignment mechanism in which the treatment vector $\boldsymbol{Z}$ is random, taking a value $\boldsymbol{z} \in \{0,1\}^N$ with probability $p(\boldsymbol{z})$.

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

In a randomized controlled experiment, Assumption 2 is ensured to hold by the researcher.

CRA in Assumption 2 implies that the canonical Difference-in-Means estimator with replies as the outcome,

$$\hat{\tau}\left(\boldsymbol{Z}, \boldsymbol{y}(\boldsymbol{Z})\right) := \left(\frac{1}{n_1}\right) \boldsymbol{Z}^{\top} \boldsymbol{y}\left(\boldsymbol{Z}\right) - \left(\frac{1}{n_0}\right) \left(\boldsymbol{1} - \boldsymbol{Z}\right)^{\top} \boldsymbol{y}\left(\boldsymbol{Z}\right), \tag{3}$$

is unbiased for τ in (1). This result follows directly from Assumptions 1 and 2. Similarly, the Difference-in-Means estimator with opens as the outcome, $\boldsymbol{m}\left(\boldsymbol{Z}\right)$, is unbiased for θ in (2).

2.2. Characterizing measurement error in opening

Unfortunately, researchers rarely have direct access to the outcome of actual opens, only *measures* of opening (based on the aforementioned tracking pixels), which are prone to error. We write $\tilde{\boldsymbol{m}}(\boldsymbol{z})$ for the potential measures of opening under assignment $\boldsymbol{z} \in \{0,1\}^N$. For each individual i, $\tilde{m}_i(\boldsymbol{z})$ indicates whether the researcher would record that individual as opening the message under assignment $\boldsymbol{z}$.

Going forward, we develop our framework on measurement error in the context of email. We focus on email because it remains the communication technology most widely used by researchers in experiments. However, as we noted in the introduction, our framework also applies to other platforms, such as social media, which may become more prevalent in future studies.

Leavitt and Rivera-Burgos (2024) identify two forms of measurement error in message-based experiments conducted via email:

(1) If an email user's software automatically scans incoming messages, it may download the tracking pixel, falsely marking the email as opened (false positive).
(2) If an email user's software blocks open tracking, it may falsely fail to register the email as opened, even if it was (false negative).

The first form of measurement error poses little threat to the design of message-based experiments (Leavitt and Rivera-Burgos, 2024). Unless a recipient has software that blocks open tracking, tracking pixels record when a recipient opens an email. If an email is logged as opened at the exact moment it is sent, the opening is presumably due to automated scanning software. Researchers can also conduct pretests by sending placebo messages at varied times; repeated immediate openings would indicate a background application that preloads messages. Such cases can be re-coded as unopened, with any subsequent genuine openings still captured by the pixel. We suppose this coding rule throughout.

For the second form of measurement error, there are no simple solutions. However, a crucial feature of this measurement error is that whether it could exist for a particular individual is a baseline covariate that is independent of treatment assignment. In other words, whether an individual has software that blocks open tracking is presumably fixed before (and, hence, independent of) whether one sends that individual a message with the treatment or control condition.

Assumption 3 below formalizes this feature of measurement error (along with that of no false positives).

Assumption 3. (Measurement error independent of treatment) *For all $i = 1, \ldots, N$ units, there exists a baseline covariate $u_i = 1$ or $u_i = 0$ such that $\tilde{m}_i(\boldsymbol{z}) = m_i(\boldsymbol{z})(1 - u_i)$ for all $\boldsymbol{z} \in \{0, 1\}^N$.*

The unobservable covariate, u_i, indicates if individual i has software that blocks open tracking ($u_i = 1$) or does not ($u_i = 0$). Assumption 3 implies that there is no measurement error for all individuals who do not block open tracking. Assumption 3 also implies that, for all individuals who do have software that blocks open tracking (i.e., all $i = 1, \ldots, N$ with $u_i = 1$), the measure of opening (correct or not) is fixed at 0 across treatment and control conditions (even if actual opening is not). Finally, note that Assumption 3 implies that, if $\tilde{m}_i(\boldsymbol{z}) = 0$ and $m_i(\boldsymbol{z}) = 1$ then u_i must be equal to 1, though the converse is not true.

With this unobservable covariate indicating whether individuals block open tracking, we now define several additional quantities. Let $N^u := \sum_{i=1}^{N} \mathbb{1}\{u_i = u\}$ for $u = 1$ or $u = 0$ denote the number of subjects who do ($u = 1$) or do not ($u = 0$) block open tracking. Also let $\bar{u} := (1/N) \sum_{i=1}^{N} u_i$ denote the proportion of subjects who block open tracking and, finally, let $\theta^u := (1/N^u) \sum_{i=1}^{N} \mathbb{1}\{u_i = u\}\theta_i$ be the conditional ATE at either $u = 1$, written as $\theta^{u=1}$, or $u = 0$, written as $\theta^{u=0}$.

We also let $\tilde{\pi}_{11}$, $\tilde{\pi}_{10}$, $\tilde{\pi}_{01}$ and $\tilde{\pi}_{00}$ denote proportions analogous to π_{11}, π_{10}, π_{01} and π_{00}, respectively, but in terms of (potentially erroneous) *measures* of opening under treatment and control. Lemma S.1 in the Supplementary Appendix shows that this representation of measures of opening in terms of principal strata is justified because SUTVA for actual opens implies SUTVA for measures of opens. To refer to quantities in terms of measures of opening under treatment and control, we henceforth affix the modifier "measurable" before any reference to a principal stratum in Table 1.

3. Estimating causal effects on opening

Under the form of measurement error described above, estimates of θ under CRA can be biased. This bias is especially consequential when opening itself is an important outcome. For example, in audit experiments to detect discrimination, the opening of emails matters because, as Hughes *et al.* (2020), p. 184 note, it is a "high volume, low-attention task" that is particularly susceptible to implicit bias (Devine, 1989; Bertrand *et al.*, 2005). Opening is also substantively important in other domains, such as political marketing, where researchers seek to infer the effects of various subject lines on open rates (e.g., Calfano, 2019; Gaynor and Gimpel, 2024).

To derive this bias, we first write the Difference-in-Means in which measures of opens are the outcome as

$$\hat{\theta}\left(\boldsymbol{Z}, \tilde{\boldsymbol{m}}\left(\boldsymbol{Z}\right)\right) = \left(\frac{1}{n_1}\right) \boldsymbol{Z}^\top \tilde{\boldsymbol{m}}\left(\boldsymbol{Z}\right) - \left(\frac{1}{n_0}\right) \left(\mathbf{1} - \boldsymbol{Z}\right)^\top \tilde{\boldsymbol{m}}\left(\boldsymbol{Z}\right), \tag{4}$$

where $\tilde{\boldsymbol{m}}(\boldsymbol{Z})$ is the collection of random, observable measures of opens for all $i = 1, \ldots, N$ subjects. Proposition 3.1 below provides the bias of this estimator in (4) for θ in (2).

Proposition 3.1. *Under Assumptions 1–3, the bias of the Difference-in-Means in (4) for the average effect in (2) is*

$$\mathrm{E}\left[\hat{\theta}\left(\boldsymbol{Z}, \tilde{\boldsymbol{m}}\left(\boldsymbol{Z}\right)\right)\right] - \theta = -\bar{u}\theta^{u=1}. \tag{5}$$

The proof of Proposition 3.1 is in the Supplementary Appendix, as are all other proofs.

