Bounding causal effects in survey experiments with noncompliance or inattention

Matthew Tyler[*]

February 17, 2025

Abstract

Survey experimentalists often want to estimate the effect of a treatment among respondents who actually “receive” treatment. To account for noncompliance and inattention, researchers frequently include post-treatment manipulation checks in their survey experiments. Unfortunately, current methods assume that compliance and attention can be measured without error. In reality, inattentive and noncompliant respondents are particularly prone to measurement errors, leading to biased causal effect estimates. In this paper, I develop a computational method for sharply bounding causal effects with manipulation checks that are prone to such measurement errors. To account for sampling variability, I also construct confidence intervals that achieve their nominal asymptotic coverage under a mild differentiability condition. The method for constructing confidence intervals solves the overcoverage problem for confidence intervals in existing optimization-based bounding methods and should be broadly applicable.

1 Introduction

Survey experiments have become a cornerstone of empirical research in political science, providing a powerful tool for understanding how different treatments influence respondents' attitudes and behaviors. The ability to manipulate key variables and observe their impact in a controlled setting makes survey experiments an invaluable approach for testing theories and generating robust empirical evidence across the subfields of political science. Indeed, a recent study found that survey experiments comprise about 20% of *all* recently published articles in top political science journals (Briggs et al. 2025). Other meta analyses have shown that survey experiments are the majority of preregistered studies in political science (**slough**).

Manipulation checks, including post-treatment attention and comprehension checks, play a crucial role in interpreting the results of survey experiments. These checks are designed

[*] Assistant Professor, Department of Political Science, Rice University, United States. Email: mdtyler@rice.edu

to verify that participants have actually received and understood the treatment as intended by the researchers. Without manipulation checks, it would be difficult to determine whether any observed effects can be attributed to the content manipulated by the survey experiment or the causal mechanisms identified anticipated by the researcher (Mutz 2021). By including questions that assess participants' comprehension and engagement with the treatment, researchers can confirm that the experimental manipulation has been successfully implemented and/or test causal mechanisms.

However, survey measurement error poses a significant challenge to the reliability of manipulation checks. Respondents may not always provide accurate answers to manipulation check questions due to various reasons, such as lack of attention or comprehension of either the manipulation check item(s) or the experimantal stimulus (e.g., a long vignette). Survey measurement error is pervasive in political science (Clayton et al. 2023; Blair, Chou, and Imai 2019; Westwood et al. 2022). In this case, measurement error can lead to incorrect classifications of respondents. Consequently, some respondents who were inattentive or otherwise were not manipulated by the treatment can mistaken as attentive or having been successfully manipulated. Thus, researchers cannot take manipulation check results at face value and should account for potential errors when interpreting their findings. At present, there is no method that accounts for measurement error in manipulation checks.

This paper contributes a new method for bounding causal effects in survey experiments that accounts for measurement error in manipulation checks. Like related methods (Lee 2009; Aronow, Baron, and Pinson 2019), the proposed method attempts to bound the average causal effect of the treatment among the subset of respondents who would be compliant in both the treatment and control conditions. However, unlike related methods, the proposed method can account for measurement error in the measure of compliance (i.e., the manipulation check). The proposed method is also more assumption-flexible in that the researcher can drop or add assumptions about the data-generating process to the bounding method. For example, the reseracher can perform a sensitivity analysis on the commonly invoked monotonicity assumption (Lee 2009), or the researcher can vary the assumed false positive rate of the screener.

To faciliate this methodological contribution to survey methods, this paper makes a second contribution by developing a new computational method for bounding parameters of models of categorical random variables with missing data. Following Duarte et al. (2024), known as `autobounds`, I represent the bounding problem as a pair of numerical optimization problems. However, I show that, in this setting, the bounding problem can be formulated as a pair of *convex* optimization problems, namely linear and second-order cone programs, which can be solved extremely quickly with Newton-method-like convergence guarantees (Boyd and Vandenberghe 2004). This computational approach enables researchers to perfor sensitivity analyses with many different assumptions and/or parameter values in a reasonable amount of time.

Furthermore, I derive confidence intervals for the bounds that achieve their nominal asymptotic coverage under a differentiability condition. The method for modifying the optimization problems to produce a confidence interval is similar to the "asymptotic bounds" approach of Duarte et al. (2024). However, I derive the relationship between the observed data confidence region radius and the resulting confidence interval width, enabling the calibration of the confidence interval to achieve the desired asymptotic coverage rate. This solves the overcoverage problem of existing methods (Duarte et al. 2024). The proof is surprisingly general, enabling researchers to derive calibrated confidence intervals for other optimization-based methods.

The rest of the paper is organized as follows. In Section 2, I introduce the framework for survey experiments with manipulation checks and measurement error. Section 3 presents the method for bounding causal effects of interest using linear programming. Section 4 extends the method to account for sampling variability and construct confidence intervals using second-order cone programming. Section 5 applies the method to two empirical examples, demonstrating its practical utility. Finally, I conclude with a summary of contributions and directions for future research.

2 Framework

2.1 Survey experiments with noncompliance

Consider a survey experiment sample with n respondents, who I treat as though they were drawn i.i.d. from an infinite superpopulation (Assumption A0). Each respondent in the sample is randomly assigned to a binary treatment condition $D \in \{0, 1\}$. The researcher is interested in the effect of the treatment condition on a categorical outcome $Y \in \{y_1, y_2, \ldots, y_K\}$ with K possible values.[1] Of course, with no further considerations, standard experimental methods can be used to quantify the average causal effect of D on Y.

Generally, the researcher is most interested in respondents who actually received treatment stimulus. I call these compliant respondents, meaning the respondent is attentive to and/or comprehends the stimulus and/or was successfully manipulated by the stimulus.[2] For example, in a survey experiment on media effects, the researcher may want to estimate the effect of being assigned to a particular media source on respondents' political attitudes. However, if some respondents do not actually read the assigned media source, then the effect of the treatment on the outcome is not actually the effect of the media source, but rather a composite effect of the media source and the decision to read it. To estimate the effect of the media source on attitudes, the researcher would ideally restrict their analysis to just those compliant respondents who would always read the media source.

To classify respondents as compliant or not, researchers often rely on manipulation checks as measures of compliance. Manipulation checks, which could include post-treatment attention and/or comprehension checks, are used to verify that the experimental manipulation has the intended effect on the participants. For example, if the experiment involves exposing participants to a particular message, a manipulation check might involve asking participants questions to confirm that they understood and internalized the message. Naturally, manipulation checks are measured post-treatment.

[1] In practice, most survey outcomes used in political science are categorical, either unordered or ordered. In exceptional cases, researchers can usually convert continuous outcomes into categorical outcomes by binning.

[2] Note that this definition of compliance is distinct from the notion of compliance used in the literature on instrumental variables.

In this paper, I refer to all types of manipulation checks as "screeners" for simplicity.[3] I treat screeners as being measured post-treatment since pre-treatment screeners can be represented as a special case of post-treatment screeners in my model (see below). The question of whether pre- or post-treatment screeners should be preferred over the other is outside the scope of this paper.[4]

Formally, I model the screener measure as a binary indicator $S \in \{0,1\}$, where $S = 1$ indicates that the respondent passed the screener and $S = 0$ indicates that they failed. This binary screener could potentially be a measure composed of multiple survey items.

Like almost any other type of survey measure, the screener is most likely an (imperfect) measure of the underlying concept of compliance that the screener seeks to measure. Let $A \in \{0,1\}$ denote the unobserved indicator of compliance. A respondent satisfies $A = 1$ if they have exerted enough effort to automatically pass the screener. Consequently, $A = 1$ implies $S = 1$. Respondents who don't exert enough effort to pass the screener automatically are considered noncompliant, $A = 0$.

Ideally, the screener would indicate compliance perfectly, $S = A$. However, due to survey measurement error, the screener could misclassify some respondents. For example, if the screener's survey item is a multiple choice question with one correct answer (screener pass if correct, screener fail otherwise), then some fraction of noncompliant respondents might accidentally pass the screener by chance. This would result in $A = 0$ even though $S = 1$ (i.e., a "false positive"). Note that, given how compliance is defined based on the screener measure, any measurement error can only result in false positives. Measurement error in this model cannot produce "false negatives" ($A = 1$ and $S = 0$).

Critically, compliance A is not observed. The researcher only observes the screener S, and so the researcher must rely on the screener as a proxy for compliance. The fact that the screener is an imperfect measure of compliance frustrates existing methods, which implicitly assume compliance is measured perfectly (c.f. Aronow, Baron, and Pinson 2019), and suggests a new approach might be necessary.

2.2 Potential outcomes and principal strata

To characterize the causal effect of the treatment stimulus among compliant respondents, I use a potential outcomes model. For respondent i, they are first assigned to a binary treatment $D_i \in \{0,1\}$. Then (simultaneously) their compliance status $A_i \in \{0,1\}$ is realized, their screener $S_i \in \{0,1\}$ is measured, and their categorical outcome $Y_i \in \{y_1, \ldots, y_K\}$ is measured. I impose the following (largely untestable) assumptions on this process.

