A Methodological Approach for Causal Inference under Uncontrolled (and Possibly Latent) Pre-exposure to Treatment

Diogo Ferrari[*]

July 8, 2024

[*] Assistant Professor, Dept. of Political Science, 2234 Watkins Hall, University of California, Riverside (diogo.ferrari@ucr.edu)

A Methodological Approach for Causal Inference under Uncontrolled (and Possibly Latent) Pre-exposure to Treatment

Abstract

Experiments have become one of the main tools for causal inference in political science, but scholars have raised concerns regarding the uncontrolled real-world pre-exposure of subjects to the information manipulated during the experiment. This pre-exposure can mislead conclusions about the treatment effects, and the problem is exacerbated because direct measurement of pre-exposure is not feasible in many studies due to the risk of confounding the experiment. This paper presents a method to estimate causal effects when pre-exposure is uncontrolled and directly unobservable. It formalizes the problem using the potential outcomes framework; decomposes the average treatment effect (ATE) into pre-exposure and exposure components; derives the bias that emerges when pre-exposure is ignored; establishes sufficient identification conditions; and introduces a bias-corrected estimator for the relevant causal parameters. The method is applied to analyze the impact of party cues on voters' policy attitudes.

Word count: 7,180

Introduction

Recently, researchers have raised concerns about threats to causal inference in information experiments due to the possibility of real-world pre-exposure to the information manipulated during the experiment (Gaines, Kuklinski, and Quirk 2006; Druckman and Leeper 2012; Linos and Twist 2018). Druckman and Leeper (2012) discuss cases where pre-exposure, if disregarded, can lead to erroneous conclusions that the information had no or minimal effect when, in reality, it had a substantial impact, and this impact would have appeared in the analysis if the pre-exposure status of the subjects in the experiment were considered. Linos and Twist (2018, p. 149) succinctly captures the problem, stating that when an experiment repeats information already received by respondents in the real world, failing to observe further changes in opinion during the study might lead to the incorrect conclusion that the message had no effect when in fact the effect happened during the prior exposure (Gaines, Kuklinski, and Quirk 2006; Chong and Druckman 2010; Slothuus 2016).

In many applications, measuring pre-exposure directly from the experimental units at the time of the experiment is not feasible or may lead to unintended consequences. One reason is that the act of measuring pre-exposure can serve as a form of treatment exposure. For instance, when investigating the impact of party cues on public support for policies (Barber and Pope 2019), asking individuals which party supports a particular policy (i.e., measuring pre-exposure to the party position on the issue) can prime individuals to think along partisan lines (exposure), potentially biasing subsequent responses and the cue-taking in the experiment. Furthermore, measuring pre-exposure after the outcome of interest is often impractical. In the example of partisan cues, it is not viable to measure pre-exposure after the experiment, as individuals in the treatment group have already been informed about the party's position on the issue, which can affect the answers to questions measuring pre-exposure.

This paper proposes a methodological solution to address the problem of pre-exposure in information experiments when measuring pre-exposure among experiment subjects is not feasible. Specifically, it a formalizes the pre-exposure problem using the potential outcomes framework, and uses that formalization to decompose the causal effect of information into its pre-exposure and exposure components. Then, I derive a close-form expression for the bias that arises when a naive average treatment effect (ATE) estimation is employed to investigate the effect of information, neglecting the pre-exposure status of the subjects in the experiment. The formalization and the ATE decomposition lead to formal definition of causal parameters that capture different quantities of interest related to information exposure and pre-exposure. Sufficient identification condition to estimate those parameters are presented. A bias-corrected estimator is introduced, and a Monte Carlo experiment is conducted to demonstrate the finite sample performance of the proposed solution. I illustrate the method with an application to investigate the effect of party cues on public policy support.

Decomposing the Average Treatment Effect

To examine the issues caused by pre-exposure to the treatment information, and how it can lead to a misinterpretation of the ATE when pre-exposure is ignored, I start by formulating the problem using the potential outcomes framework. Let $Y_i(d)$ indicate the outcome value for unit i under treatment value $D_i = d$. For simplicity, let us consider a binary treatment $d \in \{0, 1\}$, where $D_i = 1$ signifies that unit i received the information treatment, and $D_i = 0$ indicates that it did not. I assume SUTVA, positivity, and consistency (see online supplement for deatils) throughout the study, as costumary (Imbens and Rubin 2015).

A common practice in applied research is to estimate the ATE, which is defined as the difference between the average potential outcomes across all units if everyone had received the treatment and the average potential outcome if no one had. This causal parameter can be expressed as:

$$\tau^{\text{ATE}} = \mathbb{E}\left[Y_i(1) - Y_i(0)\right] = \mathbb{E}_X\left[\mathbb{E}\left[Y_i \mid D_i = 1, X_i\right]\right] - \mathbb{E}_X\left[\mathbb{E}\left[Y_i \mid D_i = 0, X_i\right]\right] \tag{1}$$

When researchers want to measure the impact of information exposure, the ATE fails to capture some quantities of interest. For instance, Druckman and Leeper (2012) examined whether support for a proposal to establish a state-owned casino is influenced by framing the proposal in positive or negative terms. They conducted an experiment where participants were randomly assigned to either positive or negative framing. The authors highlight the key problem with pre-exposure and using the ATE to access the information/framing effect: some individuals may have already formed strong opinions about the proposal before the experiment due to real-life prior information exposure. For these individuals, “another exposure [during the experiment] would have minimal additional effect” (pg. 886). This means that estimating the ATE, as shown in equation (1), will underestimate the effect of information as such because a portion of this effect—the one among pre-exposed subjects—occurred prior to the experiment during pre-exposure. The treatment and control groups in the experiment consist of individuals who had already been pre-exposed, resulting in minimal effects during the experiment for this pre-exposed group. What the ATE captures is a weighted average effect of exposure (among subjects not pre-exposed) and re-exposure (among pre-exposued subjects) to information, whereas the focus of interest often lies in the effect of exposure itself.

Since Druckman and Leeper (2012), the literature has empirically demonstrated that using a ATE naively can result in either an overestimation or underestimation of treatment exposure effects, due to unaccounted pre-exposure (Gaines, Kuklinski, and Quirk 2006; Chong and Druckman 2010; Slothuus 2016; Linos and Twist 2018; Clifford, Leeper, and Rainey 2023).

The main complication is that pre-exposure is often both uncontrolled by the researchers *and* not directly measurable among experimental subjects, as it can lead to treatment contamination. For instance, when investigating the impact of party cues on public support for policies (Barber and Pope 2019), asking individuals which party supports a particular policy (i.e., measuring pre-exposure to the party position on the issue) can prime individuals to think along partisan lines (exposure), potentially biasing subsequent responses and the cue-taking in the experiment. Furthermore, measuring pre-exposure after the outcome of interest is often impractical. It is not viable to measure pre-exposure after the treatment either, as individuals in the treatment group have already been informed about the party's position on the issue, which can affect the answers to questions measuring pre-exposure.

In sum, real-life information pre-exposure matters, and it typically cannot be measured or controlled by researchers during the experiment. This problem has not yet been analyzed using modern causal modeling tools. In the following discussion, I formalize this problem using the potential outcome framework. By formalizing the problem, we can gain a clearer understanding of its nature and define the relevant causal quantities involved. Furthermore, this formalization allows us to derive sufficient identification conditions and propose a feasible bias-corrected estimator that can be employed in cases where pre-exposure is unobservable among experimental subjects.

To start, denote t as the period that unit i was exposed to the information treatment. For simplicity, consider two periods, and use $t = 1$ to denote the time prior to the experiment and $t = 2$ the time of the experiment. I am not dealing with practical issues related to the decay of the pre-exposure effects, which require more substantive research and can be case-specific. The approach proposed is general enough, and these problems do not prevent its application. Future extensions can deal with the effect decay issue. Here, I use $t = 1$ and $t = 2$ as indexes to indicate pre-exposure status at the time of the experiment rather than the actual time that the person was pre-exposed.