Proposition 3.1 states that the bias depends on two quantities: the proportion of subjects who have software that blocks open tracking and the ATE among this subgroup of subjects. In expectation, an experiment may either overstate or understate the magnitude of θ depending on whether the ATE among the subgroup of individuals who block open tracking is negative or positive. The bias will be 0 when there are no subjects who block open tracking or the ATE is 0 among the individuals who block open tracking.

3.1. Estimating a subgroup ATE on opening

Despite the bias in estimating the ATE on opening due to measurement error, it is still possible to reliably estimate another quantity. The Difference-in-Means with measures of opening as the outcome is informative about the ATE among the subgroup of subjects who do not block open tracking. Recall that this ATE is formally defined as

$$\theta^{u=0} := \left(\frac{1}{N^{u=0}}\right) \sum_{i=1}^{N} \mathbb{1}\{u_i = 0\}\theta_i. \tag{6}$$

This target in (6) can be substantively important in that individuals who do not block open tracking may make up a large majority of all experimental subjects. For example, in an experiment with $1,400$ mayors across all 50 U.S. states, Moy (2021) states that the overall open rate is 0.78, which (if one presumes the approach to measurement described thus far) implies that the proportion of mayors who do not block open tracking is *at least* 0.78.

Whether measurement error could exist is a baseline covariate (albeit unobserved) and measures of opens are always 0 when error does exist (Assumption 3). Therefore, we can express $\tilde{\pi}_{11}$, $\tilde{\pi}_{10}$ and $\tilde{\pi}_{01}$ in terms of actual opens as

$$\tilde{\pi}_{11} = \left(\frac{1}{N}\right) \sum_{i=1}^{N} (1 - u_i)\mathbb{1}\left\{m_i(1) = 1,\ m_i(0) = 1\right\},$$

$$\tilde{\pi}_{10} = \left(\frac{1}{N}\right) \sum_{i=1}^{N} (1 - u_i)\mathbb{1}\left\{m_i(1) = 1,\ m_i(0) = 0\right\},$$

$$\tilde{\pi}_{01} = \left(\frac{1}{N}\right) \sum_{i=1}^{N} (1 - u_i)\mathbb{1}\left\{m_i(1) = 0,\ m_i(0) = 1\right\}.$$

Measurable Never-Openers, by contrast, include two groups: (i) Never-Openers who do not use software that blocks open tracking and (ii) individuals who do use such software (belonging to any of the four principal strata in Table 1). Hence, the proportion of measurable Never-Openers is

$$\tilde{\pi}_{00} = \left(\frac{1}{N}\right) \sum_{i=1}^{N} (1 - u_i)\mathbb{1}\{m_i(1) = 0,\ m_i(0) = 0\} + u_i. \tag{7}$$

Proposition 3.2 derives bounds on the ATE among individuals who do not block open tracking in terms of this quantity, $\tilde{\pi}_{00}$.

Proposition 3.2. *Under Assumptions 1 and 3, the lower and upper bounds (in magnitude) of the ATE among subjects who do not block open tracking, denoted by* $\underline{\theta}^{u=0}$ *and* $\overline{\theta}^{u=0}$*, respectively, are*

$$\underline{\theta}^{u=0} = \tilde{\pi}_{10} - \tilde{\pi}_{01} \tag{8}$$

$$\overline{\theta}^{u=0} = \left(\frac{1}{1 - \tilde{\pi}_{00}}\right)(\tilde{\pi}_{10} - \tilde{\pi}_{01}). \tag{9}$$

The lower bound in (8) corresponds to the case where none of the measurable Never-Openers block open tracking, while the upper bound in (9) corresponds to the case where all of them do. The Difference-in-Means in (4) is unbiased for the lower bound. Researchers can then assess sensitivity to different numbers of individuals with blocking software using

$$\left(\frac{N}{N - N^{u=1}}\right) \hat{\theta}\left(\boldsymbol{Z}, \tilde{\boldsymbol{m}}\left(\boldsymbol{Z}\right)\right), \tag{10}$$

where, given the observed data, the possible values of the unknown $N^{u=1}$ range from 0 to the total number of units (across treatment and control conditions) recorded as not opening the emails,

$\sum_{i=1}^{N} z_i[1-\tilde{m}_i(1)] + (1-z_i)[1-\tilde{m}_i(0)]$. Setting $N^{u=1} = 0$ produces the lower bound in (8). The maximum value of $N^{u=1}$ produces the upper bound in (9) in which all measurable Never-Openers are individuals who block open tracking, thereby making the number of individuals who do not, $N - N^{u=1} = N^{u=0}$, equal to $N(\tilde{\pi}_{11} + \tilde{\pi}_{10} + \tilde{\pi}_{01}) = N(1 - \tilde{\pi}_{00})$, which is the denominator in (9).

Proposition 3.2 is valuable because, in a randomized experiment (i.e., under Assumption 2), the lower bound in (8) is equal to the expected value of the Difference-in-Means in (4). The bound in (8) is the conditional ATE with the smallest magnitude. Hence, the results of an experiment, even if with measurement error, can be interpreted as a conservative estimate of the ATE among subjects who do not block open tracking.

3.2. Incorporating auxiliary information

Direct measures are not the only way to determine whether an individual has opened an email. Certain final outcomes can also reveal whether an opening has occurred. For example, replying to a message or clicking a link embedded in the message are actions that require an individual to have opened the email.

More formally, suppose that the final outcome is binary $\boldsymbol{y}(\boldsymbol{z}) \in \{0,1\}^N$ for all $\boldsymbol{z} \in \{0,1\}^N$, as is common with final outcomes, such as email replies. Then consider the following assumption.

Assumption 4. (No Positive Outcome without Opening) *For all $i = 1, \ldots, N$ units, $y_i(z) \leq m_i$ for $z = 1$ and $z = 0$.*

A logical consequence of Assumption 4, together with Assumption 3, is the following: If $\tilde{m}_i(\boldsymbol{z}) = 0$ and $y_i(\boldsymbol{z}) = 1$, then $u_i = 1$. To see this, note that when $y_i(\boldsymbol{z}) = 1$, Assumption 4 implies $m_i(\boldsymbol{z}) = 1$. Given $m_i(\boldsymbol{z}) = 1$, Assumption 3 implies that observing $\tilde{m}_i(\boldsymbol{z}) = 0$ requires $u_i = 1$.

Assumption 4 points to two ways in which researchers might incorporate auxiliary information about opening from final outcomes. First, one could change the measure of opening so that it is equal to 1 if either $y_i(\boldsymbol{z}) = 1$ or $\tilde{m}_i(\boldsymbol{z}) = 1$. Second, one could draw on final outcomes to tighten the bounds of the proportion of measurable Never-Openers in (7) and, consequently, the bounds of the ATE on opening among individuals who do not block open tracking.