Assumption A1 (SUTVA)**.** The Stable Unit Treatment Value Assumption (SUTVA) holds, meaning there is no interference between units and no hidden versions of the treatment. Formally, for any unit i, the realized values of the random variables are given by

$$A_i = A_i(D_i), \quad S_i = S_i(D_i), \quad Y_i = Y_i(D_i). \tag{1}$$

[3] In some cases, there is no clear distinction between a manipulation check and a post-treatment attention check anyway.

[4] In Study 1 below, I show that a pre-treatment screener fails to remove many noncompliant respondents from the sample that were caught by a post-treatment screener.

Assumption A2 (Random Assignment)**.** The treatment D is randomly assigned, independent of all potential outcomes.

$$D_i \perp\!\!\!\perp (A_i(0), A_i(1), S_i(0), S_i(1), Y_i(0), Y_i(1)). \tag{2}$$

Furthermore, $0 < P(D_i = 1) < 1$, meaning all respondents have a positive probability of being assigned to either condition.

Note that invoking SUTVA and random assignment is usually uncontentious in the survey experimental literature since the survey experiment setting is carefully controlled and respondents often take surveys independently. I leave extensions to survey experiments with interference to future work.

These baseline assumptions allow us to define the principal strata needed to characterize the causal estimand of interest. Based on their potential compliance statuses, each respondent can be categorized into one of four principal strata: always-compliant ($A_i(0) = A_i(1) = 1$), never-compliant ($A_i(0) = A_i(1) = 0$), activated-compliant ($A_i(0) = 0, A_i(1) = 1$), and suppressed-compliant ($A_i(0) = 1, A_i(1) = 0$). As alluded to above, the substantive effect of the treatment on the outcome is most scientifically interesting among the always-compliant stratum, where $A_i(1) = A_i(0) = 1$. These are the respondents who are attentive to and/or manipulated by the treatment as intended by the researcher *in both conditions*. By contrast, in the activated-compliant and suppressed-compliant strata, the effect of the treatment on the outcome is confounded by the effect of the treatment on compliance, a potentially strong mediator of the treatment's effect on the outcome. Likewise, the treatment effect among the always-noncompliant is ignored because it is unclear what respondents are reacting to (if they are reacting at at all) given they are inattentive and/or not manipulated.

Therefore, the causal estimand of interest is the average treatment effect among the always-compliant stratum (ATAC).

$$\text{ATAC} = E[Y_i(1) - Y_i(0) \mid A_i(1) = 1, A_i(0) = 1] \tag{3}$$

This estimand captures the effect of the treatment on the outcome while holding compliance fixed across conditions. In contrast, the basic average treatment effect (ATE) is contaminated by the other principal strata.

Unfortunately, there are two major difficulties to identifying (and thus estimating) the ATAC. First, the researcher can only observe one set of potential outcomes per respondent (i.e., the fundamental problem of causal inference). It is therefore impossible to assign respondents to principal strata to calculate the conditional expectation even when compliance A_i is measured without error. Second, compliance A_i is generally measured with error, which would make it difficult to assign respondents to principal strata even if the fundamental problem of causal inference could be resolved.

To parse what elements of the data-generating process are obfuscating the ATAC, it is useful to decompose the bias of using the "naive" difference-in-means among those who pass the screener. Let

$$\text{DiMS} = E[Y_i \mid S_i = 1, D = 1] - E[Y_i \mid S_i = 1, D = 0] \tag{4}$$

denote the difference-in-means among those who pass the screener (DiMS). For most researchers, this parameter would be the (identifiable) quantity that they would—understably—estimate as a proxy for the ATAC. The following result clarifies the bias that results from equating the DiMS with the ATAC.

Lemma 1.

$$\text{DiMS} - \text{ATAC} = B_1 - B_2 + B_3 - B_4, \tag{5}$$

where the bias terms are defined as

$$\begin{aligned}
B_1 &= \frac{P[A_i(0) = 0, A_i(1) = 1]}{P[A_i(0) = 0, A_i(1) = 1] + P[A_i(0) = 1, A_i(1) = 1]} && (6)\\
&\quad \times (E[Y_i(1) \mid A_i(0) = 0, A_i(1) = 1] - E[Y_i(1) \mid A_i(0) = 1, A_i(1) = 1]) && (7)\\
B_2 &= \frac{P[A_i(0) = 1, A_i(1) = 0]}{P[A_i(0) = 1, A_i(1) = 0] + P[A_i(0) = 1, A_i(1) = 1]} && (8)\\
&\quad \times (E[Y_i(0) \mid A_i(0) = 1, A_i(1) = 0] - E[Y_i(0) \mid A_i(0) = 1, A_i(1) = 1]) && (9)\\
B_3 &= E[Y_i(1) \mid S_i(1) = 1] - E[Y_i(1) \mid A_i(1) = 1] && (10)\\
B_4 &= E[Y_i(0) \mid S_i(0) = 1] - E[Y_i(0) \mid A_i(0) = 1]. && (11)
\end{aligned}$$

See proof on page 22.

Lemma 1 shows the bias can decomposed into two components. The first component, principal strata misclassification arises because respondents with treated compliance $A_i(1) = 1$ may belong to either the always-compliant or activated-compliant strata (B_1). Similarly, respondents with untreated compliance $A_i(0) = 1$ may belong to either the always-compliant strata or suppressed-compliant strata. When these strata diverge in the relevant potential outcomes, the bias terms B_1 and B_2 increase in magnitude. Unfortunately, the researcher cannot adjust for these differences or calculate the relevant probabilities because the principal strata memberships are unknown.

The second component, screener error (B_3 and B_4) arises because $S_i = 1$ does not generally mean that $A_i = 1$. Therefore, some noncompliant respondents can contaminate the subset of respondents who pass the screener relative to the always-compliant stratum. This is the bias that results from measurement error in the screener.

Given the difficulties in correcting for the bias, the best that the researcher can do is to try and partially identify (i.e., bound) the ATAC by combining the observed survey experimental data with the researcher's assumptions about the data-generating process. The next section lays out some assumptions on the data-generating process that will make the partial identification bounds tighter (i.e., more precise).

2.3 Further identification assumptions

To make the bounds on the ATAC as precise as possible, I consider a range of assumptions that either constrain the relationship between compliance A_i and the screener S_i or constrain the treatment's causal effects on compliance.

Assumption A3 (No False Negatives, Known False Positive Rate)**.** Compliant respondents always pass the screener. Noncompliant respondents pass the screener with a known false positive rate $\alpha_d \in [0, 1)$ that may depend on the treatment condition d. Formally, for $d \in \{0, 1\}$,

$$P[S_i(d) = 1 \mid A_i(d) = 1] = 1 \tag{12}$$

$$P[S_i(d) = 1 \mid A_i(d) = 0] = \alpha_d. \tag{13}$$

There are two ideas embedded in this assumption. First, there are no false negatives: if a respondent is compliant, then they will pass the screener without fail. As discussed above, this is part of the definition of compliance.

Second, the false positive rate α_d is known and constant across respondents. It is usually straightforward to guess a reasonable value of α_d. For example, if the screener is a multiple choice question with one correct answer, and we imagine noncompliant respondents respond to the screener by selecting a response option at random, then α_d is equal to one over the number of response options (e.g., $1/5$ if there are five response options). However, if we are unsure of the value of α_d, then the researcher can perform a sensitivity analyses by considering a range of values for α_d. I provide examples of this approach in the applications section.

Of course, in many cases, the chosen value of α_d could be the same for both conditions. However, if different screener items are used across conditions, then α_0 and α_1 should be calibrated separately. Another possibility is that the treatment makes false positives more or less likely by encouraging noncompliant respondents to pick different response options on the screener across conditions. For example, maybe the manipulation check screener item has a popular response option for idiosyncratic reasons (e.g., acquiescence bias) that results in a pass for one condition but a fail for the other condition. In this case, the researcher should calibrate α_0 and α_1 accordingly.

The next two assumptions are optional and can be dropped from the bounding method if strictly necessary. However, invoking these assumptions typically reduces the width of the resulting partial identification bounds by a significant degree. Therefore, they are often worth entertaining.

Assumption A4 (No Differential Measurement Error)**.** The known false positive error rate $\alpha_d \in [0, 1)$ in condition d does not vary based on the measured outcome Y.

$$P[S_i(d) = 1 \mid A_i(d) = 0, Y_i(d) = y] = \alpha_d. \tag{14}$$

This assumption says that, in the aggregate, the response behaviors of noncompliant respondents are not systematic enough to induce a correlation between the screener and the outcome. This might be violated if, for example, all noncompliant respondents pick the first (or last) response option and the response options for both screener and outcome are not presented in a random order.[5] Another potential threat to this assumption is if both the screener and outcome items are prone to acquiescence bias, which could induce a correlation between the screener and outcome among noncompliant respondents.