We can write $D_i = (D_{1i}, D_{2i})$, and D_{ti} to denote the value of the exposure to the treatment for unit i at period t. Likewise, $d = (d_1, d_2) \in \{0, 1\}^2$ is a specific treatment status pair. I use $D_{1i} = 1$ to indicate that i had been pre-exposed to the treatment information, that is, had already been exposed to the information manipulated during the experiment prior to the experiment and therefore knew it already by the time of the experiment. I use $D_{2i} = 1$ to indicate that i was in the treatment group, that is, received the information during the experiment. The potential outcome of i becomes $Y_i(d_1, d_2)$, and there are four possibilities: $Y_i(0, 0)$ is the potential outcome if i had not received the information treatment prior to or during the experiment; $Y_i(1, 0)$ is the potential outcome if she was in the control group in the experiment but had been pre-exposed, $Y_i(0, 1)$ is the potential outcome if i had been treated but not pre-exposed, and finally, $Y_i(1, 1)$ is the potential outcome if i had been pre-exposed and then exposed again during the experiment.

Using that notation, we can define different causal parameters of interest. Let us consider a typical case in which the researcher controls exposure to the treatment during the experiment and can randomly assign units to each treatment group, but pre-exposure happens in real-life outside experimental conditions and is not under the researcher's control. Researchers are often interested in estimating the average causal effect of information exposure, which would require no prior exposure (Druckman and Leeper 2012; Clifford, Leeper, and Rainey 2023; Gaines, Kuklinski, and Quirk 2006). I call this parameter the *average information effect* (AIE). This is the implicitly parameter that is being underestimated by the ATE according to Druckman and Leeper (2012) discussion about the casino proposal framing effect under pre-exposure. AIE is also the target in Clifford, Leeper, and Rainey (2023), which investigates the effect of party cues on policy support under voters pre-exposure to party positions on salient issues. The AIE can be defined as:

$$\tau^{\text{AIE}} = \mathbb{E}\left[Y_i(0,1) - Y_i(0,0)\right] \tag{2}$$

In the same vein, we can define the average effect of re-exposure, that is, the effect in the experiment among those who have already been exposed to the treatment information in real life. This parameter is referred to as the average information re-exposure effect (AIRE). It can be defined as follows:

$$\tau^{AIRE} = \mathbb{E}\left[Y_i(1,1) - Y_i(0,0)\right] \tag{3}$$

In order to facilitate the subsequent analyses, it is also useful to define the average information pre-exposure effect (AIPE). The AIPE represents the average pre-exposure effect among the control group units. In a randomized experiment, this effect should be equivalent to the pre-exposure effect in the treatment group.

$$\tau^{AIPE} = \mathbb{E}\left[Y_i(1,0) - Y_i(0,0)\right] \tag{4}$$

Now, the first result of this paper establishes the relationship between the ATE and the other causal parameters just defined:

Proposition 2.1 (Average Treatment Effect Decomposition)**.** *Denote π_1 the proportion of people pre-exposed. The ATE can be decomposed into its AIE, AIPE, and AIPE parts as follows:*

$$\tau^{ATE} = \tau^{AIE} - \pi_1(\tau^{AIRE} - \tau^{AIE} - \tau^{AIPE}) \tag{5}$$

The proof of Proposition 2.1 can be found in the Appendix. Equation (5) demonstrates that the ATE is a result of combining the AIE, the AIPE, and the AIRE. However, if one were to use an unbiased estimator of the ATE to estimate the AIE, it would lead to a *pre-exposure* *bias*, whose magnitude is exactly $\pi_1(\tau^{AIRE} - \tau^{AIE} - \tau^{AIPE})$. Thus, the bias is a function of the proportion of people who were pre-exposed as well as the average effects of pre-exposure, re-exposure, and exposure under no pre-exposure. Equation (5) provides an explicit expression for what Druckman and Leeper (2012), Linos and Twist (2018), Slothuus (2016) and other authors allude to when they discuss overestimation or underestimation due to ignoring pre-exposure. Equation (5) makes it is evident that a "naive" ATE can lead to either overestimation or underestimation of the AIE unless there is no pre-exposure ($\pi_1 = 0$).

Identification Conditions

This section discusses the conditions that, if met, are sufficient for the identification of AIE, AIPE, and AIRE. Let us start by assuming that during the experiment, pre-exposure (D_{1i}) can be measured directly from each unit i, and that researchers only control the exposure (D_{2i}), while pre-exposure happens in real life and is beyond researchers' control. I will address cases in which pre-exposure is not directly observable in the following section.

For identification of the pre-exposure causal parameters (AIE, AIPE, and AIRE), a sufficient condition is that the pre-exposure is conditionally ignorable given a set of measurable background variables X_i. This is a particular version of the strong ignorability assumption proposed by Rosenbaum and Rubin (1983), which is widely used in observational studies and matching methods. It can be stated as follows:

Assumption 3.1 (*Pre-exposure conditional ignorability* (PECI))**.** *Pre-exposure is independent of the potential outcomes given a set of covariates X_i. Formally,*

$$Y_i(d_1, d_2) \perp\!\!\!\perp D_{1i} \mid X_i = x \quad , \quad \forall d = (d_1, d_2) \in \{0,1\}^2, x \in \mathcal{X}$$

Theorem 3.1 provides identification results for the general causal parameter $\tau^{PE}(d, d')$ and its special cases (AIE, AIPE, and AIRE). The proof is in the Appendix.

Theorem 3.1 (Nonparametric identification of causal effects with observed pre-exposure)**.** *Under assumption 3.1, the general causal parameter $\tau^{PE-ATE}(d, d')$ defined in* (**??**) *and its special cases (AIE, AIPE, and AIRE) are nonparametric identifiable with $\tau^{PE-ATE}_{PECI}(d, d')$ defined as follows:*

$$\tau_{PECI}^{PE-ATE}(d,d') = \mathbb{E}_X\left[\mathbb{E}\left[Y_i \mid D_{1i}=d, D_{2i}=1, X_i\right]\right] - \mathbb{E}_X\left[\mathbb{E}\left[Y_i \mid D_{1i}=d', D_{i2}=0, X_i\right]\right] \quad (6)$$

When pre-exposure is observed and PECI holds, it is straighforward to estimate the causal parameters. One can estimate them non-parametrically using a plug-in estimator by comparing averages based on exposure and pre-exposure status for each demographic group $X_i = x$, or parametrically using a linear model with an interaction term between exposure (D_{1i}) and pre-exposure (D_{2i}).

Now, let us consider a scenario where pre-exposure cannot be directly observed. As discussed previously, this is a common situation due to the risks that the process of measuring pre-exposure poses in terms of contaminating or being contaminated by the treatment assignment or outcome measurement. In such cases, the central problem for applied research is to determine whether and how the pre-exposure average treatment effect (PE-ATE) parameters can be estimated. When pre-exposure cannot be observed, the fundamental problem of causal inference (Holland 1986) gets aggravated because none of the potential outcomes are observable.

Table 1 illustrates this problem using three hypothetical subjects (rows). Consider the first row (subject $i = 1$), which has been assigned to receive the treatment during the experiment ($D_{21} = 1$). The observed outcome for that subject is $Y_1 = 2$. As we cannot observe pre-exposure (D_{11}), we do not know if this outcome corresponds to $Y_1(0,1)$ or $Y_1(1,1)$. The same reasoning applies to the other cases. Hence, we cannot recover any pre-exposure causal parameters. To estimate the PE-ATE, we need to recover the pre-exposure status of the subjects. Thus, the challenge lies in determining the PE-ATE parameters when direct observations of subjects' pre-exposure status are not feasible.

Table 1: Hypothetical scenario illustrating the fundamental problem of causal inference (Holland 1986) under missing information (N/A) on treatment pre-exposure (D_{i0}).

Hypothetical Data					Potential Outcomes			
i	X_i	D_{i0}	D_{i1}	Y_i	$Y_i(0,0)$	$Y_i(1,0)$	$Y_i(0,1)$	$Y_i(1,1)$
1	7	N/A	1	2	×	×	?	?
2	-3	N/A	1	3	×	×	?	?
3	4	N/A	0	6	?	?	×	×

Note: Assuming SUTVA and consistency, for each row we know with certainty that the observed outcome Y_i do not correspond to the potential outcomes marked with × in the respective row, but we don't know which ones with quotation mark (?) in the respective row Y_i captures.