In the Supplementary Appendix, we show that the first approach can lead to biased estimates of the ATE on opening and of the ATE on opening among the subgroup of subjects who do not block open tracking. We focus here on the second approach. Recall that the lower bound in (8) arises when the unknown $N^{u=1}$ takes its minimum value of 0. The upper bound in (9) arises when $N^{u=1}$ takes its maximum value, equal to the total number of individuals recorded as not opening the email across treatment and control, $\sum_{i=1}^{N} z_i[1-\tilde{m}_i(1)] + (1-z_i)[1-\tilde{m}_i(0)]$. Under Assumption 4, we can increase the lower bound to the number of individuals who replied to the email and were recorded as not opening it. Hence, we write the lower and upper bounds of $N^{u=1}$, given the observed data, as

$$\underline{N}^{u=1} = \sum_{i=1}^{N} z_i(1)[1-\tilde{m}_i(1)]y_i(1) + (1-z_i)[1-\tilde{m}_i(0)]y_i(0) \tag{11}$$

$$\overline{N}^{u=1} = \sum_{i=1}^{N} z_i[1-\tilde{m}_i(1)] + (1-z_i)[1-\tilde{m}_i(0)]. \tag{12}$$

Researchers can then deploy the estimator in (10) over the tighter feasible range for $N^{u=1}$, from the lower bound in (11) to the upper bound in (12).

Table 2. Email opening and reply outcomes for election officials in dataset block 623

arab_name	open	reply
1	0	1
0	0	0
1	1	1
0	1	1

3.3. Empirical application: audit experiment on racial bias

For a straightforward application of this approach, consider the audit experiment from Hughes *et al.* (2020) in which the opening of emails is an important outcome for the detection of implicit racial bias among local election officials. In this experiment, Hughes *et al.* (2020) construct blocks of local election officials based on a range of their baseline covariates. Within these blocks, the researchers assign election officials to emails from one of four randomly chosen aliases that cue either White, African-American, Latino, or Arab identity.

For simplicity, we condition our analysis on election officials assigned either a White or Arab identity cue, along with each block's realized number of officials in each of these two conditions. After this conditioning, the experiment includes 3,201 local election officials across 1,599 blocks. We excluded from our analysis all blocks lacking at least one treated official (Arab alias) and one control official (White alias).

To provide intuition for our analysis, Table 2 presents the officials in the block with dataset label 623.

Table 2 shows four officials, two of whom are marked as not opening the email. Hence, the upper bound on the number of subjects who block open tracking is 2. For the lower bound, note that one official is marked as not opening the email but nevertheless replied. Hence, there is at least one official whose email blocks open tracking. Therefore, when we apply the estimator in (10) to block 623, the factor, $N/(N - N^{u=1})$, can be equal to either $4/2$ or $4/3$. The researcher multiplies either factor by the Difference-in-Means with `open` as the outcome and `arab_name` as the treatment variable. Multiplying by $4/2$ estimates the upper bound in magnitude, while multiplying by $4/3$ estimates the lower bound.

For each block, we follow the same process to estimate lower and upper bounds. We then average these estimates across blocks, weighting by each block's share of officials. The resulting estimates of $\underline{\theta}^{u=0}$ and $\overline{\theta}^{u=0}$ are -0.12 and -0.19, both statistically significant at the $\alpha = 0.05$ level. These results corroborate the finding of Hughes *et al.* (2020), showing substantively large implicit discrimination against senders with an Arab alias relative to those with a White alias, albeit among the particular subgroup of officials who do not block open tracking.

4. Estimating causal effects on final outcomes

When the final outcome is also of interest, researchers often measure opening because they are interested in effects among the individuals who would actually receive the treatment message. A crucial feature of these experiments is that the treatment or control conditions are conveyed only in the bodies of emails. Information available to subjects before opening (e.g., in the email address or subject line) is identical across treatment and control conditions. This feature is codified in the assumption below.

Assumption 5. (Opening independent of treatment) *For all* $i = 1, \ldots, N$ *units,* $m_i(1) = m_i(0)$.

Because $m_i(1) = m_i(0)$ under Assumption 5, the opening of an email does not depend on treatment and, hence, is equivalent to a fixed baseline covariate. Therefore, we write an individual's email opening or not as m_i and the collection of all such values over all N subjects by $\boldsymbol{m}$, now with the dependence of opening on treatment assignment removed. Under this assumption, the proportion of openers is the same as π_1, which is what we now use to denote the proportion of openers, i.e., $(1/N)\sum_{i=1}^{n} m_i$.

Assumption 5 implies that every individual falls into one of two categories: Always-Openers or Never-Openers. Therefore, we now refer to the Always-Openers as Openers and the Never-Openers as Non-Openers. The ATE on final outcomes among Openers is then defined as

$$\tau^{m=1} := \left(\sum_{i=1}^{N} m_i\right)^{-1} \sum_{i=1}^{N} m_i \left[y_i(1) - y_i(0)\right]. \tag{13}$$

Assumption 5 is unlikely to hold in many studies. It may be especially tenuous in audit experiments that aim to detect discrimination based on racially distinctive names in email addresses (Leavitt and Rivera-Burgos, 2024). For example, in the experiment by Hughes *et al.* (2020) discussed above, Assumption 5 is implausible because officials can observe the senders' names without first opening the emails.

In many other message-based experiments, however, Assumption 5 is plausible, particularly when experimental conditions manifest in only the body text. In practice, researchers also frequently design treatments to ensure this assumption is satisfied. For example, Schiff and Schiff (2023), p. 826 note that they ensured "the symmetry of the emails before opening (e.g., same email subject line)," and Incerti (2024), p. 1605 reports that "[a]ll treatments included identical subject lines and preview texts to ensure equal compliance rates across treatment arms." That said, email technologies continue to evolve, and researchers seeking to satisfy Assumption 5 must carefully design experiments that account for variation across subjects' devices and email clients.

The following assumption is highly plausible—indeed, trivially true—under Assumption 5 and is central to deriving the ATE on final outcomes among Openers.

Assumption 6. (No effect among Non-Openers) *For all* $i = 1, \ldots, N$ *units with* $m_i = 0$, $y_i(1) - y_i(0) = 0$.

To see why Assumption 6 follows from Assumption 5, suppose the final outcome measures a behavior, such as participation in a city council meeting (Incerti, 2024), in which a positive outcome does not require first opening the email. For Non-Openers, treatment cannot affect this behavior because the information available without opening is identical across conditions (which is what justifies Assumption 5 in the first place). Moreover, even if one thought that the mere act of being sent an email—adding to inbox clutter or triggering a phone vibration—could influence the outcome independently of the message itself, Assumption 6 would still be plausible because those message-independent features are identical across treatment and control.

Researchers typically estimate the ATE on final outcomes among Openers, as defined in (13), using two main strategies. The first, employed by Moy (2021) and Schiff and Schiff (2023), follows the standard approach for randomized experiments with one-sided noncompliance (see Gerber and Green, 2012, Chapter 5, pp. 131–171). The second strategy conditions directly on the subjects recorded as Openers (Incerti, 2024; Lee, 2024). We now consider each approach in turn.

4.1. Message opening as imperfect compliance

Analogous to experiments with one-sided noncompliance, the ATE among Openers can be interpreted as the complier average causal effect (CACE). Compliers—unlike Always-Takers, Never-Takers, and Defiers—receive the treatment if and only if assigned to it. When treatment receipt is

defined as opening the message (as in Moy 2021 and Schiff and Schiff 2023), all Openers must be Compliers because opening under control yields only the control message, ruling out Always-Takers. Non-Openers, by contrast, must all be Never-Takers: When assigned to treatment, they never receive the treatment message, and under control they likewise do not (hence, no Defiers). Thus, every subject is either a Complier or a Never-Taker, corresponding to the one-sided noncompliance setting in which Always-Takers and Defiers are absent.