[5] Many survey instruments already randomize the order of response options to avoid response order effects.

Assumption A5 (Compliance Monotonicity)**.** One of the following two conditions holds. (1) For all respondents, $A_i(1) \geq A_i(0)$; or (2) for all respondents, $A_i(0) \geq A_i(1)$.

In practice, the researcher would assume condition (1) when treatment increases the screener pass rate and condition (2) when the treatment decreases the screener pass rate.[6] Effectively, this assumption makes it so that any compositional effects of treatment on compliance is known to be just a fraction $E[A_i(1) - A_i(0)]$ of respondents moving from noncompliance to compliance under condition (1), reversed for condition (2). No other compositional changes are possible. This assumption parallels the monotonicity assumption of Lee (2009).

Whether Assumption A5 holds might be context dependent. The monotonicity assumption is not always taken for granted (Aronow, Baron, and Pinson 2019). However, in many survey experiments, the experimental stimuli are probably not strong enough to induce a large enough change in compliance, much less that required to violate the monotonicity assumption to a significant degree. For example, changing a few key words in the treatment stimulus probably does not alter attention that much. More generally, A5 is most plausible when (a) compliance is thought to be largely stable across conditions or (b) it is expected that the treatment (weakly) increases or decreases compliance broadly among all types of respondents—as opposed to increasing compliance among some respondents while simultaneously decreasing compliance among others. Assumption A5 seems to be least plausible with the treatment stimulus makes large changes to the length and topic of any text or prompt that respondents are expected to read.

Finally, for the sake of completeness, there is another optional assumption which assumes compliance is either measured before treatment assignment or otherwise perfectly stable across conditions. This assumption is not necessary for the bounding method, but it is useful for comparing pre- versus post-treatment attention checks.

Assumption A6 (No Compliance or Screener Effects)**.** Both compliance and the screener are determined and measured before treatment assignment or otherwise do not vary with treatment. Formally, for all respondents,

$$A_i(1) = A_i(0), \quad S_i(1) = S_i(0). \tag{15}$$

This assumption clearly holds when the screener is measured pre-treatment. It is perhaps also plausible when the treatment stimulus is nearly unchanged across conditions (e.g., changing only one or two words).

3 Sharp bounds via linear programming

To sharply bound the ATAC, I propose a numerical method that calculates the minimum and maximum data- and assumption-consistent values of the ATAC using convex optimization techniques. In particular, this section shows that when sampling variability is ignored, the bounding problem is equivalent to a pair of linear programs. Later sections extend the

[6] More precisely, the researcher should account for the potentially different false positive rates. If p_d is the screener pass rate in condition d, then $E[A_i(d)] = (p_d - \alpha_d)/(1 - \alpha_d)$. The researcher should assume condition (1) if $(p_1 - \alpha_1)/(1 - \alpha_1) > (p_0 - \alpha_0)/(1 - \alpha_0)$ and condition (2) otherwise.

optimization method to account for sampling error to produce confidence intervals. Setting up the problem as an optimization problem requires new notation, but the payoff is that sharp bounds can be computed quickly with convergence guarantees.

The optimization problem I propose is based on searching over possible values of the joint distribution of all potential outcomes. In particular, define the 6-tensor π^* as follows.

$$\begin{aligned} &\pi^*(a_0, a_1, s_0, s_1, j, k) && (16) \\ \overset{\text{def.}}{=}\ & P[A_i(0) = a_0, A_i(1) = a_1, S_i(0) = s_0, S_i(1) = s_1, Y_i(0) = y_j, Y_i(1) = y_k]. && (17) \end{aligned}$$

Of course, if the value of π^* were known, then the ATAC could be derived from it immediately. Given it is unknown, I will derive bounds on the ATAC by restricting π^* with a set of constraints that corresponding to some subset of assumptions A1-A6.[7]

Immediately, one can impose the constraint that $\pi^* \geq 0$ and the sum of π^* across all dimensions is one. These constraints are necessary for π^* to be a valid joint distribution.

Other resitrctions are imposed on the marginal or marginal joint distributions of π^*. In particular, I use the notation ":" to indicate marginalizing (i.e., summing) over the corresponding index. For example,

$$\begin{aligned} \pi^*(a_0, :, s_0, s_1, :, k) &= \sum_{a_1 \in \{0,1\}} \sum_{j=1}^{K} \pi^*(a_0, a_1, s_0, s_1, j, y_k) && (18) \\ &= P[A_i(0) = a_0, S_i(0) = s_0, S_i(1) = s_1, Y_i(1) = y_k]. && (19) \end{aligned}$$

Furthermore, let the notation "*" indicate a slice. For example, $\pi^*(a_0, *, s_0, s_1, *, k)$ is a 2-by-K matrix and $\pi^*(:, :, :, :, :, *)$ is a K-vector. These notations are used below to operationalize the assumptions A2-A6 as constraints on the tensor π^*.

3.1 Constraints from assumption A2

Consider the constraints implied by assumption A2. If treatment is randomly assigned, then the researcher has a simple random sample of the joint distribution of the untreated screener and outcome: $(S_i(0), Y_i(0))$, namely the control group. Similarly, the treatment group is a simple random sample of the joint distribution of the treated screener and outcome: $(S_i(1), Y_i(1))$. Using the new notation introduced above, $\pi^*(:, :, *, :, *, :)$ corresponds to the joint distribution of $(S_i(0), Y_i(0))$, and $\pi^*(:, :, :, *, :, *)$ corresponds to the joint distribution of $(S_i(1), Y_i(1))$. Note that both of these are 2-by-K matrices.

To collect these marginal joint distributions, define the reduced form parameter μ^*, a function of π^*, as follows

$$\mu^* = \begin{bmatrix} \text{vec}(\pi^*(:, :, *, :, *, :)) \\ \text{vec}(\pi^*(:, :, :, *, :, *)) \end{bmatrix}, \tag{20}$$

where vec(M) denotes the vector formed by stacking the columns of the matrix M. By constraining μ based on the observable survey experimental data, one effectively constrains the possible values of π^*.

[7] Assumption A1 is required to even define π^*.

To simplify the rest of this section, I temporarily assume that the value of μ^* is known and observed by the researcher. In the following section, I explain how to estimate μ^* and incorporate estimation error into bounding the ATAC by generating a confidence interval. With known μ^*, the sharp bounds can be understood as the population partial identification bounds for the ATAC—analogous to a population parameter in point estimation theory.

3.2 Constraints from assumptions A3-A4

Assumption A3 says that there are no false negatives and that the false positive rates in each condition are known. Assumption A4 is a stronger version of A3 that is similarly operationalized.

If there are no false negatives, then all respondents with $A_i(d) = 1$ must satisfy $S_i(d) = 1$. By definition of π^* it follows that

$$P[S_i(0) = 1 \mid A_i(0) = 1] = \frac{P[A_i(0) = 1, S_i(0) = 1]}{P[A_i(0) = 1]} = \frac{\pi^*(1,:,1,:,:,:)}{\pi^*(1,:,:,:,:,:)} = 1 \tag{21}$$

$$P[S_i(1) = 1 \mid A_i(1) = 1] = \frac{P[A_i(1) = 1, S_i(1) = 1]}{P[A_i(1) = 1]} = \frac{\pi^*(:,1,:,1,:,:)}{\pi^*(:,1,:,:,:,:)} = 1. \tag{22}$$

Note that these are linear constraints on π^* since they are equivalent to specifying $\pi^*(1,:,1,:,:,:) = \pi^*(1,:,:,:,:,:)$ and $\pi^*(:,1,:,1,:,:) = \pi^*(:,1,:,:,:,:)$.

The second component of assumption A3 is that the false positive rates are known, with values α_0 and α_1 in the control and treatment conditions, respectively. This implies that

$$P[S_i(0) = 1 \mid A_i(0) = 0] = \frac{P[A_i(0) = 0, S_i(0) = 1]}{P[A_i(0) = 0]} \tag{23}$$

$$= \frac{\pi^*(0,:,1,:,:,:)}{\pi^*(0,:,:,:,:,:)} = \alpha_0 \tag{24}$$

$$P[S_i(1) = 1 \mid A_i(1) = 0] = \frac{P[A_i(1) = 0, S_i(1) = 1]}{P[A_i(1) = 0]} \tag{25}$$

$$= \frac{\pi^*(:,0,:,1,:,:)}{\pi^*(:,0,:,:,:,:)} = \alpha_1. \tag{26}$$

These are equivalent to the linear constraints $\pi^*(0,:,1,:,:,:) = \alpha_0 \pi^*(0,:,:,:,:,:)$ and $\pi^*(:,0,:,:,:,:) = \alpha_1 \pi^*(:,0,:,1,:,:)$.