Feasiable Estimators under Unobserved Pre-Exposure

When subjects' pre-exposure cannot be directly observed among experimental subjects, one feasible methodological solution—there might be others that can be developed in future work, relaxing parametric assumptions—is to use a parametric regression model and subjects' pre-exposure predicted status (PEPS). This combination and the method proposed below leads to an unbiased estimator of the AIE.

To motivate the solution, consider first the case in which is possible to directly observe pre-exposure. Let us focus on cases with continuous outcome y. In this case, we can capture the causal quantities using the following parametrization, with additive separable covariates X_i: $\mathbb{E}\left[Y_i(0,0) \mid X_i\right] = \beta_0+\beta_x X_i$, $\mathbb{E}\left[Y_i(0,1) \mid X_i\right] = \beta_0+\beta_x X_i+\tau^{AIE}$, $\mathbb{E}\left[Y_i(1,0)\right] = \beta_0+\beta_x X_i+\tau^{AIPE}$, and $\mathbb{E}\left[Y_i(1,1)\right] = \beta_0+\beta_x X_i+\tau^{AIE}+\tau^{AIPE}+\gamma$. Let ϵ_i denote the individual-specific deviation from the outcome's mean for subject i. If pre-exposure was observed and assumption 3.1 holds, we can recover the causal parameters by estimating the following model using an ordinary least square (OLS) estimator of the regression coefficients:

$$y_i = \beta_0 + \tau^{AIE} D_{2i} + \tau^{AIPE} D_{1i} + \gamma D_{1i} D_{2i} + \beta_x X_i + \epsilon_i \tag{7}$$

When pre-exposure is not observed, the corresponding model for the ATE that averages out pre-exposure is:

$$y_i = \beta_0' + \tau^{ATE} D_{2i} + \beta_x' X_i + \epsilon_i' \tag{8}$$

When D_{2i} is randomized, the ATE can be estimated without bias using the OLS estimator. However, if the aim is to estimate the AIE parameter, for instance, but pre-exposure could not be measured directly for each unit i in the experiment, then using the ATE to estimate the AIE results in *pre-exposure bias*, as discussed in Druckman and Leeper (2012). Clearly, from model (7), we have $\gamma = \tau^{AIRE} - \tau^{AIE} - \tau^{AIPE}$. From proposition 2.1 and the fact that $\widehat{\tau}^{ATE}$ is an unbiased estimator of τ^{ATE}, we see that the *pre-exposure bias* is $\pi_1 \gamma$, that is, the product of the proportion of pre-exposure individuals and the parameter capturing the interaction effects between exposure and pre-exposure from model (7) (proof in the online supplement):

$$\text{Bias}(\widehat{\tau}^{ATE}, \tau^{AIE}) = \gamma \pi_1 \tag{9}$$

To address the issue of *pre-exposure bias* when we cannot observe pre-exposure directly, I propose a solution that utilizes the pre-exposure probability as a function of subject background characteristics, denoted as $p(D_{1i} \mid X_i)$. This probability is not influenced by the treatment status (D_{2i}) or outcome (y_i). Hence, we can calculate it by drawing an auxiliary random sample from the same population as the experimental sample. In this ancillary sample, we solely need to measure pre-exposure (D_{i1}) and the background variables (X_i), without any interference from treatment or outcome, avoiding contamination of the treatment or outcome due to measuring pre-exposure, or vice-versa. We can then apply the probability $p(D_{1i} \mid X_i)$, computed using the auxiliary sample, to classify the subjects in the experiment as pre-exposed or not, based on their observed background features X_i.

Define the predicted pre-exposure status $\widehat{D}_{1i}$ as:

$$\widehat{D}_{1i} = \underset{d}{\operatorname{argmax}}\, p(D_{1i} = d \mid X_i) \tag{10}$$

Note that 11 chooses the pre-exposure status d that maximizes the probability of pre-exposure, that is, $\widehat{D}_{1i} = 1$ whenever $p(D_{1i} = 1 \mid X_i) > p(D_{1i} = 0 \mid X_i)$. Next, define the following estimators as a function of the predicted pre-exposure status:

Definition 4.1 (PEPS-ATE estimator)**.** *The pre-exposure predicted status average treatment effect (PEPS-ATE) estimator is an estimator constructed using the predicted pre-exposure status $\widehat{D}_{1i}$ instead of the true status D_{1i} to produce the estimates. That is,*

$$(\widehat{\tau}^{AIE}, \widehat{\tau}^{AIPE}, \widehat{\tau}^{AIRE}, \widehat{\beta}) = f(D_2, \widehat{D}_1, X)$$

Using the predicted status $\widehat{D}_1$ is useful because it allows us to derive an unbiased estimator of the target parameters. Note first that if there is no misclassification of the subjects in the experiment, then we can use $\widehat{D}_{1i}$ instead of D_{1i} and still recover all parameters, such as AIE, using model (7). But in practice, there is likely to be some misclassification of pre-exposure with this approach. However, even with severe misclassification, when aiming to recover the AIE, utilizing the predicted pre-exposure status $\widehat{D}_{1i}$ and the PEPS estimator bring significant advantages. The main advantage is that it changes the nature of the bias. Instead of a *pre-exposure bias* due to completely omitting pre-exposure (omitted variable bias), the PEPS estimator leads to a *misclassification bias*, which arises due to the misclassification associated with both false positive and false negative cases. Therefore, the problem caused by unmeasured pre-exposure can be reformulated in terms of a misclassified covariate and measurement error. And while we may not be able to eliminate the misclassification error, we can eliminate the asymptotic bias by using a misclassification bias-corrected PEPS estimator. That is, reformulating the unmeasured pre-exposure problem as a misclassified pre-exposure problem enables us to extend existing solutions that address the inconsistency of OLS estimators in the case of misclassified binary covariates (Savoca 2000; Black, Sanders, and Taylor 2003; Greene 2012; Yi and He 2017; Aigner 1973), and use it to deal with bias due to uncontrolled and unmeasured pre-exposure. This new problem is formally equivalent to dealing with the asymptotic bias caused by covariates that are subject to measurement error.

To see why this is advantageous, I follow Aigner (1973) and model the misclassification using a misclassification error $U \in \{-1, 0, 1\}$.

$$\widehat{D}_{1i} = D_{1i} + U \tag{11}$$

Let us define the probabilities of false positive and false negative as $P(\widehat{D}_{1i} = 1 \mid D_{1i} = 0) = \alpha_0$ and $P(\widehat{D}_{1i} = 0 \mid D_{1i} = 1) = \alpha_1$, respectively. These probabilities correspond to a subject being misclassified as pre-exposed when they are not, and as not pre-exposed when they are, respectively. Consider the following *operational model* obtained from the true model (7) after replacing the unobserved pre-exposure variable D_{1i} with its error-prone predicted value $\widehat{D}_{1i}$ (from equation (11)):

$$y_i = \beta_0 + \tau^{AIE} D_{2i} + \tau^{AIPE} \widehat{D}_{1i} + \gamma \widehat{D}_{1i} D_{2i} + \beta_x X_i + (\epsilon_i - \tau^{AIPE} U_i - \gamma U_i D_{2i}) \tag{12}$$

Denote $\widehat{M} = (D_2, \widehat{D}_1, D_2\widehat{D}_1, X)$ a $n \times k$ matrix where k is the number of covariates, including $\widehat{D}_1$, D_2, and their interaction. Denote $\widehat{\theta}_m = (\widehat{\tau}_m^{AIE}, \hat{\tau}_m^{AIPE}, \widehat{\gamma}_m, \widehat{\beta}_x)$ the PEPS estimator with misclassified cases of pre-exposure. Then, the asymptotic bias of $\widehat{\theta}_m$ is (proof in the Appendix):

$$\text{Asy.Bias}(\widehat{\theta}_m) = -\operatorname{plim}\left[(\widehat{M}^T\widehat{M})^{-1}(\tau^{AIPE}\widehat{M}^T U + \gamma \widehat{M}^T (U^T \boldsymbol{I} D_2))\right] \tag{13}$$

where $\boldsymbol{I}$ denotes the $n \times n$ identity matrix. Thus, the asymptotic bias depends on the misclassification error U and the misclassified pre-exposure. In the absence of these factors (e.g., with no misclassification error), the misclassification bias would not exist.