Also analogous to experiments with one-sided noncompliance, Proposition 4.1 shows that the average effect among Openers in (2)—equivalently, the CACE—can be expressed as the ratio of two quantities: the average effect on the final outcome among all subjects and the proportion of Openers.

Proposition 4.1. *Under Assumptions 1, 5 and 6 and supposing that* $\pi_1 > 0$*, the ATE among Openers—equivalently, the ATE among Compliers—is*

$$\frac{\tau}{\pi_1}. \tag{14}$$

The proof of this proposition relies on Assumption 6, which plays the role of the conventional excludability assumption by requiring that all Non-Openers (Never-Takers) have zero treatment effect. This proposition therefore provides a new formal justification for the CACE estimand in Moy (2021) and Schiff and Schiff (2023).

The problem that measurement error poses for estimation of the target in (14) has to do with estimation of the denominator, π_1. The usual instrumental variables regression via two-stage least squares, adopted in, e.g., Moy (2021), essentially estimates π_1 through the first term of the Difference-in-Means in (4), which—under Assumptions 3 and 5—reduces to $(1/n_1)\boldsymbol{Z}^\top \tilde{\boldsymbol{m}}$. Under CRA in Assumption 2, the expectation of this estimator, $(1/n_1)\boldsymbol{Z}^\top \tilde{\boldsymbol{m}}$, is equal to the overall proportion of measurable openers, $\tilde{\pi}_1$.

However, an insight from this paper is that the proportion of measurable openers does not require estimation because it can be directly calculated. Under Assumptions 3 and 5, measurable opens remain fixed across assignments. Thus, we can express the proportion of measurable openers as

$$\tilde{\pi}_1 = (1/N)\sum_{i=1}^{N} \tilde{m}_i, \tag{15}$$

which is composed of only observable quantities without dependence on the individual treatment assignment variables.

This proportion of measurable openers, $\tilde{\pi}_1$, must be less than or equal to π_1 under Assumption 3. As a result, the denominator in (15) is smaller than the true proportion of openers. Therefore, dividing the Difference-in-Means in (3) by the proportion of measurable openers will, in expectation, overstate the magnitude of the ATE among openers.

Nevertheless, as Leavitt and Rivera-Burgos (2024) also show, researchers can assess the sensitivity of estimates over the possible values of π_1. The bounds on π_1 can be tightened by the observed outcomes under Assumption 4. If it is impossible to have a positive response (e.g., a reply to an email) without first opening the email, any individuals with $\tilde{m}_i = 0$ who replied to the email must have $m_i = 1$. Incorporating this information implies a lower bound of the proportion of openers given by

$$\left(\frac{1}{N}\right)\left[\sum_{i=1}^{N} \tilde{m}_i + \sum_{i=1}^{N}(1-\tilde{m}_i)\left(z_i y_i(1) + (1-z_i)y_i(0)\right)\right]. \tag{16}$$

Hence, researchers can estimate the ATE among openers via the estimator in (3) divided by the possible values of π_1, ranging from the lower bound in (16) to 1.

4.2. Conditioning on measurable openers

Researchers can not only directly calculate the proportion of measurable openers; they can also discern exactly *which* subjects are measurable openers. With this information, it would be straightforward to estimate the ATE among openers in (14) by conditioning on measurable openers before employing the Difference-in-Means in (3). We write this post-stratified estimator as

$$\begin{aligned}\hat{\tau}^{\text{Open}}\left(\boldsymbol{Z}, \tilde{\boldsymbol{m}}, \boldsymbol{y}(\boldsymbol{Z})\right) := &\left(\frac{1}{\boldsymbol{Z}^{\top}\tilde{\boldsymbol{m}}}\right)\boldsymbol{Z}^{\top}\left(\tilde{\boldsymbol{m}} \odot \boldsymbol{y}\left(\boldsymbol{Z}\right)\right) \\ &- \left(\frac{1}{(\boldsymbol{1}-\boldsymbol{Z})^{\top}\tilde{\boldsymbol{m}}}\right)(\boldsymbol{1}-\boldsymbol{Z})^{\top}\left(\tilde{\boldsymbol{m}} \odot \boldsymbol{y}\left(\boldsymbol{Z}\right)\right),\end{aligned} \tag{17}$$

where $\odot$ denotes the element-wise (Hadamard) product of two matrices of the same dimension, which produces another matrix of the same dimension. This post-stratified estimator is the strategy that, e.g., Incerti (2024) and Lee (2024) employ, which is standard in placebo-controlled designs (Nickerson, 2008b; Gerber *et al.*, 2010).

Proposition 4.2 shows that, although the post-stratified estimator in (17) can be biased for the ATE among openers in (14), this estimator is unbiased for the ATE among openers who do not block open tracking.

Proposition 4.2. *Under Assumptions 1–3 and 5–6, the post-stratified Difference-in-Means in (17) is equal, in expectation, to the ATE on the final outcome among openers who do not block open tracking, i.e.,*

$$\mathrm{E}\left[\hat{\tau}^{Open}\left(\boldsymbol{Z}, \tilde{\boldsymbol{m}}, \boldsymbol{y}(\boldsymbol{Z})\right)\right] = \left(\sum_{i=1}^{N}(1-u_i)m_i\right)^{-1}\sum_{i=1}^{N}(1-u_i)m_i\tau_i. \tag{18}$$

To reiterate, this proposition, like Proposition 4.1, relies on Assumptions 5 and 6 in which opening is independent of treatment, and there are no effects among Non-Openers. The resulting subgroup ATE on the right-hand side of (18) can be substantively meaningful, particularly in settings where measurable openers constitute a large share of all openers.

4.3. Empirical application: experiment on social pressure primes

A randomized experiment from Moy (2021) consists of emails requesting public records to city executives across all 50 states. Each message came from the same sender (Bryant J. Moy, then at Washington University in St. Louis) with the same subject line. The body of the email varied by condition: a *duty* prime mentioning the obligation to be responsive to the public, a *peer effects* prime mentioning requests to other executives and the public reporting of responses, or a pure *control* with no prime. The study found evidence for a negative ATE of the peer effects prime, consistent with a potential "backfire" response to peer pressure (Ringold, 2002; Gerber *et al.*, 2008; Panagopoulos, 2014a; 2014b; Terechshenko *et al.*, 2019)

In our analysis of the data from this experiment, we condition on the 940 city executives assigned to either the *peer effects* or pure *control* conditions. The lower bound of the proportion of Openers is the share marked as opening the email or marked as not opening but replying, which is approximately 0.78. The Difference-in-Means estimate of the ATE on replies is roughly −0.07. As Proposition 4.1 shows, this Difference-in-Means corresponds to the lower bound (in magnitude) of the ATE on replies among Openers. To estimate the upper bound, we divide the Difference-in-Means by the proportion of measurable Openers (0.78). This yields an estimate of roughly −0.09, a substantively meaningful difference of two percentage points in magnitude relative to the lower bound.

In assessing statistical significance, our approach differs slightly from that of Moy (2021). Under Assumptions 3 and 5, the proportion of Openers is fixed across assignments. This property makes

variance estimation simpler via the approach in Leavitt and Rivera-Burgos (2024), Eq. 18, p. 457 because the estimator of the ATE among Openers need not be a ratio of two random quantities. This distinction, while subtle, is important for inference.