Assumption A4 is a stronger version of Assumption A3 that further conditions on $Y_i(d)$. This is equivalent to the following K pairs of constraints: for each value of $k \in \{1, \dots, K\}$,

$$P[S_i(0) = 1 \mid A_i(0) = 0, Y_i(0) = y_k] = \frac{P[A_i(0) = 0, S_i(0) = 1, Y_i(0) = k]}{P[A_i(0) = 0, Y_i(0) = y_k]} \tag{27}$$

$$= \frac{\pi^*(0,:,1,:,k,:)}{\pi^*(0,:,:,:,k,:)} = \alpha_0 \tag{28}$$

$$P[S_i(1) = 1 \mid A_i(1) = 0, Y_i(1) = y_k] = \frac{P[A_i(1) = 0, S_i(1) = 1, Y_i(1) = y_k]}{P[A_i(1) = 0, Y_i(1) = y_k]} \tag{29}$$

$$= \frac{\pi^*(:,0,:,1,:,k)}{\pi^*(:,0,:,:,:,k)} = \alpha_1. \tag{30}$$

3.3 Constraints from assumptions A5-A6

Assumption A5 constrains the joint distribution of $(A_i(0), A_i(1))$, which is a marginal joint distribution of π^*. Assumption A6 is a stronger version of A5 that includes both versions of A5.

Assumption A5, compliance monotonicity, is equivalent to the condition that there are no respondents who violate monotonicity. In particular, if version A5(1) is assumed and $A_i(1) \geq A_i(0)$, then

$$P[A_i(0) = 1, A_i(1) = 0] = \pi^*(1, 0, :, :, :, :) = 0. \tag{31}$$

Alternatively, if version A5(2) is assumed and $A_i(0) \leq A_i(1)$, then

$$P[A_i(0) = 0, A_i(1) = 1] = \pi^*(0, 1, :, :, :, :) = 0. \tag{32}$$

Assumption A6, which stipulates $A_i(0) = A_i(1)$, is equivalent to both version A5(1) and A5(2) holding simultaneously since that imposes $A_i(0) = A_i(1)$ with probability one.

3.4 Specifying the objective function

To finish the task of bounding the ATAC, one needs to specify the ATAC as a function of π^* and then minimize/maximize this function with respect to the constraints.

In particular, the ATAC can be expressed as a function of π^* as follows.

$$\begin{aligned}
\text{ATAC} &= E[Y_i(1) - Y_i(0) \mid A_i(1) = 1, A_i(0) = 1] && (33)\\
&= E[Y_i(1) \mid A_i(1) = 1, A_i(0) = 1] - E[Y_i(0) \mid A_i(1) = 1, A_i(0) = 1] && (34)\\
&= \sum_{k=1}^{K} y_k (P[Y_i(1) = k \mid A_i(1) = 1, A_i(0) = 1] - P[Y_i(0) = k \mid A_i(1) = 1, A_i(0) = 1]) && (35)\\
&= \sum_{k=1}^{K} y_k \, \frac{\pi^*(1, 1, :, :, :, k) - \pi^*(1, 1, :, :, k, :)}{\pi^*(1, 1, :, :, :, :)}. && (36)
\end{aligned}$$

To define the objective function for optimization, this expression is rewritten as a function of a candidate value of π, namely

$$\tau(\pi) = \sum_{k=1}^{K} y_k \, \frac{\pi(1, 1, :, :, :, k) - \pi(1, 1, :, :, k, :)}{\pi(1, 1, :, :, :, :)}. \tag{37}$$

Note that ATAC $= \tau(\pi^*)$. Critically, this objective function $\tau(\pi)$ is a linear-fractional function of π and can be converted into a linear objective function for the optimization problems considered in this paper. This will enable convex optimization methods, namely linear and second-order cone programming, even though $\tau(\pi)$ is not convex.

3.5 Bounding the ATAC

Conceptually, producing sharp bounds for the ATAC is straightforward. The goal is to find the minimum and maximum values of the objective function $\tau(\pi)$ where the input π is compatible with both the researcher's assumptions and the observable-data distribution μ^*.

To limit the search to compatible values of π, let $F(\mu)$ denote the feasible set of π values that are consistent with the candidate observable-data distribution μ (mandatory assumptions A1-A2) and any subset of the optional assumptions A3-A6. Consistency between π and μ is assessed using equation (20), replacing the true values μ^* and π^* with their candidate values μ and π. Consistency with each optional assumptions is assessed by checking whether the assumption's constraints on π^* are satisfied by the candidate π; see equations (21)-(32).

It is convenient at this point to introduce the functions L and U which respectively give the lower and upper bound on the ATAC consistent with the candidate value of μ.

$$L(\mu) = \min\{\tau(\pi) : \pi \in F(\mu)\} \tag{38}$$

$$U(\mu) = \max\{\tau(\pi) : \pi \in F(\mu)\}. \tag{39}$$

Note that L and U are well-defined when the feasible set $F(\mu)$ is nonempty since $F(\mu)$ is compact and $\tau(\pi)$ is continuous and bounded for $\pi \in F(\mu)$.[8] It follows that, by definition, the left endpoint $L^* = L(\mu^*)$ and right endpoint $U^* = U(\mu^*)$ form the tightest bound $[L^*, U^*]$ on the ATAC consistent with both μ^* and the researcher's assumptions. That is, $[L^*, U^*]$ are the population partial identification bounds on the ATAC. The bounds are sharp in the sense that they cannot be tighter while still containing all values of the ATAC that are consistent with μ^* and the researcher's assumptions.

Of course, the functions $L(\mu)$ and $U(\mu)$ are only useful in practice if they can be calculated. The following result shows that this can be done using fast computational methods.

Theorem 1. Calculating either $L(\mu)$ or $U(\mu)$ is equivalent to a linear program.

See proof on page 22. Since membership in the feasible set $\pi \in F(\mu)$ is equivalent to just a series of linear constraints on π, the crux of the proof is showing how to transform the objective function $\tau(\pi)$ into a linear objective function using the Charnes and Cooper (1962) transformation.

The payoff of Theorem 1 is that the functions $L(\mu)$ and $U(\mu)$ can be calculated quickly and reliably using standard computational methods. Critically, they can be solved with interior point methods that have convergence properties to the global minimum/maximum similar to Newton's method (Boyd and Vandenberghe 2004). There are multiple software packages that can be used to calculate $L(\mu)$ and $U(\mu)$ nearly instantly. In the empirical applications, I use the `CVXR` package in the R programming language (Fu, Narasimhan, and Boyd 2017). Calculating both $L(\mu)$ and $U(\mu)$ this way takes less than a second. To enable others to use the proposed method, I intend to make software that implements the method publicly available in the form of an R package.

[8] In practice, the optimization algorithm will simply inform the researcher if the feasible set $F(\mu)$ is empty.

4 Confidence intervals

Theorem 1 shows that the functions $L(\mu)$ and $U(\mu)$ can be calculated quickly and reliably. However, these functions are only relevant if the researcher knows the true value of μ^*. In practice, the researcher estimates the observable-data distribution μ^* from the sample $\{(D_i, S_i, Y_i)\}_{i=1}^n$ with error. This section introduces a method for generating asymptotically valid pointwise confidence intervals for the ATAC that account for both the lack of identifiability and sampling error. The proposed confidence intervals can be computed using second-order cone programming, which has similar performance to linear programming in this setting.

4.1 Proposing a confidence interval

My approach is based on searching over plausible μ^* values using a confidence region for μ^* whose volume depends on the desired coverage rate γ (e.g., $\gamma = .95$ for a 95% confidence interval).

The center of this confidence region is the standard method-of-moments estimator $\hat{\mu}$. Recall the i.i.d. assumption introduced at the start of the model exposition. Suppose $n_1 = \sum_{i=1}^n D_i$ units are assigned to treatment, and $n_0 = n - n_1$ units to control. Define $q = n_1/n$ and

$$\hat{\mu}_{sy}^0 = n^{-1} \sum_{i=1}^n \frac{(1 - D_i)}{1 - q} I[S_i = s, Y_i = y] \tag{40}$$

$$\hat{\mu}_{sy}^1 = n^{-1} \sum_{i=1}^n \frac{D_i}{q} I[S_i = s, Y_i = y] \tag{41}$$

$$\hat{\mu} = \begin{bmatrix} \text{vec}(\hat{\mu}^0) \\ \text{vec}(\hat{\mu}^1) \end{bmatrix}, \tag{42}$$

where $I[\dots]$ is the usual indicator function. Reweighting the terms in the sums $\hat{\mu}_{sy}^0$ and $\hat{\mu}_{sy}^1$ means that $\hat{\mu}$ is consistent for μ^* by the law of large numbers. Furthermore, by the central limit theorem, it follows that

$$\sqrt{n}(\hat{\mu} - \mu^*) \rightsquigarrow \mathcal{N}(0, \Sigma), \tag{43}$$

where $\rightsquigarrow$ denotes convergence in distribution as the sample size tends to infinity and $\mathcal{N}$ denotes the (multivariate) normal distribution. It is well known that the estimator $\hat{\Sigma} = \text{diag}(\hat{\mu}) - \hat{\mu}\hat{\mu}^T$ is consistent for Σ as $n \to \infty$.