Measurement error in one covariate is sufficient to bias the OLS estimates of all coefficients, except for the coefficient of a covariate that is orthogonal to the other covariates and it is correctly measured (Aigner 1973; Greene 2012). Although I am assuming randomization and therefore orthogonality of exposure during the experiment, the interaction between exposure and pre-exposure is sufficient to bias the OLS estimator of the AIE if pre-exposure is ignored or misclassified, as demonstrated in expession (9).

The proposed solution, as presented in Theorem 4.1, accounts for the asymptotic bias in OLS estimates resulting from the misclassification of subjects' pre-exposure status. Theorem

4.1, together with the formulation of the problem in terms of potential outcomes, the causal identification analysis, and the identification assumption stated in 3.1, allow us to estimate the AIE consistently, and the estimated quantity has the desired causal effect interpretation.

Theorem 4.1 (Bias-corrected PEPS-ATE Estimator)**.** *Denote the biased and inconsistent OLS PEPS estimators* $(\widehat{\tau}_m^{AIE}, \widehat{\tau}_m^{AIPE}, \widehat{\gamma}_m)$ *computed from model* (7) *using the predicted pre-exposure status* $\widehat{D}_{1i}$ *from* (11) *instead of the true pre-expoure* D_{1i}*, where* $\widehat{D}_{1i}$ *is subject to false negative and false positive classification rates* α_1 *and* α_0*, respectively. Denote* $\hat{\alpha}_0$*,* $\hat{\alpha}_1$*, and* $\widehat{\pi}_1^{(M)}$ *consistent estimators of the respective parameters, where* $\pi_1^{(M)}$ *is the proportion of subjects classified as pre-exposed. Let* $\hat{\pi}_2$ *be the proportion of subjects exposed (i.e., in the treatment group) during the experiment. The following estimators* $(\widehat{\tau}_c^{AIE}, \widehat{\gamma}_c)$ *of the causal parameters* (τ^{AIE}, γ) *defined in* (2) *and* (7) *are asymptotically unbiased and asymptotically Gaussian:*

$$\widehat{\tau}_c^{AIE} = \widehat{\tau}_m^{AIE} + a\left[\left(\frac{1}{1-c}\right)\widehat{\tau}_m^{AIPE} + \left(\frac{d}{1-c}\right)\widehat{\gamma}_c\right] + b\widehat{\gamma}_c \tag{14}$$

$$\widehat{\gamma}_c = \left[\frac{1}{(1-f)(1-c)-ed}\right]\left[(1-c)\widehat{\gamma}_m + e\widehat{\tau}_m^{AIPE}\right] \tag{15}$$

where,

$$\widehat{M} = (D2, \widehat{D}_1, D_2\widehat{D}_1, X)^T; W = \text{plim}(\widehat{M}^T\widehat{M}/n)^{-1}; w_k = [W]_{k\cdot}\,\textit{(}k^{th}\textit{ row of }W\textit{)}$$

$$a = w_1\hat{\rho}_{AIPE} \quad , \quad b = w_1\hat{\rho}_{\gamma} \quad , \quad c = w_2\hat{\rho}_{AIPE} \quad , \quad d = w_2\hat{\rho}_{\gamma} \quad , \quad e = w_3\hat{\rho}_{AIPE} \quad , \quad f = w_3\hat{\rho}_{\gamma}$$

$$\hat{\rho}_{AIPE} = \begin{bmatrix} 0 \\ (\hat{\eta}+\hat{\nu})\hat{\sigma}_p^2 \\ \hat{\pi}_2(\hat{\eta}+\hat{\nu})\hat{\sigma}_p^2 \\ \hat{\Phi}\hat{\rho}_{px} \end{bmatrix} \quad ; \quad \hat{\rho}_{\gamma} = \begin{bmatrix} \hat{\Psi}\hat{\sigma}_E^2 \\ \hat{\pi}_2(\hat{\eta}+\hat{\nu})\hat{\sigma}_p^2 \\ \hat{\phi} \\ \hat{\pi}_2\hat{\Phi}\hat{\rho}_{px} \end{bmatrix} \quad ; \quad \hat{\eta} = \frac{\hat{\alpha}_1\left(\hat{\pi}_1^{(m)} - \hat{\alpha}_0\right)}{(1-\hat{\pi}_1^{(m)})(1-\hat{\alpha}_0-\hat{\alpha}_1)}$$

$$\hat{\nu} = \frac{\hat{\alpha}_0(1-\hat{\alpha}_1-\hat{\pi}_1^{(m)})}{\hat{\pi}_1^{(m)}(1-\hat{\alpha}_0-\hat{\alpha}_1)} \quad ; \quad \hat{\Phi} = -\left(\frac{\hat{\alpha}_0+\hat{\alpha}_1}{1-\hat{\alpha}_0-\hat{\alpha}_1}\right) \quad ; \quad \hat{\phi} = \hat{\pi}_2\hat{\pi}_1^{(m)}(\hat{\nu}-\hat{\pi}_2\hat{\Psi})$$

$$\hat{\Psi} = \hat{\nu}\hat{\pi}_1^{(m)} - \hat{\eta}(1-\hat{\pi}_1^{(m)}) \quad ; \quad \hat{\rho}_{px} = \mathbb{C}ov\left[X, \widehat{D}_1\right] \quad ; \quad \hat{\sigma}_p^2 = \hat{\pi}_1^{(m)}(1-\hat{\pi}_1^{(m)}) \quad ; \quad \hat{\sigma}_E^2 = \hat{\pi}_2(1-\hat{\pi}_2)$$

The proof for Theorem 4.1 is extensive and is provided in the online supplement. In essence, the Theorem indicates that we can obtain a consistent estimator for the *average information effect* (AIE) by utilizing the PEPS estimators, even without directly observing pre-exposure and in spite of potentially inaccurate predictions of pre-exposure status due to misclassification. One important advantage is that there is no *a priori* restriction on how severe the misclassification might be. The approach remains viable as long as we have consistent estimators for (1) the (misclassified) proportion of pre-exposed subjects ($\pi_1^{(M)}$), which can be achieved through a plugin estimator, and (2) the misclassification rates α_0 and α_1, which can be easily obtained via maximum likelihood estimator (Hausman, Abrevaya, and Scott-Morton 1998) or cross-validation techniques (Fu, Carroll, and Wang 2005; Ounpraseuth et al. 2012; Bates, Hastie, and Tibshirani 2023). The application in this paper uses the latter.

In summary, the main contributions of this paper are: (1) formulating the pre-exposure problem in terms of potential outcomes; (2) defining the main target parameters of interest that have been informally discussed in the previous literature (AIE, AIPE, and AIRE); (3) demonstrating that the ATE can be decomposed in terms of those parameter; (4) deriving a sufficient identification condition to estimate these parameters; (5) deriving a closed-form expression for the *pre-exposure bias* that emerge if one ignores pre-exposure; and (6) proposing an estimator for consistently recovering the AIE when pre-exposure cannot be measured directly among subjects in the experimental sample due to risk of contamination. The proposed method functions by predicting the pre-exposure status of the experiment participants to deal with the *pre-exposure bias* when using a “naive” ATE to estimate the AIE, and then employing an estimator (bias-corrected PEPS-ATE) that corrects the remaining *misclassification bias* due to pre-exposure prediction misclassification. I introduced an asymptotic bias-corrected PEPS estimator that is consistent even when there is severe pre-exposure misclassification. Standard errors can be obtained through the delta method or bootstrap (Davison and Hinkley 1997; Wasserman 2013). The application in the paper uses the latter.

Sampling Procedures

When practical implementation is concerned, two decisions need to be made when using the method proposed here. The first is selecting a model for $p(\widehat{D}_{1i} \mid X_i)$. This model remains unspecified in this work since the bias-corrected estimator remains applicable even with an model that produces severe misclassification. Therefore, researchers can select a model of their preference, such as logistic regression.