Finally, we also implement an alternative approach that conditions directly on measurable Openers via the estimator in (17), rather than dividing by the proportion of measurable Openers. Using this approach, we obtain a similar estimate of about −0.09. This estimate can be interpreted as the ATE among Openers who do not block open tracking. While the value is nearly identical to the upper bound estimate of the ATE among Openers, in general the two approaches may yield different results.

5. Discussion and conclusion

This paper has demonstrated how measurement error poses problems for two methods that researchers use to incorporate intermediary variables in message-based experiments. When opening itself is an outcome of interest, measurement error implies that the canonical estimator of the ATE on opening can be biased. In other settings, researchers may be interested in the final outcome, pertaining to either online or offline political behavior. Measurement of opening enables researchers to estimate the ATE among openers (who are also Compliers in the standard instrumental variable framework). However, measurement error can lead to biased estimates in this setting, too.

We show that, despite these issues, researchers can still draw reliable inferences about important causal targets. When opening is the outcome of interest, researchers can estimate informative bounds of the ATE among individuals who do not block open tracking. When the ATE on the final outcome among openers is of interest, researchers can estimate informative bounds of this ATE, and can assess sensitivity to varying proportions of Openers consistent with the observed data. Moreover, the common approach of conditioning on measurable openers is unbiased for the ATE among openers who do not block open tracking.

These results explicate the actual targets that estimators deployed in the literature are able to unbiasedly estimate. In addition, this paper shows how researchers can improve upon existing practice by assessing sensitivity to varying assumptions about the proportion of openers. Nevertheless, crucial open questions remain.

One is an empirical question about how pervasive blocking of open tracking is among common experimental subjects, such as state bureaucrats, employers, voters, etc. For example, do public officials tend to have outdated email servers that may be less likely to block open tracking? Future research might benefit from empirical answers to this question, generated via clever experimental designs.

Additional open questions pertain to how this paper's framework translates to message-based experiments conducted via technologies other than email. In some alternative settings, this paper's framework is readily transferable. For example, in experiments conducted via physical letters, as in Gaikwad and Nellis (2021), there are presumably no measures of whether experimental subjects open the envelopes addressed to them. Nevertheless, if one is interested in the ATE among openers, the sensitivity analysis to the proportion of openers, which can be bounded from below by the proportion of replies to the letters, can be used to estimate potentially informative bounds of this quantity. The same logic applies in messages delivered via social media platforms, which allow for measurement of opening to varying degrees. In other settings, however, the nature of measurement error might be quite different (as in measurement of text message opening in, e.g., Chivers and Barnes, 2018). Nevertheless, this paper's framework underscores the importance of addressing measurement errors in intermediary variables and charts a path forward as these measures become increasingly common across various settings.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2025.10082. To obtain replication material for this article, https://doi.org/10.7910/DVN/X3CORT.

Conflicts of interest. The authors state no conflicts of interest.

References

Adams WC and Smith DJ (1980) Effects of telephone canvassing on turnout and preferences: A field experiment. *The Public Opinion Quarterly* **44**, 389–395.

Bail CA, Argyle LP, Brown TW, Bumpus JP, Chen H, Hunzaker MBF, Lee J, Mann M, Merhout F and Volfovsky A. (2018) Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences of the United States of America* **115**, 9216–9221.

Bennion EA and Nickerson D (2011) The cost of convenience: An experiment showing e-mail outreach decreases voter registration. *Political Research Quarterly* **64**, 858–869.

Bergner C, Desmarais BA and Hird JA (2019) Speaking truth in power: Scientific evidence as motivation for policy activism. *Journal of Behavioral Public Administration* **2**, 1–11.

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

Bertrand M and Mullainathan S (2004) Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review* **94**, 991–1013.

Calfano B (2019) Power lines: Unobtrusive assessment of e-mail subject line impact on organization website use. *Journal of Political Marketing* **18**, 179–195.

Chivers B and Barnes G (2018) Sorry, wrong number: Tracking court attendance targeting through testing a 'nudge' text. *Cambridge Journal of Evidence-Based Policing* **2**, 4–34.

Coppock A, Guess A and Ternovski J (2016) When treatments are tweets: A network mobilization experiment over Twitter. *Political Behavior* **38**, 105–128.

Crabtree C. 2018. *An introduction to conducting email audit studies*. *Methodos Series*, Chapter 5, In SM Gaddis Ed., 14, Cham, Switzerland: Springer pp.103–117.

Devine PG (1989) Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology* **56**, 5–18.

Frangakis CE and Rubin DB (2002) Principal stratification in causal inference. *Biometrics* **58**, 21–29.

Gaikwad N and Nellis G (2021) Do politicians discriminate against internal migrants? Evidence from nationwide field experiments in India. *American Journal of Political Science* **65**, 790–806.

Gaynor SW and Gimpel JG (2024) Small donor contributions in response to email outreach by a political campaign. *Journal of Political Marketing* **23**, 51–73.

Gerber AS and Green DP (2012) *Field Experiments: Design, Analysis, and Interpretation.* New York, NY: W.W. Norton.

Gerber AS, Green DP, Kaplan EH and Kern HL (2010) Baseline, placebo, and treatment: Efficient estimation for three-group experiments. *Political Analysis* **18**, 297–315.

Gerber AS, Green DP and Larimer CW (2008) Social pressure and voter turnout: Evidence from a large-scale field experiment. *The American Political Science Review* **102**, 33–48.

Holland PW (1986) Statistics and causal inference. *Journal of the American Statistical Association* **81**, 945–960.

Hughes DA, Gell-Redman M, Crabtree C, Krishnaswami N, Rodenberger D and Monge G (2020) Persistent bias among local election officials. *Journal of Experimental Political Science* **7**, 179–187.

Incerti T (2024) Countering capture in local politics: Evidence from eight field experiments. *The Journal of Politics* **86**, 1603–1607.

Leavitt T and Rivera-Burgos V (2024) Audit experiments of racial discrimination and the importance of symmetry in exposure to cues. *Political Analysis* **32**, 445–462.

Lee DDI (2024) *Political candidacy as a spectrum: How minorities develop political ambition.* OSF Preprint. https://doi.org/10.31219/osf.io/bxrv3.

Malhotra N, Michelson MR and Valenzuela AA (2012) Emails from official sources can increase turnout. *Quarterly Journal of Political Science* **7**, 321–332.

McClendon GH (2014) Social esteem and participation in contentious politics: A field experiment at an LGBT pride rally. *American Journal of Political Science* **58**, 279–290.

Moy BJ (2021) Can social pressure foster responsiveness? An open records field experiment with mayoral offices. *Journal of Experimental Political Science* **8**, 117–127.

Neyman J (1923) Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* **10**, 1–51.

Nickerson D (2008a) Does email boost turnout?. *Quarterly Journal of Political Science* **2**, 369–379.

Nickerson D (2008b) Is voting contagious? Evidence from two field experiments. *The American Political Science Review* **102**, 49–57.

Panagopoulos C (2014a) I've got my eyes on you: Implicit social-pressure cues and prosocial behavior. *Political Psychology* **35**, 23–33.

Panagopoulos C (2014b) Watchful eyes: implicit observability cues and voting. *Evolution and Human Behavior* **35**, 279–284.