To specify the confidence region conveniently for the optimization problem, I use the square root of Σ, denoted Ψ. Without loss of generality, one can write $\Sigma = QDQ'$ where D is a diagonal matrix (of eigenvalues) and Q is an orthogonal matrix (of eigenvectors). The matrix square root Ψ is defined as $QD^{\frac{1}{2}}Q'$, which guarantees that $\Psi^2 = \Sigma$. Similarly, define its (consistent) estimator $\hat{\Psi}$ as the square root of the estimator $\hat{\Sigma}$.[9]

[9] Due to numerical instability, one or more diagonal elements of $\hat{D}$ might be within machine precision of zero but negative. To correct for this, I set $\hat{D}_{ii}^{1/2} = \max(0, \hat{D}_{ii})^{1/2}$, which means $\hat{\Psi}$ is the matrix square root of the positive semidefinite matrix which is closest to $\hat{\Sigma}$ in spectral norm.

Finally, I construct the confidence region for μ^* based on a radius $r \geq 0$, to be calibrated based on γ. The confidence region is given by the set

$$\left\{\hat{\mu} + n^{-\frac{1}{2}}\hat{\Psi}z : \|z\| \leq r\right\}. \tag{44}$$

The radius r controls how often, asymptotically, the confidence region covers the entire vector μ^*. Larger values of r lead to a higher coverage rate at the cost of expanding the feasible set of μ (and thus π) values, resulting in wider bounds and less precise inferences. Therefore, tuning the value of r is critical.

A seemingly reasonable choice would be to tune r so that the confidence region covers the entire vector μ^* with the nominal coverage rate γ of the final ATAC confidence interval (e.g., a 95% confidence region for μ^*). In this approach, since Σ is rank $4K - 2$, the radius r^2 should be set to the γ quantile of the χ^2 distribution with $4K - 2$ degrees of freedom (Duarte et al. 2024). For example, if $\gamma = .95$ and $K = 2$, then set $r = 3.55$. However, as I show below, this choice of r is far too conservative since the entire vector μ^* doesn't need to be covered for the ultimate goal of covering the unidmensional ATAC (or any other unidmensional parameter).

Given the $L(\mu)$ and $U(\mu)$ functions and a radius $r \geq 0$, I propose a confidence interval $[\hat{L}_r, \hat{U}_r]$ for the ATAC as follows.

$$\hat{L}_r = \min\left\{L\Big(\hat{\mu} + n^{-\frac{1}{2}}\hat{\Psi}z\Big) : \|z\| \leq r\right\} \tag{45}$$

$$\hat{U}_r = \max\left\{U\Big(\hat{\mu} + n^{-\frac{1}{2}}\hat{\Psi}z\Big) : \|z\| \leq r\right\} \tag{46}$$

This is simply an extension of the functions $L(\mu)$ and $U(\mu)$ where μ is required to lie in the confidence region generated by the radius r.

The following result yields a lower bound on the asymptotic coverage rate of $[\hat{L}_r, \hat{U}_r]$, enabling the researcher to tune the radius r accordingly to their desired coverage rate γ.

Theorem 2. *Suppose the functions $L(\mu)$ and $U(\mu)$ are differentiable at μ^* with gradients G_L, G_U satisfying $\Psi G_L \neq 0$, $\Psi G_U \neq 0$. Let Φ^{-1} denote the quantile function of the standard normal distribution.*

a) *If $L^* < U^*$, then setting $r = \Phi^{-1}(\gamma)$ for $\gamma \geq .5$ results in*

$$\lim_{n\to\infty} P\Big[\hat{L}_r \leq \text{ATAC} \leq \hat{U}_r\Big] \geq \gamma.$$

b) *If $L^* = U^*$, then setting $r = \Phi^{-1}((1 + \gamma)/2)$ results in*

$$\lim_{n\to\infty} P\Big[\hat{L}_r \leq \text{ATAC} \leq \hat{U}_r\Big] \geq \gamma.$$

See proof on page 22. Requiring $L(\mu)$ and $U(\mu)$ to be differentiable leads to a unique delta-method-style result for the estimators $\hat{L}_r, \hat{U}_r$.[10] The logic of the proof is surprisingly

[10] Requiring $\Psi G_L \neq 0$ and $\Psi G_U \neq 0$ allows rules out a presumably obscure edge case—which may not ever occur—where $(\hat{L}_0, \hat{U}_0)$ converge to (L^*, U^*) even *faster* than the standard $n^{\frac{1}{2}}$ rate. I ignore this case since it would require higher-order asymptotic theory to characterize the coverage rate.

general and does not depend on the specifics of L and U except for the differentiability assumption, suggesting that the result applies to many other optimization-based methods that meet the condition.

Given this result, one naturally wonders whether the theorem's assumptions are likely to hold in practice. We are not flying completely dark: since $L(\mu)$ and $U(\mu)$ are the solutions to linear programs, we know that their solutions are obtained at the vertices of the same convex polytope (i.e., the feasible set). The shape and position of this polytope in space, and thus its vertices, depend on the argument μ. Therefore, the differentiability assumption is really a smoothness condition on the vertices of this convex polytope as a function of μ near $\mu = \mu^*$. However, it is presently unclear how one could verify the differentiability condition in practice.

Additionally, while the cases in Theorem 2 are specified in terms of the unknown values L^* and U^*, the researcher can in practice check these conditions by comparing $\hat{L}_0$ and $\hat{U}_0$, obtained by setting $r = 0$. This is because $L^* = U^*$ is equivalent to the ATAC being point identified, and that is largely dependent on the assumptions determining the feasible set $F(\mu)$ rather than the the value of the observable-data distribution μ^*. For example, the ATAC is point identified ($L^* = U^*$) when $\alpha_0 = \alpha_1 = 0$ and A1-A6 are assumed, no matter the value of μ^*. Otherwise, the ATAC is not point identifiable and thus $L^* < U^*$.[11]

4.2 Calculating the confidence interval

To make this result actionable, I argue that calculating the lower and upper bounds of the confidence interval $[\hat{L}_r, \hat{U}_r]$ is, in practice, just as straightforward as calculating the partial identification region $[L^*, U^*]$ when μ^* is known. This arugment is predicated on the following result.

Theorem 3. Calculating either $\hat{L}_r$ or $\hat{U}_r$ is equivalent to a second-order cone program.

See proof on page 25. The proof relies on the same rescaling technique that I used for Theorem 1 to transform the objective into a linear function. I then show how to transform the radius constraint $\|z\| \leq r$ into a second-order cone constraint based on the rescaled probabilities.

From the analyst's perspective, Theorem 3 implies that calculating the confidence interval $[\hat{L}_r, \hat{U}_r]$ is no harder than calculating $[L^*, U^*]$. Like linear programs, interior point methods can solve second-order cone programs quickly with convergence guarantees similar to Newton's method (Boyd and Vandenberghe 2004). For example, using the `CVXR` package in `R`, I can compute a confidence interval on a standard survey dataset in about a second. I intend to make this code publicly available in the form of an R package.

5 Applied examples

I apply the proposed method to two example studies. The first example is a reanalysis of Dancygier and Wiedemann (2024), who examine support for expropriating corporate

[11] In practice, the resarcher can always check whether $\hat{L}_0 = \hat{U}_0$ to test for point identifiability.

landlords using a survey experiment of German residents (Study 1). The second example is in progress.

5.1 Study 1

The Dancygier and Wiedemann (2024) survey experiment randomly assigned a 5087 German resident respondents to one of three treatment conditions and one control condition. For simplicity, I limit my reanalysis to just the control and one treatment condition (T2: Financialization), hereafter "the treatment condition". Respondents in the control condition were asked to read a vignette about the health benefits of eating broccoli. Meanwhile, respondents in the treatment condition instead read a vignette about how large real estate companies have are increasingly buying apartments in German cities as investments motivated entirely by profit. See Dancygier and Wiedemann (2024) for the full text of each vignette. After reading either vignette, respondents were asked to indicate their support for "the idea of expropriating large real estate companies in exchange for compensation and transferring them into public ownership," with the response coded as a binary outcome variable (1 = supports the policy, 0 = otherwise).

In line with the authors' expectations, the difference-in-means suggests that the treatment condition increased support for expropriating corporate landlords by about 9 percentage points (from 60 to 69 percentage points). However, it remains somewhat unclear how much of this difference can be attributed to the authors' manipulation of the content of the vignettes, for the reasons articulated above.