The second important decision concerns the sampling procedure. One possible approach is to use a *split-module sampling scheme*. Essentially, this scheme randomly allocates subjects into two distinct samples. In the first sample, which I call the *experimental sample*, subjects participate the experiment, following usual procedures. In the second sample, which I call the *PEPS sample*, subjects do not answer the outcome question and are not exposed to any treatment condition. This module's sole purpose is to gauge background variables (X_i) along with prior exposure information. Note that this procedure isn't strictly necessary for implementing the method, with the only requirement being that assumption 3.1 is satisfied. Nonetheless, this sampling scheme facilitates balanced samples in both observable and unobservable variables. At a minimum, subjects in both samples need to originate from the same base population and be sampled utilizing the same procedure. Therefore, the *PEPS sample* can still be gathered post the completion of the experiment.

Next, for each sample, one needs to decide about the sample size. Let us consider the *experimental sample* first. Although several studies have presented results for sample size calculation in regression analysis with misclassified variables (Lachenbruch 1968; Rahme, Joseph, and Gyorkos 2000; Edwards et al. 2005; Cheng, Stamey, and Branscum 2009; Beleites et al. 2013; Riley et al. 2020), none of these results are directly applicable to the proposed model. Due to the complexity of sample size calculation in this case, addressing it formally is beyond the limits and scope of this paper.

To aid practitioners with the sample size selection for their study, I conducted various simulation analyses whose details are in the online supplement. Briefly, I evaluated the average increase in the size of the confidence interval of the bias-corrected PEPS estimator relative to the confidence interval of the corresponding model where the true pre-exposure was correctly observed. This enables us to determine the additional sample size required for analyses using predicted pre-exposure to achieve a confidence interval that is similar in size to what is needed for analyses using OLS with true pre-exposure observed. The exercise showed that in order for the bootstrap-based confidence interval of the corrected pre-exposure predicted status average treatment effect (PEPS-ATE) estimator to match the confidence interval of a OLS estimator calculated using real pre-exposure measures, the sample size must be at least 4.1 times greater. Therefore, practitioners can conduct a standard power analysis for analyzing a single coefficient

Table 2: Steps to estimate the *average information effect* (AIE) using the asymptotic bias-corrected pre-exposure predicted status average treatment effect (PEPS-ATE) estimatior and a *split-module sampling procedure*. Simulations indicate a good performance with $n_{peps} = 150$ and a factor $k = 4.1$.

```
Steps
 1: Compute the experimental sample size $n_{\text{exp}}$ using usual power analysis to test hypothesis
    about a regression coefficient
 2: Split-sampling scheme:
 3:    Randomly assign k $\times n_{exp}$ subjects to the experimental sample
 4:    Randomly assign $n_{peps}$ subjects to the PEPS sample
 5: Using the PEPS sample, measure $X$ (covariates) and $D_1$ (pre-exposure) only
 6: Using the experimental sample, measure $X$ (covariates), $D_2$ (exposure), and $y$ (outcome)
 7: Estimate the probability of pre-exposure $p(D_{1i} \mid X_i)$ using the PEPS-sample
 8: Compute the predicted probabilities $\widehat{p}(D_{1i} \mid X_i)$ for subjects in the experimental sample
 9: Compute the predicted pre-exposure status of experiment subjects using expression (11)
10: Estimate model (7) using $\widehat{D}_{1i}$ instead of $D_{1i}$
11: Estimate the misclassification rates $\alpha_0$ and $\alpha_1$ using a consistent estimator
12: Estimate the proportion classified as pre-exposed $\pi_1^{(M)}$ using a consistent estimator
13: Compute the bias-corrected estimators in Theorem 4.1
14: Compute the confidence intervals
```

in a multivariate linear regression (Cohen et al. 2013) and then increase the sample size by a factor of 4.1 or more.

Regarding the PEPS sample, researchers can follow the usual sampling size guidelines for classification methods (Figueroa et al. 2012). The proposed method, however, does not require accurate classification, meaning the sample size may be much smaller than when predictive performance is required. The simulations show that a sample size of 150 with three background covariates is sufficient for good performance. Table 2 outlines the recommended steps.

Simulation Study

This section shows a small-scale Monte Carlo exercise as in Imai, Keele, and Yamamoto (2010) to demonstrate the results presented in the previous section and investigate the finite sample performance of the three main estimators discussed, namely, the ATE, which leads to *pre-exposure bias* when used to estimate the *average information effect* (AIE); the PEPS-ATE, which is not subject to *pre-exposure bias* but to *misclassification bias* because it uses predicted pre-exposure but does not correct for prediction misclassification; and the PEPS-ATE (BC), which corrects both biases. I assume the goal is to recover the AIE, as in Druckman and Leeper

(2012) and Clifford, Leeper, and Rainey (2023), and include the both PEPS-ATE estimators to demonstrate the difference between the *pre-exposure bias* and the *misclassification bias.* I assume that 3.1 holds in the data generating process throughout. The details of the simulation can be found in the online supplement due to space constraints.

Figure 1 compares the estimators for different true values of the parameters τ^{AIE}, τ^{AIPE}, and γ. As discussed in the previous section, the *pre-exposure bias* due to using the ATE to estimate the AIE is exactly $\gamma\pi_1$. The solid black lines with square marks in Figure 1 demonstrate this bias. Regardless of the true values of τ^{AIPE} (column panels) or τ^{AIE} (row panels), the bias can be positive or negative depending on the value of γ, which captures the interaction effect of pre-exposure and exposure. This simulation result confirms the formal derivation from prior sections and supports arguments in the existing literature that ATE not only underestimates the true AIE, as discussed in Druckman and Leeper (2012), but can also overestimate it, as shown in applied examples by Linos and Twist (2018). The results here provide a more general closed-form expression supporting their empirical findings, and Figure 1 illustrates that the bias can be zero if γ is zero. It can also be zero if there is no pre-exposure ($\pi_1 = 0$).

The Figure 1 reveals some other important results. The PEPS-ATE estimator improves over (i.e., decreases the bias when compared to) the ATE. However, this improvement does not hold in the exceptional cases that I will discuss below. To emphasize this improvement, I have highlighted the *pre-exposure bias* in the top-left panel, which originates from omitting pre-exposure information when utilizing a “naive” ATE estimator to learn about the AIE. I have also highlighted the *misclassification bias* resulting from utilizing the PEPS-ATE when there is pre-exposure misclassification of experiment subjects. As we can see, the PEPS-ATE (BC) estimator eliminates both biases.

Figure 2 analyzes the estimators’ performance as a function of the misclassification rates α_0 and α_1. Notably, the pre-exposure bias (ATE bias) is lower than the misclassification bias (PEPS-ATE bias) only in cases where $\alpha_0 + \alpha_1 > 1$. The right panel of Figure 2 highlights this difference in absolute bias values. As the on-line supplment shows, $\alpha_0 + \alpha_1 < 1$ is equivalent to a positive correlation between $\widehat{D}_1$ and D_1. The PEPS-ATE (BC) doesn’t suffer from this limitation and rectifies both the misclassification and the pre-exposure biases, irrespective of how poor the pre-exposure classification is, as evident in Figure 2.