Persian R, Adityawarman GP, Bogiatzis-Gibbons D, Kurniawan MH, Subroto G, Mustakim M, Scheunemann L, Gandy K and Sutherland A (2023) Behavioural prompts to increase early filing of tax returns: A population-level randomised controlled trial of 11.2 million taxpayers in Indonesia. *Behavioural Public Policy* **7**, 701–720.

Ringold D (2002) Boomerang effects in response to public health interventions: Some unintended consequences in the alcoholic beverage market. *Journal of Consumer Policy* **25**, 27–63.

Rivera MU, Hughes DA and Gell-Redman M (2023) Email mobilization messages suppress turnout among Black and Latino voters: Experimental evidence from the 2016 general election. *Journal of Experimental Political Science* **10**, 261–266.

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688.

Schiff DS and Schiff KJ (2023) Narratives and expert information in agenda-setting: Experimental evidence on state legislator engagement with artificial intelligence policy. *Policy Studies Journal* **51**, 817–842.

Terechshenko Z, Crabtree C, Eck K and Fariss CJ (2019) Evaluating the influence of international norms and shaming on state respect for rights: An audit experiment with foreign embassies. *International Interactions* **45**, 720–735.

Van Remoortere A, Vermeer S and Kruikemeier S (2024) Contact us! An audit study to examine the responsiveness of political elites on social media during a Dutch election. *Electoral Studies* **90**, 102815.

Woollaston-Webber V (2016) Twitter adds read receipts to direct messages (accessed 13 August 2025).

Cite this article: Leavitt T and Rivera-Burgos V (2026) Navigating the mismeasurement of intermediary variables in message-based experiments. *Political Science Research and Methods*, 1–14. https://doi.org/10.1017/psrm.2025.10082

Supplementary Appendix for "Navigating the Mismeasurement of Intermediary Variables in Message-Based Experiments"

Thomas Leavitt and Viviana Rivera-Burgos

A Proofs

A.1 Lemma S.1

Lemma S.1. *Under Assumption 3, SUTVA in Assumption 1 implies that SUTVA also holds for measures of opening, i.e., for all $i = 1, \ldots, N$ individuals (with either $u_i = 0$ or $u_i = 1$), $\tilde{m}_i(\boldsymbol{z})$ takes on a fixed value, $\tilde{m}_i(1)$, for all $\boldsymbol{z} : z_i = 1$ and takes on a fixed value, $\tilde{m}_i(0)$, for all $\boldsymbol{z} : z_i = 0$.*

Proof. Assumption 3 implies that, for all $i = 1, \ldots, N$ with $u_i = 0$, $\tilde{m}_i(\boldsymbol{z}) = m_i(\boldsymbol{z})$ for all $\boldsymbol{z} \in \{0, 1\}^N$. Therefore, Assumption 1 implies that, for all $i = 1, \ldots, N$ units with $u_i = 0$, $\tilde{m}_i(\boldsymbol{z})$ also takes on a fixed value, $m_i(1)$ for all $\boldsymbol{z} : z_i = 1$ and another, $m_i(0)$, for all $\boldsymbol{z} : z_i = 0$. For all $i = 1, \ldots, N$ units with $u_i = 1$, note that Assumption 3 imply that $\tilde{m}_i(\boldsymbol{z}) = 0$ for all $\boldsymbol{z} \in \{0, 1\}^N$, thereby completing the proof. □

A.2 Proof of Proposition 1

Proof. By Lemma S.1, let $\tilde{\theta}_i$ denote the individual effect when measures of opening (not actual opening) is the outcome, and write the ATE for measures of opens as

$$\begin{aligned}\tilde{\theta} &\coloneqq \left(\frac{1}{N}\right) \sum_{i=1}^{N} \tilde{\theta}_i \\ &= \left(\frac{N^{u=1}}{N}\right) \left(\frac{1}{N^{u=1}}\right) \sum_{i=1}^{N} u_i \tilde{\theta}_i + \left(\frac{N^{u=0}}{N}\right) \left(\frac{1}{N^{u=0}}\right) \sum_{i=1}^{N} (1 - u_i) \tilde{\theta}_i. \end{aligned} \tag{19}$$

Under Assumption 3, the first term of (19) is equal to 0 and $\tilde{\theta}_i$ in the second term can be replaced with θ_i since for all $i = 1, \ldots, N$ with $u_i = 0$, $\tilde{m}_i(z) = m_i(z)$ for $z = 1$ and $z = 0$. Consequently,

$$\tilde{\theta} = \left(\frac{N^{u=0}}{N}\right)\left(\frac{1}{N^{u=0}}\right)\sum_{i=1}^{N}(1-u_i)\theta_i, \tag{20}$$

which, under CRA in Assumption 2, is equal to the expected value of the Difference-in-Means in (4).

We can therefore express the bias of (4) for θ as

$$\begin{aligned}
\mathrm{E}\left[\hat{\tau}\left(\boldsymbol{Z}, \tilde{\boldsymbol{m}}\left(\boldsymbol{Z}\right)\right)\right] - \theta &= \tilde{\theta} - \theta \\
&= \left(\frac{N^{u=0}}{N}\right)\left(\frac{1}{N^{u=0}}\right)\sum_{i=1}^{N}(1-u_i)\theta_i - \left(\frac{N^{u=1}}{N}\right)\frac{1}{N^{u=1}}\sum_{i=1}^{N}u_i\theta_i \\
&\quad - \left(\frac{N^{u=0}}{N}\right)\frac{1}{N^{u=0}}\sum_{i=1}^{N}(1-u_i)\theta_i \\
&= -\left(\frac{N^{u=1}}{N}\right)\frac{1}{N^{u=1}}\sum_{i=1}^{N}u_i\theta_i,
\end{aligned}$$

which, by the definitions of $\bar{u}$ and $\theta^{u=1}$, is

$$-\bar{u}\,\theta^{u=1},$$

thereby completing the proof. □

A.3 Proof of Proposition 2

Proof. First note that, under Assumptions 1 and 3,

$$\begin{aligned}
\tilde{\pi}_{10} &= 1 - \tilde{\pi}_0 - \tilde{\pi}_{00} = \left(\frac{1}{N}\right)\sum_{i=1}^{N}(1-u_i)\mathbb{1}\left\{m_i(1) = 1,\, m_i(0) = 0\right\} \\
\tilde{\pi}_{01} &= \tilde{\pi}_0 - \tilde{\pi}_1 + \tilde{\pi}_{10} = \left(\frac{1}{N}\right)\sum_{i=1}^{N}(1-u_i)\mathbb{1}\left\{m_i(1) = 0,\, m_i(0) = 1\right\}.
\end{aligned}$$

Hence, we can express the respective sums of measurable Only-Treated-Openers and measurable Only-Control-Openers as

$$N\tilde{\pi}_{10} = \sum_{i=1}^{N}(1-u_i)\mathbb{1}\left\{m_i(1)=1,\, m_i(0)=0\right\} \text{ and}$$
$$N\tilde{\pi}_{01} = \sum_{i=1}^{N}(1-u_i)\mathbb{1}\left\{m_i(1)=0,\, m_i(0)=1\right\}.$$

Under Assumption 3, units with $\tilde{m}_i(z)=1$ for $z=1$ or $z=0$ consist of only units with $u_i=0$, so we can express the conditional ATE in (6) as

$$\theta^{u=0} = \left(\frac{N}{N^{u=0}}\right)\left(\tilde{\pi}_{10}-\tilde{\pi}_{01}\right). \tag{21}$$