To gauage whether respondents were actually exposed to the information contained in the treatment, the authors asked respondents to complete a manipulation check. In particular, respondents were required to indicate what information they had been exposed to in the vignette.[12] Respondents successfully passed the manipulation check if they selected the correct response option out of the four that were available.[13] Additionally, before administering the treatment, the authors measured a pre-treatment attention check for all responsdents. I limit my reanalysis to respondents who passed this pre-treatment attention check ($n_0 = 860$, $n_1 = 920$).

Taking a difference in means approach implies that the treatment increased aggregate support for expropriation but also decreased aggregate compliance. In terms of the outcome, the average support for expropriation in the treatment group was 69% versus 60% in the control group ($p < .05$). Meanwhile, about 78% respondents passed the manipulation check in the control condition, and 74% passed it in the treatment condition ($p < .05$). This slight decrease in compliance is not surprising given the treatment vignette is slightly more complex and the manipulation check is probably easier for respondents in the control condition

[12] The authors kindly shared the details of the manipulation check with me in a separate correspondence. The manipulation check prompt, in German: "Wenn Sie jetzt an die Aussage zurückdenken, die Sie auf der vorherigen Seite gelesen haben, welche haben Sie gesehen? Bitte markieren Sie alle Aussagen, an die Sie sich erinnern."

[13] Respondents in the control condition passed if they selected "Brokkoli ist ein gesundes Lebensmittel." Respondents in the treatment condition passed if they selected "Konzerne kaufen Wohnungen in deutschen Städten als Spekulationsinstrumente." The other two response options were "Umweltverschmutzung ist ein Problem." and "Mieten in deutschen Städten steigen."

given the manipulation check only has one Broccoli-related response option. In contrast, the treatment vignette has slightly more complex information and the manipulation check has two housing-related response options. I note that there is still a significant number of respondents who failed the manipulation check in the survey experiment (at least 30% in each condition), implying that noncompliance remains an experimental concern even after applying a pre-treatment screener.

What then is the effect of the treatment among respondents who are compliant in both conditions, the ATAC? As a baseline estimate, the difference-in-means among those who pass the screener (DiMS) is estimated at 13 (SE = 0.03, $p < .05$) percentage points. However, this estimator could be biased by both principal strata misclassification and screener measurement error. To address these concerns, I apply the proposed method to construct sharp bounds on the ATAC. I consider different types of bounds, each different in which subset of assumptions A1-A6 are assumed.

Table 1 reports the bounding results. For six combination of assumptions, including the Lee (2009) bounds which are a special case of the proposed method with for assumptions A1-A5 and $\alpha = 0$.[14] The estimated bounds (EB) refer to $[\hat{L}_0, \hat{U}_0]$, or a confidence interval with radius $r = 0$. The estimated bounds are effectively point estimates of the population partial identification bounds $[L^*, U^*]$. The confidence intervals (CI) are constructed using the radius prescribed by Theorem 2 with $r = 1.64$. Except for the Lee (2009) bounds, all bounds are estimated assuming $\alpha_0 = \alpha_1 = \alpha = .25$ given the manipulation check has one correct response option and three incorrect response options.

Table 1: ATAC bounds under various assumptions (Study 1)

Type	**Assumptions**	α	**EB Lower**	**EB Upper**	**CI Lower**	**CI Upper**
I	A1-A3	.25	-.57	.87	-.66	.99
II	A1-A4	.25	-.48	.8	-.55	.91
III	A1-A3, A5	.25	-.024	.3	-.093	.36
IV	A1-A5	.25	.097	.18	.032	.24
Lee (2009)	A1-A5	0	.094	.15	.043	.2
V	A1-A6	.25	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$

Note: EB = estimated bounds. CI = confidence interval. The $\emptyset$ indicates that the bounds are infeasible given the assumptions.

Looking at Table 1, we see that the estimated bounds and confidence intervals become progresively tighter as more assumptions are imposed. This is not surprising, but it is interesting to note which assumptions appear to be most consequential. In particular, compliance monotonicity (A5) appears to be the most impactful assumption, reducing the bound width by at least half (Type II to Type IV). Additionally, differential measurement error (A4) is consequential in combination with compliance monotonicity (A5); together they exclude 0 from both the EB and CI bounds. Note that both the estimated bounds for Type IV and

[14] I verified that the proposed method returns the same estimated bounds and 95% confidence interval as the `leebounds` module in Stata (Tauchmann 2014).

Lee (2009) just barely rule out the conventional 9 percentage-point ATE estimate (which ignores the screener).

Additionally, note that the Type V bounds are infeasible, and thus cannot be computed, given the addition of the fixed compliance assumption (A6). This is expected given the modest but significant difference in screener pass rates between the two conditions. The data indicate that some respondents, although perhaps only a few, violate the fixed compliance assumption.

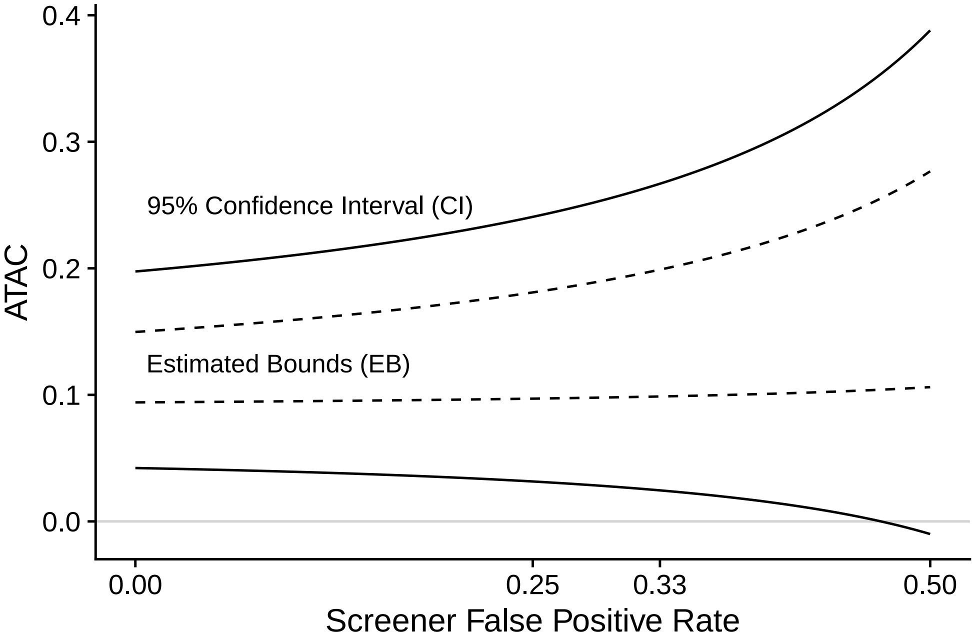
Which ATAC bounds are most credible? Naturally, the researcher wants the tightest bounds that don't require any untenable assumptions. Which assumptions are tenable depends on the case. Since the Type V bounds are infeasible, I can confidently rule out assumption A6. Beyond A6, assumption A3-A4 might be controversial if only because I have asserted a false positive rate of .25, but I perform a sensitivity analysis below to examine the robustness to the specific false positive rate. Additionally, assumption A4 could be controversial if the screener and outcome response options were not presented in any kind of randomized order—since a fixed response order could induce a conditional correlation between the outcome Y and screener S among noncompliant respondents. But, since it is common practice, one assumes that the authors preempted this concern by randomizing the order of response options.

Given the treatment appears to decrease compliance, assumption A5 in this study stipulates that there is no respondent whose compliance was increased by the treatment. This is perhaps not exactly true for every respondent, but perhaps it holds approximately, with very few respondents (say, <1%) in violation. A significant violation of A5 would require enough respondents to become noticeably more engaged when tasked to read about the financialization of the urban housing supply. I am therefore skeptical that A5 is violated by a meaningful number of respondents.

Consider assumptions A3-A4 again. To gauge how sensitive the estimated bounds and confidence intervals are to the assumed false positive rate α, I vary the assumed value of α from 0 to 1/3 in Figure 1 for the Type IV bounds. The estimated bounds are represented by the dashed lines, and the confidence intervals are represented by the solid lines. On the horizontal axis, I have indicated the false positive rate α values .25 (=1/4), and .33 (=1/3), and .5 (1=/2). The value $\alpha = .33$ is a conceptually interesting landmark because it is the value of α consistent with noncompliant respondents being able to perfectly rule out one incorrect repsonse option from the manipulation check, perhaps because it is implausible, and then they selected randomly from the remaining options. Similarly, the $\alpha = .5$ landmark is noted because it assumes noncompliant respondents can rule out two incorrect response options.[15] Note that, in the special case where $\alpha = 0$, the bounds are exactly equal to the Lee (2009) bounds.

[15] By definition, being able to rule out the remaining incorrect response option would mean that a respondent is compliant.