Table 3 presents the results of a Monte Carlo experiment consisting of ten thousand it-

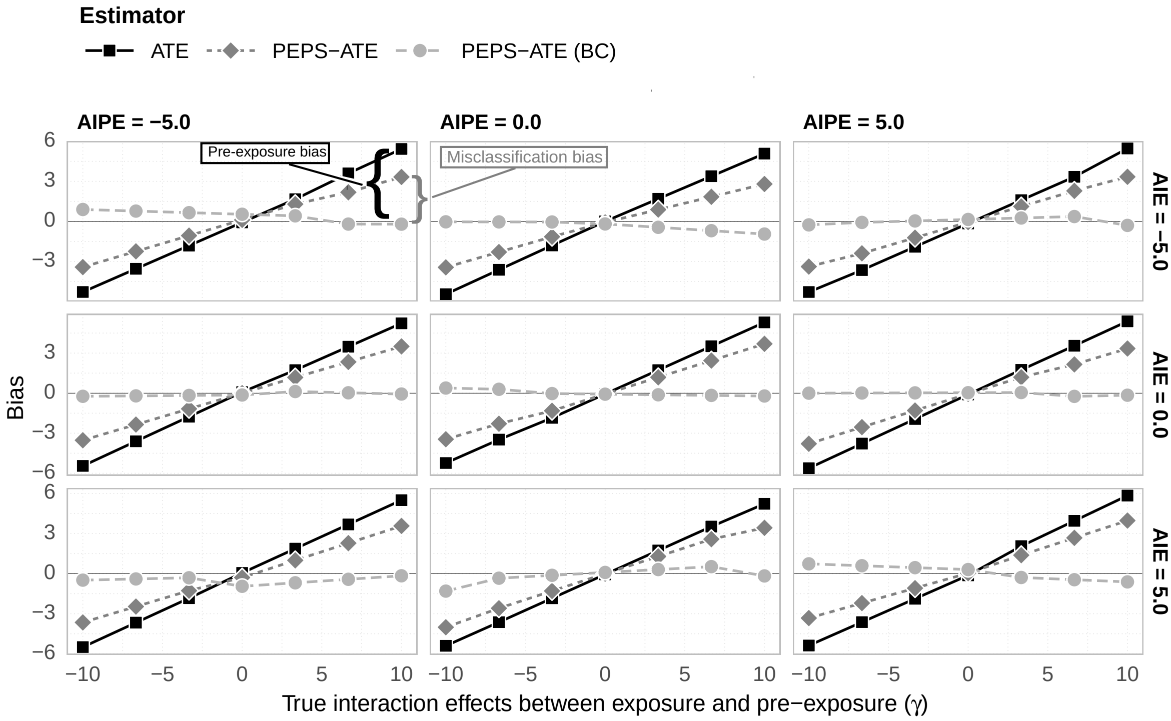

Figure 1: Comparing the bias in the estimators of the AIE for various true values of the causal parameters τ^{AIE}, τ^{AIPE}, and γ. Simulated values use $\pi_1 = 0.54$ (min: 0.52, max: 0.56); $\alpha_0 = 0.42$ (min: 0.38, max: 0.46); $\alpha_1 = 0.28$ (min: 0.25, max: 0.31).

erations. This table compares the performance of the ATE, the PEPS-ATE, the PEPS-ATE (BC), and an OLS estimator computed using the true pre-exposure status of subjects in the experiment. The performance of the PEPS-ATE (BC) estimators are comparable to the OLS estimator in terms of coverage and average estimated bias. The coverage levels reach the expected rate of 95%, and it is evident that this is not due to excessively large standard errors. Although the standard error of the PEPS-ATE (BC) estimator is expected to be greater than those from the true model, the difference is not particularly large.

In summary, the PEPS-ATE (BC) estimator corrects bias even in cases of severe misclassification. It recovers the true value of the target parameter (AIE), even when pre-exposure is not observable, at a rate comparable to situations in which true pre-exposure was observed.

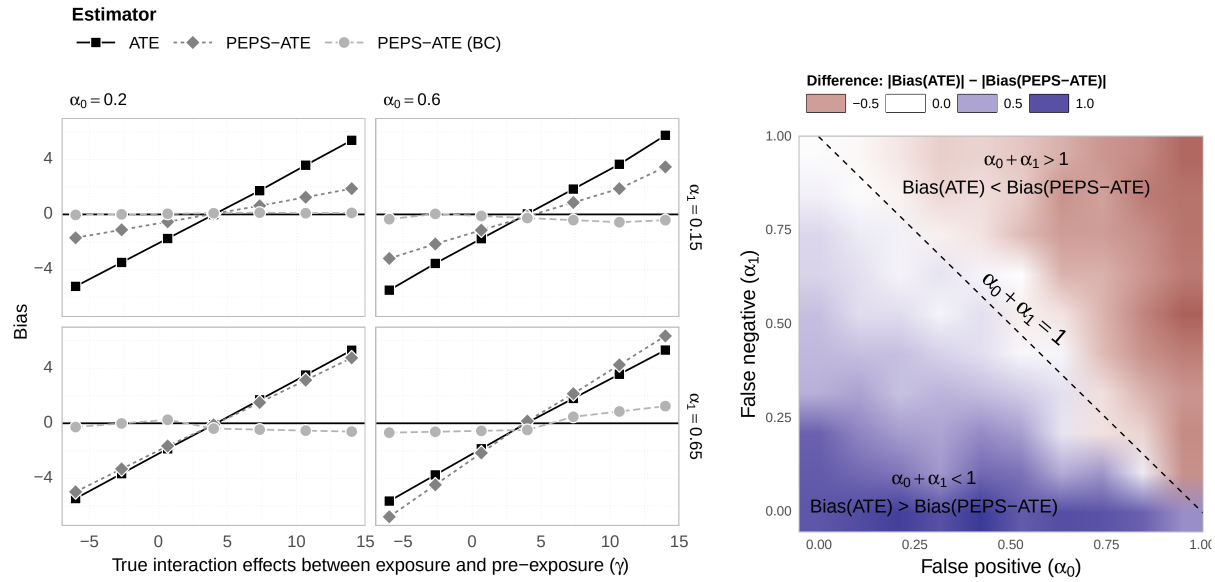

Figure 2: Comparing estimators of the average information effect (AIE) as function of misclassification rates (α_0 and α_1) and interaction effects between exposure and pre-exposure (λ). Simulated values use $\tau^{AIE} = \tau^{AIPE} = 2$ for the left panel and $\tau^{AIPE} \in \{-2, 0, 2\}$ for the right panel, with average pre-exposure rate ($pi_1^{(m)}$) of 0.54 (min: 0.53, max: 0.56).

Table 3: Finite Sample Performance of the ATE, PEPS-ATE, Bias-corrected PEPS-ATE, and True Model OLS Estimators for Various Sample Sizes. The a Monte Carlo Experiment Used Ten Thousand Iterations with $(\tau^{AIE}, \tau^{AIPE}, \gamma) = (2.1, 1.2, 0.7)$ and the same other procedures as in Figure 1.

	Sample	95% CI Coverage		Average Std. Error		Bias	
Estimator	Size	τ^{AIE}	γ	τ^{AIE}	γ	τ^{AIE}	γ
ATE	1000	0.0020	—	0.078776	—	0.378441	—
	1500	0.0000	—	0.064237	—	0.380618	—
	2000	0.0000	—	0.055619	—	0.379430	—
PEPS-ATE	1000	0.4573	0.1355	0.123852	0.160822	0.250684	-0.483209
	1500	0.3004	0.0460	0.100968	0.130907	0.248058	-0.475531
	2000	0.1817	0.0103	0.087285	0.113300	0.249331	-0.479456
PEPS-ATE (BC)	1000	0.9479	0.9504	0.298010	0.527781	-0.005066	0.004977
	1500	0.9459	0.9450	0.239372	0.423504	-0.012302	0.022308
	2000	0.9477	0.9496	0.205802	0.364184	-0.004196	0.006024
True model (pre-exposure observed)	1000	0.9545	0.9605	0.093882	0.127482	0.003797	-0.005349
	1500	0.9450	0.9517	0.076526	0.103859	0.000329	0.001525
	2000	0.9486	0.9515	0.066256	0.089928	-0.000110	-0.000610

Applied Example

I apply the proposed method to real data using the experimental study conducted by Clifford, Leeper, and Rainey (2023). The authors explore the impact of party cues on partisans' policy support. Party cues are brief pieces of information regarding the policy stances of political parties. The term "party cues effects" refers to the inclination of partisans to support (or reject) a policy when they are informed that their favoured party advocates (or opposes) that policy (Bullock 2020, 2011; Barber and Pope 2019).

Clifford, Leeper, and Rainey (2023) note that there is a substantial heterogeneity in party cue effects depending on the policy topic. They demonstrate that this dispersion in experimental results is strongly associated with variations in subjects' pre-exposure to party positions, which can vastly differ depending on the policy salience. They measured the effect of party cues and pre-exposure to party policy positions for 48 policies with varying degree of salience.