The quantity $N^{u=0}$ is unknown, but it can be bounded: If all measurable Never-Openers were actual Never-Openers with $u_i=0$, then $N^{u=0}$ would be equal to N. By contrast, if all measurable Never-Openers are individuals with $u_i=1$, then $N^{u=0}$ would be

$$N\left(\tilde{\pi}_{11}+\tilde{\pi}_{10}+\tilde{\pi}_{01}\right) = N\left(1-\tilde{\pi}_{00}\right),$$

which is the lower bound. Dividing $\theta^{u=0}$ in (21) by the upper bound of $N^{u=0}$ yields the lower bound (in magnitude) of $\theta^{u=0}$ and dividing $\theta^{u=0}$ by the lower bound of $N^{u=0}$ yields the upper bound (in magnitude) of $\theta^{u=0}$. Hence, the lower- and upper bounds of $\theta^{u=0}$, denoted by $\underline{\theta}^{u=0}$ and $\overline{\theta}^{u=0}$, respectively, are

$$\underline{\theta}^{u=0} = \left(\frac{N}{N^{u=0}}\right)\left(\tilde{\pi}_{10}-\tilde{\pi}_{01}\right) = \left(\frac{N}{N}\right)\left(\tilde{\pi}_{10}-\tilde{\pi}_{01}\right) = \tilde{\pi}_{10}-\tilde{\pi}_{01}$$

$$\overline{\theta}^{u=0} = \left(\frac{N}{N^{u=0}}\right)\left(\tilde{\pi}_{10}-\tilde{\pi}_{01}\right) = \left(\frac{N}{N\left(1-\tilde{\pi}_{00}\right)}\right)\left(\tilde{\pi}_{10}-\tilde{\pi}_{01}\right) = \left(\frac{1}{1-\tilde{\pi}_{00}}\right)\left(\tilde{\pi}_{10}-\tilde{\pi}_{01}\right),$$

thereby completing the proof. □

A.4 Proof of Proposition 3

Under Assumptions 1 and 5, the ATE among openers is

$$\left(\frac{1}{N\pi_1}\right)\sum_{i=1}^{N} m_i\tau_i. \tag{22}$$

Assumption 6 then implies that the sum of individual effects among openers is the sum of individual effects among all individuals (openers and non-openers). Hence, the ATE among openers in (22) is

$$\left(\frac{1}{N\pi_1}\right)\sum_{i=1}^{N} \tau_i.$$

Then, by the definition of τ in (1), the ATE among openers in (22) is

$$\frac{\tau}{\pi_1},$$

which completes the proof.

A.5 Proof of Proposition 4

Proof. Under Assumptions 1, 2 and 5, the expected value of the Difference-in-Means among measurable openers in (17) is equal to the ATE among measurable openers:

$$\frac{1}{N\tilde{\pi}_1}\sum_{i=1}^{N} \tilde{m}_i\tau_i. \tag{23}$$

Assumption 3 then implies that the ATE among measurable openers in (23) is

$$\left(\sum_{i=1}^{N}(1-u_i)m_i\right)^{-1}\sum_{i=1}^{N}(1-u_i)m_i\tau_i. \tag{24}$$

□

A.6 Bias from Using Replies as an Auxiliary Measure of Opening

Under Assumption 4, any positive response in terms of the final outcome implies that the email was opened. We now formally write the estimator incorporating the final outcome under Assumption 4 as

$$\begin{aligned}\hat{\theta}\left(\boldsymbol{Z}, \tilde{\boldsymbol{m}}\left(\boldsymbol{Z}\right), \boldsymbol{y}\left(\boldsymbol{Z}\right)\right) = & \left(\frac{1}{n_1}\right) \boldsymbol{Z}^{\top}\left[\tilde{\boldsymbol{m}}\left(\boldsymbol{Z}\right) + \left(\mathbf{1} - \tilde{\boldsymbol{m}}(\boldsymbol{Z})\right) \odot \boldsymbol{y}(\boldsymbol{Z})\right] \\ & - \left(\frac{1}{n_0}\right) \left(\mathbf{1} - \boldsymbol{Z}\right)^{\top}\left[\tilde{\boldsymbol{m}}\left(\boldsymbol{Z}\right) + \left(\mathbf{1} - \tilde{\boldsymbol{m}}(\boldsymbol{Z})\right) \odot \boldsymbol{y}(\boldsymbol{Z})\right],\end{aligned} \tag{25}$$

where $\tilde{\boldsymbol{m}}(\boldsymbol{Z}), \boldsymbol{y}(\boldsymbol{Z}) \in \{0,1\}^N$, i.e., both are binary vectors of length N. Unlike the estimator in (4), this estimator in (25) makes sure to count as opens the emails that received a positive final response (e.g., a reply to an email) but were measured as not opened. Note that this estimator could be expressed as a conventional Difference-in-Means in which the outcome is an augmented measure of opening that incorporates replies: $\check{m}_i(\boldsymbol{Z}) = \tilde{m}_i(\boldsymbol{Z}) + (1 - \tilde{m}_i(\boldsymbol{Z}))y_i(\boldsymbol{Z})$.

This estimator in (25) can also be biased for θ in (2) and, unlike the Difference-in-Means in (4), cannot be interpreted as a conservative (i.e., lower bound in magnitude) estimate of the ATE among a meaningful subgroup of subjects. To see why, suppose the final outcome is binary, $\boldsymbol{y}(\boldsymbol{z}) \in \{0,1\}^N$ for all $\boldsymbol{z} \in \{0,1\}^N$, and Assumption 4 holds. Then define principal strata on the basis of the final outcome, analogous to those in Table 1 for opens, as follows:

	$y_i(1) = 1$	$y_i(1) = 0$
$y_i(0) = 1$	*Always-Responder*	*Only-Control-Responder*
$y_i(0) = 0$	*Only-Treatment-Responder*	*Never-Responder*

Table 3: Principal Strata of experimental subjects in terms of final outcomes under treatment and control

We define the proportion of units in each principal stratum in Table 3 as

$$\varphi_{11} \coloneqq \left(\frac{1}{N}\right)\sum_{i=1}^{N}\mathbb{1}\left\{y_i(1)=1,\, y_i(0)=1\right\}, \quad \varphi_{10} \coloneqq \left(\frac{1}{N}\right)\sum_{i=1}^{N}\mathbb{1}\left\{y_i(1)=1,\, y_i(0)=0\right\},$$
$$\varphi_{01} \coloneqq \left(\frac{1}{N}\right)\sum_{i=1}^{N}\mathbb{1}\left\{y_i(1)=0,\, y_i(0)=1\right\}, \quad \varphi_{00} \coloneqq \left(\frac{1}{N}\right)\sum_{i=1}^{N}\mathbb{1}\left\{y_i(1)=0,\, y_i(0)=0\right\}.$$

We also let $\bar{m}_{11}(z)$, $\bar{m}_{10}(z)$, $\bar{m}_{01}(z)$ and $\bar{m}_{00}(z)$ denote the average potential opens under treatment ($z=1$) or control ($z=0$) among individuals belonging to the Always-Responder, Only-Treatment-Responder, Only-Control-Responder and Never-Responder strata, respectively. Consequently, the ATE in (2) can be expressed as

$$\begin{aligned}\theta = {} & \varphi_{11}[\bar{m}_{11}(1)-\bar{m}_{11}(0)] + \varphi_{10}[\bar{m}_{10}(1)-\bar{m}_{10}(0)] \\ & + \varphi_{01}[\bar{m}_{01}(1)-\bar{m}_{01}(0)] + \varphi_{00}[\bar{m}_{00}(1)-\bar{m}_{00}(0)].\end{aligned} \tag{26}$$

With this setup, Proposition S.1 below shows the bias of the estimator in (25) for the ATE in (2).