Figure 1: Sensitivity analysis for type IV ATAC bounds (Study 1)



Looking at Figure 1, one observes that the bounds become, unsuprisingly, wider as the assumed false positive rate α increases. That is, the bounds become wider with more measurement error. However, this width increase is asymmetrical: the upper estimated bound increases at a faster rate than the lower estimated bound, implying that a higher false positive rate is consistent with higher values of the ATAC than are the Lee (2009) bounds.[16] Somewhere before $\alpha = .5$, the confidence interval starts to include 0, which would be consistent with a null effect among the always-compliant. However, such a value of α assumes that most noncompliant respondents can rule out two incorrect response options, which seems implausible—perhaps some can, but not most. At best, respondents in the treatment condition might be able to quickly rule out the incorrect response option that mentions broccoli, but that still only implies that $\alpha_1 = .33$.[17] The qualitative conclusion that the financialization vignette increased support for expropriation is not sensitive to varying α in the range $[.25, .33]$.

6 Conclusion

In this paper, I have introduced a novel method for bounding the average treatment effect among always-compliant respondents (ATAC) in survey experiments with manipulation checks. Critically, these manipulation checks are prone to survey measurement error, just like any other construct measured with surveys. By leveraging convex optimization techniques, I have shown that the bounding problem can be formulated as a pair of linear programs,

[16] In this case, the lower estimated bound also increases, but at a much slower rate.

[17] If $\alpha_1 = .33$ but $\alpha_0 = .25$, then the estimated bounds are $[0.062, 0.22]$ and the confidence interval is $[-0.012, 0.29]$.

allowing for fast and reliable computation of sharp bounds. Additionally, I have extended the method to account for sampling variability, providing asymptotically valid confidence intervals for the ATAC using second-order cone programming. The proposed method is flexible, allowing researchers to include different sets of assumptions about the data-generating process.

To solve the overcoverage problem of existing optimization-based methos (Duarte et al. 2024), I derived the asymptotic coverage rate of the proposed confidence intervals. The proof is sufficiently general to apply to any optimization-based method that satisfies the assumptions of Theorem 2. This result allows researchers to choose a radius for optimization-based confidence intervals that achieves a desired asymptotic coverage rate.

Finally, I have demonstrated the method's utility with a reanalysis of Dancygier and Wiedemann (2024). I showed which assumptions are most consequential for bounding the ATAC and how (in)sensitive the qualitative conclusions are in that case to the assumed false positive rate of the manipulation check. Of course, in general, other survey experimental results may not be as robust.

A promising avenue for future work is in generating different types of assumptions which can be seen as substitutes for compliance monotonicity. In particular, one promising approach would be to restrict the degree to which principal strata are correlated with the potential outcomes, drawing on the insights of Lemma 1. This would allow for the researcher to limit the data-generating process, and thus achieve tighter bounds, without requiring the brittle assumption of compliance monotonicity. If this new assumption can be operationalized as a convex constraint, then the proposed optimization-based approach can be immediately extended to account for the new assumption and compute more credible bounds.

References

Aronow, Peter M, Jonathon Baron, and Lauren Pinson. 2019. A note on dropping experimental subjects who fail a manipulation check. *Political Analysis* 27 (4): 572–589.

Blair, Graeme, Winston Chou, and Kosuke Imai. 2019. List experiments with measurement error. *Political Analysis* 27 (4): 455–480.

Boyd, Stephen P, and Lieven Vandenberghe. 2004. *Convex optimization.* Cambridge university press.

Briggs, Ryan, John Mellon, Vincent Arel-Bundock, and Tim Larson. 2025. *We used llms to track methodological and substantive publication patterns in political science and they seem to do a pretty good job.* https://osf.io/v7fe8. Unpublished manuscript.

Charnes, Abraham, and William W Cooper. 1962. Programming with linear fractional functionals. *Naval Research logistics quarterly* 9 (3-4): 181–186.

Clayton, Katherine, Yusaku Horiuchi, Aaron R Kaufman, Gary King, Mayya Komisarchik, Danny Ebanks, Jonathan N Katz, Gary King, Georgina Evans, Gary King, et al. 2023. Correcting measurement error bias in conjoint survey experiments. *American Journal of Political Science* 12 (B2): 1–11.

Dancygier, Rafaela, and Andreas Wiedemann. 2024. The financialization of housing and its political consequences. *American Journal of Political Science.*

Duarte, Guilherme, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. 2024. An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association* 119 (547): 1778–1793.

Fu, Anqi, Balasubramanian Narasimhan, and Stephen Boyd. 2017. Cvxr: an r package for disciplined convex optimization. *arXiv preprint arXiv:1711.07582.*

Lee, David S. 2009. Training, wages, and sample selection: estimating sharp bounds on treatment effects. *Review of Economic Studies* 76 (3): 1071–1102.

Mutz, Diana C. 2021. Improving experimental treatments in political science. *Advances in Experimental Political Science* 219.

Tauchmann, Harald. 2014. Lee (2009) treatment-effect bounds for nonrandom sample selection. *The Stata Journal* 14 (4): 884–894.

Westwood, Sean J, Justin Grimmer, Matthew Tyler, and Clayton Nall. 2022. Current research overstates american support for political violence. *Proceedings of the National Academy of Sciences* 119 (12): e2116870119.

A Proofs

Lemma 1.

$$\text{DiMS} - \text{ATAC} = B_1 - B_2 + B_3 - B_4, \tag{5}$$

where the bias terms are defined as

$$\begin{aligned}
B_1 &= \frac{P[A_i(0)=0, A_i(1)=1]}{P[A_i(0)=0, A_i(1)=1] + P[A_i(0)=1, A_i(1)=1]} && (6)\\
&\quad \times \left(E[Y_i(1) \mid A_i(0)=0, A_i(1)=1] - E[Y_i(1) \mid A_i(0)=1, A_i(1)=1]\right) && (7)\\
B_2 &= \frac{P[A_i(0)=1, A_i(1)=0]}{P[A_i(0)=1, A_i(1)=0] + P[A_i(0)=1, A_i(1)=1]} && (8)\\
&\quad \times \left(E[Y_i(0) \mid A_i(0)=1, A_i(1)=0] - E[Y_i(0) \mid A_i(0)=1, A_i(1)=1]\right) && (9)\\
B_3 &= E[Y_i(1) \mid S_i(1)=1] - E[Y_i(1) \mid A_i(1)=1] && (10)\\
B_4 &= E[Y_i(0) \mid S_i(0)=1] - E[Y_i(0) \mid A_i(0)=1]. && (11)
\end{aligned}$$

Proof of Lemma 1. XYZ □

Theorem 1. Calculating either $L(\mu)$ or $U(\mu)$ is equivalent to a linear program.

Proof of Theorem 1. XYZ □

Theorem 2. Suppose the functions $L(\mu)$ and $U(\mu)$ are differentiable at μ^* with gradients G_L, G_U satisfying $\Psi G_L \neq 0$, $\Psi G_U \neq 0$. Let Φ^{-1} denote the quantile function of the standard normal distribution.

a) If $L^* < U^*$, then setting $r = \Phi^{-1}(\gamma)$ for $\gamma \geq .5$ results in

$$\lim_{n\to\infty} P\Big[\hat{L}_r \leq \text{ATAC} \leq \hat{U}_r\Big] \geq \gamma.$$

b) If $L^* = U^*$, then setting $r = \Phi^{-1}((1+\gamma)/2)$ results in

$$\lim_{n\to\infty} P\Big[\hat{L}_r \leq \text{ATAC} \leq \hat{U}_r\Big] \geq \gamma.$$

Proof of Theorem 2. Let $T_n = \hat{\mu} - \mu^* = \mathcal{O}_p(n^{-\frac{1}{2}})$ and suppose $W \sim \mathcal{N}(0, I)$ where I is the appropriate identity matrix. From the central limit theorem, it follows that $T_n = n^{-\frac{1}{2}}\Psi W + o_p(n^{-\frac{1}{2}})$. Additionally, let $V_n = n^{-\frac{1}{2}}\hat{\Psi} = \mathcal{O}_p(n^{-\frac{1}{2}})$.

Define the remainder function $R(h)$ as

$$R(h) = \frac{L(\mu^* + h) - L^* - G_L' h}{\|h\|} \tag{47}$$

when $h \neq 0$ and $R(0) = 0$. Recall that $L^* = L(\mu^*)$. Because L is differentiable at μ^*, we know that $\lim_{h\to 0} |R(h)| = 0$. Rearranging, I can write for any h

$$L(\mu^* + h) = L^* + G_L' h + \|h\| R(h). \tag{48}$$

Plugging in $h = T_n + V_n z$ yields

$$L(\hat{\mu} + V_n z) = L^* + G_L' T_n + G_L' V_n z + R_n(z), \tag{49}$$

where $R_n(z) = \|T_n + V_n z\| R(T_n + V_n z)$ is a stochastic process indexed by z.