To account for pre-exposure, the authors made use of the ancillary module (the PEPS sample), wherein participants neither partook in the experiment nor answered the outcome question. Instead, they responded to queries pertaining to their awareness of party policy positions. The goal was to obtain estimates of prior exposure "that could not be influenced by the experiment" (Clifford, Leeper, and Rainey 2023). A total of 2,764 interviews were collected for the experimental sample, and 252 for the ancillary module. More details of their analysis can be found in the online supplement for lack of space.

Figure 3 compares different estimates produced using their data. Consider the left panel first. It compares the observed policy-level rates of pre-exposure in the PEPS sample (x-axis) for each policy against the predicted pre-exposure rates in the experimental sample (y-axis) computed by Clifford, Leeper, and Rainey (2023) (gray dots) and by the proposed method (black dots). The latter uses individual-level prediction, while the fomer uses policy-level aggregate rates, and this difference explains the higher variability of the predictions (see online supplement for details). The covariates available include race, sex, age, education, partisanship (ANES), marital status, a social identity measure of party identification, an indicator of being Hispanic, and ideology. The strong correlation depicted in Figure 3 between observed pre-exposure in the PEPS sample and predicted pre-exposure in the experimental sample is remarkable. We should expect this correlation because individuals were ramdonly assigned to the experimental and PEPS samples.

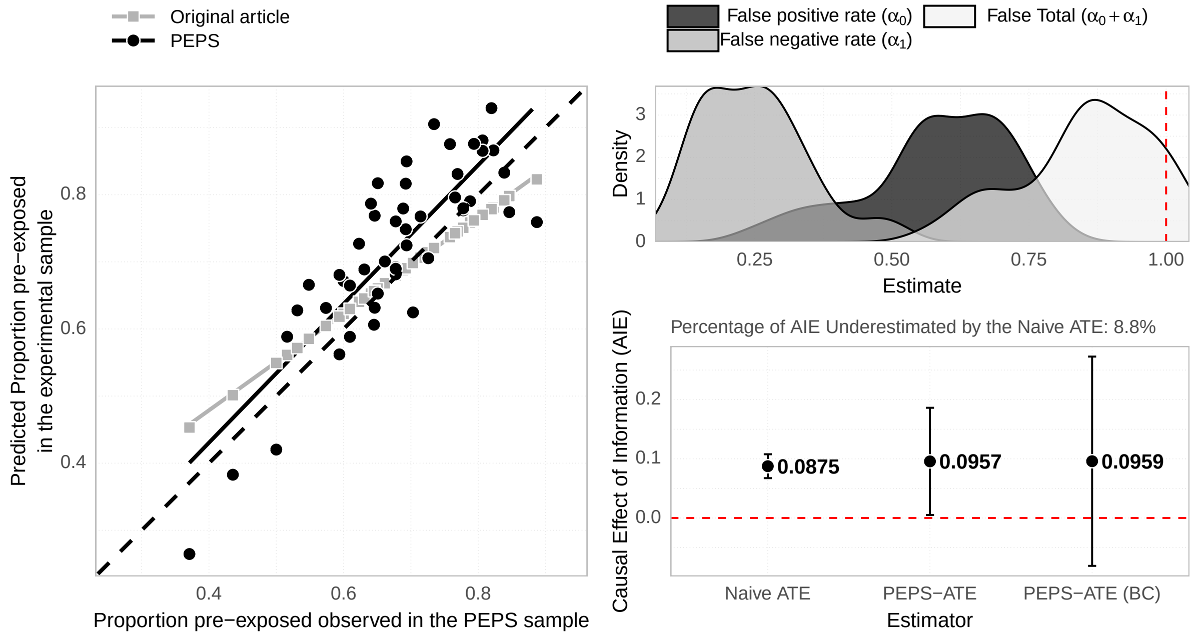

Figure 3: Left panel compares the predicted rates of pre-exposure; top-right panel shows the distribution of misclassification rates in the estimation; bottom-right compares the estimated values of the AIE using the "naive" ATE and the two estimators proposed in this article.

The top-right panel of Figure 3 shows the misclassification rates (α_0 and α_1) for each policy. As illustrated in Figures 1 and 2, using predicted pre-exposure without adjusting for misclassification can decrease pre-exposure bias, unless the combined rates of false positives and false negatives surpass 1. The top-right panel reveals that this occurred in some cases (8.3 percent, or 4 out of the 48 policies in the sample). In such a scenario, using a predictive pre-exposure without a correction for misclassification would exacerbate the bias when estimating the AIE. The bottom-right panel displays the naive ATE, PEPS-ATE, and PEPS-ATE (BC) estimates. In this application, we can observe that the naive ATE underestimates the AIE by 8.8 percent ((0.0959 - 0.0875)/0.0959).

Note that the confidence interval for the PEPS-ATE (BC) is larger and, for this particular application, the percentage change in the estimated values due to pre-exposure bias is small (8.8 percent). Given these results, one might object to the proposed method, arguing that the reduction in bias does not compensate for the increase in variance. However, this objection is incorrect for two reasons. First, it is impossible to know the size of the bias in advance for each possible case in which pre-exposure is a problem. Therefore, we cannot discard *a priori* that the bias is substantively relevant. Second, as discussed earlier, obtaining a bootstrap confidence interval comparable in size to the naive ATE would require a dataset approximately four times larger, which would enable a proper test of the difference in estimates. Hence, we should not interpret the overlap in confidence intervals as an indication that the estimates are the same. The overlap is a consequence of the different methods used to compute the standard errors using the same dataset, with the PEPS-ATE relying on bootstrap estimation and the naive ATE using a closed-form solution.

At any rate, as previously discussed, the ATE and the PEPS-ATE (BC) estimate different quantities. This application demonstrates that the proposed method offers a solution to account for pre-exposure and recover the AIE parameter, even when pre-exposure cannot be directly measured among the subjects in the experiment.

Conclusion

In this paper, I formalized the pre-exposure bias problem using the potential outcomes framework; decomposed the average treatment effect (ATE) into its pre-exposure and exposure causal effect components; identified and quantified the bias arising from neglecting pre-exposure; established sufficient identification conditions for identification of the pre-exposure causal parameters when pre-exposure is uncontrolled by the researchers; proposed a method to estimate the parameters when it is unfeaseable to measure pre-exposure directly from the experiment subjects; proposed a sampling procedure to apply the method; evaluated the finite sample performance of the bias-corrected estimator using Monte Carlo simulations; and illustrated the method with a real data application. The proposed solution can be implemented even after the experiment has been conducted and pre-exposure was not explicitly measured. It does not require repeated experiments or panel data, and can be conducted using two cross-sectional samples.

Although accounting for pre-exposure is considered an ideal procedure in information experiments (Gaines, Kuklinski, and Quirk 2006; Druckman and Leeper 2012; Hartman and Newman 2019), challenges related to causal inference that arise from pre-exposure have been largely overlooked, despite the concerns that have been raised starting at least two decades ago. The lack of attention given to pre-exposure in past research may be due to various reasons, such as the difficulty in dealing with the issue, the lack of fundational methodological work or formal results that explain pre-exposure bias, or the absence of a low-cost and practical methodological approach that addresses the problem. This paper bridges the gap between ideal and current research practices by tackling pre-exposure bias in information experiments when the goal is to evaluate the average information effect.

The method proposed has broad applicability beyond political science experiments. It can be used in observational studies that utilize machine learning classification methods to obtain binary interactive covariates. The PEPS estimator can correct for misclassification bias in such cases. The method is particularly important for information experiments (e.g. priming, framing, vignette, and cuing experiments) when addressing pressuring and salient political issues, where pre-exposure is more likely. It helps to account for the bias due to inability to directly measure pre-exposure among experimental subjects.

References

Aigner, Dennis J. 1973. "Regression with a binary independent variable subject to errors of observation." *Journal of Econometrics* 1 (1): 49–59.

Barber, Michael, and Jeremy C Pope. 2019. "Does party trump ideology? Disentangling party and ideology in America." *American Political Science Review,* 1–17.