Proposition S.1. *Under Assumptions 1 – 4, the bias of the estimator in* (25) *for* θ *in (2) is*

$$\begin{aligned}\mathrm{E}\left[\hat{\theta}\left(\boldsymbol{Z}, \tilde{\boldsymbol{m}}\left(\boldsymbol{Z}\right), \boldsymbol{y}\left(\boldsymbol{Z}\right)\right)\right] - \theta = {} & \bar{u}(\varphi_{10}^{u=1}\bar{m}_{10}^{u=1}(0) - \varphi_{01}^{u=1}\bar{m}_{01}^{u=1}(1) \\ & + \varphi_{00}^{u=1}[\bar{m}_{00}^{u=1}(1) - \bar{m}_{00}^{u=1}(0)]),\end{aligned} \tag{27}$$

where, as before, the superscript $u=1$ *denotes quantities among subjects who block open tracking, i.e., among all* $i=1,\ldots,N$ *subjects with* $u_i=1$.

Proof. Under Assumptions 1 and 3, the Difference-in-Means in (25) can be expressed as

$$\left(\frac{1}{n_1}\right)\sum_{i=1}^{N} Z_i\left[\tilde{m}_i(1) + [1-\tilde{m}_i(1)]y_i(1)\right] - \left(\frac{1}{n_0}\right)\sum_{i=1}^{N}(1-Z_i)\left[\tilde{m}_i(0) + [1-\tilde{m}_i(0)]y_i(0)\right],$$

which, by the linearity of expectation, has an expected value of

$$\begin{aligned} &\left(\frac{1}{n_1}\right)\sum_{i=1}^{N}\tilde{m}_i(1)+[1-\tilde{m}_i(1)]y_i(1)\,\mathrm{E}\left[Z_i\right] \\ -&\left(\frac{1}{n_0}\right)\sum_{i=1}^{N}\tilde{m}_i(0)+[1-\tilde{m}_i(0)]y_i(0)\,\mathrm{E}\left[(1-Z_i)\right]. \end{aligned} \tag{28}$$

Under CRA in Assumption 2, $\mathrm{E}[Z_i] = n_1/N$ and $\mathrm{E}[1-Z_i] = n_0/N$ for all $i = 1,\ldots,N$ units, thereby implying that (28) can be expressed as

$$\left(\frac{1}{n_1}\right)\left(\frac{n_1}{N}\right)\sum_{i=1}^{N}\tilde{m}_i(1)+[1-\tilde{m}_i(1)]y_i(1)-\left(\frac{1}{n_0}\right)\left(\frac{n_0}{N}\right)\sum_{i=1}^{N}\tilde{m}_i(0)+[1-\tilde{m}_i(0)]y_i(0),$$

yielding an expected value of the Difference-in-Means in (25) equal to

$$\left(\frac{1}{N}\right)\sum_{i=1}^{N}[\tilde{m}_i(1)-\tilde{m}_i(0)]+\left(\frac{1}{N}\right)\sum_{i=1}^{N}[1-\tilde{m}_i(1)]y_i(1)-[1-\tilde{m}_i(0)]y_i(0). \tag{29}$$

The first term of (29) is the ATE with measurable opens as the outcome given by $\tilde{\pi}_{10}-\tilde{\pi}_{01}$, which, under Assumption 3, is

$$(1-\bar{u})\,\theta^{u=0}.$$

Assumption 3 also implies that the second term of (29) consists of only those who block open tracking, i.e., all $i = 1,\ldots,N$ with $u_i = 1$, and Assumption 4 implies that the sum in the second term is the sum of individual effects (with replies as the outcome) among individuals who block open tracking, $N^{u=1}[\varphi_{10}^{u=1}-\varphi_{01}^{u=1}]$, where the superscript $u = 1$ denotes quantities among subjects who block open tracking.

Consequently, (29) is

$$(1-\bar{u})\,\theta^{u=0}+\bar{u}[\varphi_{10}^{u=1}-\varphi_{01}^{u=1}], \tag{30}$$

which, after noting that the ATE in (2) can be expressed as

$$\theta = (1 - \bar{u})\,\theta^{u=0} + \bar{u}\theta^{u=1}, \tag{31}$$

implies that the bias is

$$\bar{u}\left(\varphi_{10}^{u=1} - \varphi_{01}^{u=1} - \theta^{u=1}\right). \tag{32}$$

The conditional ATE, $\theta^{u=1}$, in (32) can then be decomposed as

$$\begin{aligned}
&\varphi_{11}^{u=1}\bar{m}_{11}^{u=1}(1) + \varphi_{10}^{u=1}\bar{m}_{10}^{u=1}(1) + \varphi_{01}^{u=1}\bar{m}_{01}^{u=1}(1) + \varphi_{00}^{u=1}\bar{m}_{00}^{u=1}(1) \\
&- \left[\varphi_{11}^{u=1}\bar{m}_{11}^{u=1}(0) + \varphi_{10}^{u=1}\bar{m}_{10}^{u=1}(0) + \varphi_{01}^{u=1}\bar{m}_{01}^{u=1}(0) + \varphi_{00}^{u=1}\bar{m}_{00}^{u=1}(0)\right],
\end{aligned}$$

which, under Assumption 4, is

$$\varphi_{10}^{u=1} - \varphi_{10}^{u=1}\bar{m}_{10}^{u=1}(0) + \varphi_{01}^{u=1}\bar{m}_{01}^{u=1}(1) - \varphi_{01}^{u=1} - \varphi_{00}^{u=1}(\bar{m}_{00}^{u=1}(1) - \bar{m}_{00}^{u=1}(0)). \tag{33}$$

Then substituting (33) for $\theta^{u=1}$ in the bias expression in (32) results in

$$\bar{u}\left(\varphi_{10}^{u=1}\bar{m}_{10}^{u=1}(0) - \varphi_{01}^{u=1}\bar{m}_{01}^{u=1}(1) + \varphi_{00}^{u=1}[\bar{m}_{00}^{u=1}(1) - \bar{m}_{00}^{u=1}(0)]\right),$$

which completes the proof. □

This expression for the bias is intuitive: The Difference-in-Means incorporating the final outcome in (25) has three blind spots: Control potential opens among Only-Treatment-Responders who block open tracking (in the first term of equation 27); treated potential opens among Only-Control-Responders who block open tracking (in the second term of equation 27); and Treatment and control potential opens among Never-Responders who block open tracking (in the third term of equation 27). These three blind spots are then

multiplied by the appropriate weights reflecting the proportion of individuals in the respective subgroups wherein these blind spots exist.

In addition, it is straightforward to derive the bias for the ATE on opening among subjects who do not block open tracking. From the expected value in (29) and the expression for the ATE in (31), this bias can be expressed as

$$\bar{u}[\varphi_{10}^{u=1} - \varphi_{01}^{u=1}]. \tag{34}$$

This term in (34) represents the ATE on replies among individuals who block open tracking, multiplied by the proportion of such individuals in the experimental population.