To obtain $\hat{L}_r$, I minimize both sides of the equation over $z \in \mathcal{B}_r = \{z : \|z\| \leq r\}$, yielding

$$\hat{L}_r = \inf_{z \in \mathcal{B}_r} L(\hat{\mu} + V_n z) \tag{50}$$

$$= L^* + G_L' T_n + \inf_{z \in \mathcal{B}_r} [G_L' V_n z + R_n(z)]. \tag{51}$$

Note that the first two terms on the right-hand side do not depend on z whatsoever.

Because $\inf_x f(x) + \inf_x g(x) \leq \inf_x (f(x) + g(x)) \leq \inf_x f(x) + \sup_x g(x)$, it follows that

$$-r\|V_n' G_L\| + \inf_{z \in \mathcal{B}_r} R_n(z) \leq \inf_{z \in \mathcal{B}_r} [G_L' V_n z + R_n(z)] \leq -r\|V_n' G_L\| + \sup_{z \in \mathcal{B}_r} R_n(z), \tag{52}$$

where I use the fact that $\inf_{z \in \mathcal{B}_r} v'z = -r\|v\|$. Consequently,

$$\inf_{z \in \mathcal{B}_r} R_n(z) \leq \hat{L}_r - L^* - G_L' T_n + r\|V_n' G_L\| \leq \sup_{z \in \mathcal{B}_r} R_n(z). \tag{53}$$

I combine this result with the identities $\sup_x f(x) \leq \sup_x |f(x)|$ and $\inf_x f(x) \geq -\sup_x |f(x)|$ to obtain the finite-sample bound

$$\left|\hat{L}_r - L^* - G_L' T_n + r\|V_n' G_L\|\right| \leq \sup_{z \in \mathcal{B}_r} |R_n(z)| = \sup_{z \in \mathcal{B}_r} \|T_n + V_n z\| |R(T_n + V_n z)|. \tag{54}$$

For all $z \in \mathcal{B}_r = \{z : \|z\| \leq r\}$, observe that

$$\|T_n + V_n z\| \leq \|T_n\| + \|V_n\| \|z\| \leq \|T_n\| + \|V_n\| r = \mathcal{O}_p(n^{-\frac{1}{2}}) = o_p(1). \tag{55}$$

By implication, $\sup_{z \in \mathcal{B}_r} \|T_n + V_n z\| = \mathcal{O}_p(n^{-\frac{1}{2}}) = o_p(1)$. Thus, I obtain the intermediate bound

$$\left|\hat{L}_r - L^* - G_L' T_n + r\|V_n' G_L\|\right| \leq (\|T_n\| + \|V_n\| r) \sup_{z \in \mathcal{B}_r} |R(T_n + V_n z)|. \tag{56}$$

Next, choose any ϵ and η. Because $\lim_{h \to 0} |R(h)| = 0$, one can choose δ such that $\|h\| < \delta$ implies that $|R(h)| < \eta$. Because $\sup_{z \in \mathcal{B}_r} \|T_n + V_n z\| = o_p(1)$, one can choose N such that $P(\sup_{z \in \mathcal{B}_r} \|T_n + V_n z\| > \delta) < \epsilon$ for all $n \geq N$. By implication, $P(\sup_{z \in \mathcal{B}_r} |R(T_n + V_n z)| > \eta) < \epsilon$ for all $n \geq N$. Thus, $\sup_{z \in \mathcal{B}_r} |R(T_n + V_n z)| = o_p(1)$.

Taken together, I conclude that

$$(\|T_n\| + \|V_n\| r) \sup_{z \in \mathcal{B}_r} |R(T_n + V_n z)| = \mathcal{O}_p(n^{-\frac{1}{2}}) o_p(1) = o_p(n^{-\frac{1}{2}}). \tag{57}$$

Therefore, by definition of o_p,

$$\hat{L}_r = L^* + G_L' T_n - r\|V_n' G_L\| + o_p(n^{-\frac{1}{2}}). \tag{58}$$

Because $T_n = n^{-\frac{1}{2}}\Psi W + o_p(n^{-\frac{1}{2}})$ and $\hat{\Psi} = \Psi + o_p(1)$, this further simplifies to

$$\hat{L}_r = L^* + n^{-\frac{1}{2}} G_L' \Psi W - n^{-\frac{1}{2}} r \|\Psi G_L\| + o_p(n^{-\frac{1}{2}}), \tag{59}$$

$$\implies \sqrt{n}(\hat{L}_r - L^*) = G_L' \Psi W - r\|\Psi G_L\| + o_p(1). \tag{60}$$

This establishes a delta-method-style result for $\hat{L}_r$, although note that the mean of the asymptotic normal distribution has been shifted by $-r$ standard deviations. For brevity, I omit the symmetric proof for $\hat{U}_r$ that establishes

$$\sqrt{n}(\hat{U}_r - U^*) = G_U' \Psi W + r\|\Psi G_U\| + o_p(1). \tag{61}$$

To simplify the presentation below, I introduce additional notation. Define, $\sigma_L = \|\Psi G_L\| > 0$ and $\sigma_U = \|\Psi G_U\| > 0$ (both positive by assumption). Then, define $g_L = \Psi G_L / \sigma_L$ and $g_U = \Psi G_U / \sigma_U$. Note that g_L and g_U are both unit vectors. Using this new notation,

$$\sqrt{n}(\hat{L}_r - L^*)/\sigma_L = g_L' W - r + o_p(1). \tag{62}$$

$$\sqrt{n}(\hat{U}_r - U^*)/\sigma_U = g_U' W + r + o_p(1). \tag{63}$$

Observe that both $g_L' W$ and $g_U' W$ are standard normal since $g_L' I g_L = 1$ and $g_U' I g_U = 1$.

To calculate the coverage rates, take $\tau \in [L^*, U^*]$.

$$\hat{L}_r \le \tau \iff \hat{L}_r - L^* \le (\tau - L^*) \tag{64}$$

$$\iff \frac{\sqrt{n}(\hat{L}_r - L^*)}{\sigma_L} \le \frac{\sqrt{n}(\tau - L^*)}{\sigma_L} \tag{65}$$

$$\iff g_L' W \le \frac{\sqrt{n}(\tau - L^*)}{\sigma_L} + r + o_p(1). \tag{66}$$

Similarly,

$$\hat{U}_r \ge \tau \iff g_U' W \ge -\frac{\sqrt{n}(U^* - \tau)}{\sigma_U} - r + o_p(1). \tag{67}$$

Applying the inclusion-exclusion principle reveals

$$P(\hat{L}_r \le \tau \le \hat{U}_r) = P(\hat{L}_r \le \tau) + P(\tau \le \hat{U}_r) - P(\hat{L}_r \le \tau \cup \tau \le \hat{U}_r) \tag{68}$$

$$\ge P(\hat{L}_r \le \tau) + P(\tau \le \hat{U}_r) - 1 \tag{69}$$

$$= P\left(g_L' W \le \frac{\sqrt{n}(\tau - L^*)}{\sigma_L} + r + o_p(1)\right) \tag{70}$$

$$+ P\left(g_U' W \ge -\frac{\sqrt{n}(U^* - \tau)}{\sigma_U} - r + o_p(1)\right) - 1 \tag{71}$$

$$= \Phi\left(\frac{\sqrt{n}(\tau - L^*)}{\sigma_L} + r\right) + \Phi\left(\frac{\sqrt{n}(U^* - \tau)}{\sigma_U} + r\right) - 1 + o, \tag{72}$$

where Φ is the cumulative distribution function of the standard normal.

If $L^* = U^*$, then $\tau = L^* = U^*$ and thus

$$P(\hat{L}_r \le \tau \le \hat{U}_r) \ge 2\Phi(r) - 1 + o(1) \to 2\Phi(r) - 1. \tag{73}$$

If $L^* < U^*$, there are three possibilities.

- First, if $\tau = L^* < U^*$, then

$$P(\hat{L}_r \leq \tau \leq \hat{U}_r) \geq \Phi(r) + \Phi\Big(O(\sqrt{n}) + r\Big) - 1 + o(1) \to \Phi(r). \tag{74}$$

- Second, if $\tau = U^* > L^*$, then

$$P(\hat{L}_r \leq \tau \leq \hat{U}_r) \geq \Phi\Big(O(\sqrt{n}) + r\Big) + \Phi(r) - 1 + o(1) \to \Phi(r). \tag{75}$$

- Third and finally, if $L^* < \tau < U^*$, then

$$P(\hat{L}_r \leq \tau \leq \hat{U}_r) \geq \Phi\Big(O(\sqrt{n}) + r\Big) + \Phi\Big(O(\sqrt{n}) + r\Big) - 1 + o(1) \to 1. \tag{76}$$

The proof is complete after recognizing that

$$2\Phi\Big(\Phi^{-1}((1+\gamma)/2)\Big) - 1 = \gamma, \quad \Phi\Big(\Phi^{-1}(\gamma)\Big) = \gamma. \tag{77}$$

□

Theorem 3. Calculating either $\hat{L}_r$ or $\hat{U}_r$ is equivalent to a second-order cone program.

Proof of Theorem 3. XYZ □