Bates, Stephen, Trevor Hastie, and Robert Tibshirani. 2023. "Cross-Validation: What Does It Estimate and How Well Does It Do It?" *J. Am. Stat. Assoc.* (April): 1–12. ISSN: 0162-1459.

Beleites, Claudia, Ute Neugebauer, Thomas Bocklitz, Christoph Krafft, and Jürgen Popp. 2013. "Sample size planning for classification models." *Anal. Chim. Acta* 760 (January): 25–33. ISSN: 0003-2670.

Black, Dan, Seth Sanders, and Lowell Taylor. 2003. "Measurement of higher education in the census and current population survey." *Journal of the American Statistical Association* 98 (463): 545–554.

Bullock, John G. 2011. "Elite influence on public opinion in an informed electorate." *American Political Science Review* 105 (3): 496–515.

———. 2020. "Party Cues." In *The Oxford Handbook of Electoral Persuasion,* edited by Elizabeth Suhay, Bernard Grofman, and Alexander H. Trechsel. Oxford University Press.

Cheng, Dunlei, James D. Stamey, and Adam J. Branscum. 2009. "Bayesian approach to average power calculations for binary regression models with misclassified outcomes." *Stat. Med.* 28, no. 5 (February): 848–863. ISSN: 0277-6715.

Chong, Dennis, and James N. Druckman. 2010. "Dynamic Public Opinion: Communication Effects over Time." *American Political Science Review* 104, no. 4 (November): 663–680. ISSN: 1537-5943.

Clifford, Scott, Thomas J Leeper, and Carlisle Rainey. 2023. "Generalizing Survey Experiments Using Topic Sampling: An Application to Party Cues." *Political Behavior,* 1–24.

Cohen, Jacob, Patricia Cohen, Stephen G West, and Leona S Aiken. 2013. *Applied multiple regression/correlation analysis for the behavioral sciences.* Routledge.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and their Application.* Cambridge, England, UK: Cambridge University Press, October. ISBN: 978-0-52157471-6.

Druckman, James N., and Thomas J. Leeper. 2012. "Learning More from Political Communication Experiments: Pretreatment and Its Effects." *American Journal of Political Science* 56, no. 4 (October): 875–896. ISSN: 0092-5853.

Edwards, Brian J., Chad Haynes, Mark A. Levenstien, Stephen J. Finch, and Derek Gordon. 2005. "Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies." *BMC Genet.* 6, no. 1 (December): 1–12. ISSN: 1471-2156.

Figueroa, Rosa L, Qing Zeng-Treitler, Sasikiran Kandula, and Long H Ngo. 2012. "Predicting sample size required for classification performance." *BMC medical informatics and decision making* 12:1–10.

Fu, Wenjiang J., Raymond J. Carroll, and Suojin Wang. 2005. "Estimating misclassification error with small samples via bootstrap cross-validation." *Bioinformatics* 21, no. 9 (May): 1979–1986. ISSN: 1367-4803.

Gaines, Brian J, James H Kuklinski, and Paul J Quirk. 2006. "The logic of the survey experiment reexamined." *Political Analysis* 15 (1): 1–20.

Greene, William H. 2012. *Econometric analysis.* xxxix, 1188 p. Upper Saddle River, NJ: Pearson Prentice Hall. ISBN: 0131395386; 9780131395381.

Hartman, Todd K., and Benjamin J. Newman. 2019. "Accounting for Pre-Treatment Exposure in Panel Data: Re-Estimating the Effect of Mass Public Shootings." *British Journal of Political Science* 49, no. 4 (October): 1567–1576. ISSN: 0007-1234.

Hausman, Jerry A, Jason Abrevaya, and Fiona M Scott-Morton. 1998. "Misclassification of the dependent variable in a discrete-response setting." *Journal of econometrics* 87 (2): 239–269.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto. 2010. "Identification, inference and sensitivity analysis for causal mediation effects." *Statistical Science* 25 (1): 51–71.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.

Lachenbruch, Peter A. 1968. "On expected probabilities of misclassification in discriminant analysis, necessary sample size, and a relation with the multiple correlation coefficient." *Biometrics,* 823–834.

Linos, Katerina, and Kimberly Twist. 2018. "Diverse Pre-Treatment Effects in Survey Experiments." *Journal of Experimental Political Science* 5 (2): 148–158. ISSN: 2052-2630.

Ounpraseuth, Songthip, Shelly Y. Lensing, Horace J. Spencer, and Ralph L. Kodell. 2012. "Estimating misclassification error: a closer look at cross-validation based methods." *BMC Res. Notes* 5, no. 1 (December): 1–11. ISSN: 1756-0500.

Rahme, Elham, Lawrence Joseph, and Theresa W. Gyorkos. 2000. "Bayesian Sample Size Determination for Estimating Binomial Parameters from Data Subject to Misclassification." *J. R. Stat. Soc. Ser. C. Appl. Stat.* 49, no. 1 (March): 119–128. ISSN: 0035-9254.

Riley, Richard D., Joie Ensor, Kym I. E. Snell, Frank E. Harrell, Glen P. Martin, Johannes B. Reitsma, Karel G. M. Moons, Gary Collins, and Maarten van Smeden. 2020. "Calculating the sample size required for developing a clinical prediction model." *BMJ* 368 (March): m441. ISSN: 1756-1833.

Rosenbaum, Paul R, and Donald B Rubin. 1983. “The central role of the propensity score in observational studies for causal effects.” *Biometrika* 70 (1): 41–55.

Savoca, Elizabeth. 2000. “Measurement errors in binary regressors: an application to measuring the effects of specific psychiatric diseases on earnings.” *Health Services and Outcomes Research Methodology* 1:149–164.

Slothuus, Rune. 2016. “Assessing the Influence of Political Parties on Public Opinion: The Challenge from Pretreatment Effects.” *Political Communication* 33, no. 2 (April): 302–327. ISSN: 1058-4609.

Wasserman, Larry. 2013. *All of statistics: a concise course in statistical inference.* Springer Science & Business Media.

Yi, Grace Y., and Wenqing He. 2017. “Analysis of case-control data with interacting misclassified covariates.” *J. Stat. Distrib. App.* 4, no. 1 (December): 1–16. ISSN: 2195-5832.

Appendix

To proove proposition 2.1, denote π_1 the proportion of people pre-exposed, and assume D_2 (exposure) was randomized. Assume SUTVA and consistency (Imbens and Rubin 2015), which are required for any estimation of as defined in equation (1). Then,

$$
\begin{aligned}
\tau^{ATE} &= \mathbb{E}\left[Y_i(1)\right] - \mathbb{E}\left[Y_i(0)\right] \\
&= \mathbb{E}\left[Y_i(1) \mid D_2 = 1\right] - \mathbb{E}\left[Y_i(0) \mid D_2 = 0\right] && \text{(Randomization of } D_2\text{)} \\
&= \mathbb{E}\left[Y_i \mid D_2 = 1\right] - \mathbb{E}\left[Y_i \mid D_2 = 0\right] && \text{(Consistency assumption)} \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_i \mid D_1, D_2 = 1\right]\right] - \mathbb{E}\left[\mathbb{E}\left[Y_i \mid D_1, D_2 = 0\right]\right] && \text{(Law of interated expectations)} \\
&= \left(\pi_1\mathbb{E}\left[Y_i \mid D_1 = 1, D_2 = 1\right] + (1-\pi_1)\mathbb{E}\left[Y_i \mid D_1 = 0, D_2 = 1\right]\right) \\
&\quad - \left(\pi_1\mathbb{E}\left[Y_i \mid D_1 = 1, D_2 = 0\right] + (1-\pi_1)\mathbb{E}\left[Y_i \mid D_1 = 0, D_2 = 0\right]\right) \\
&= \pi_1\mathbb{E}\left[Y_i(1,1)\right] + (1-\pi_1)\mathbb{E}\left[Y_i(0,1)\right] \\
&\quad - \pi_1\mathbb{E}\left[Y_i(1,0)\right] + (1-\pi_1)\mathbb{E}\left[Y_i(0,0)\right] && \text{(Consistency assumption)} \\
&= (1-\pi_1)\left(\mathbb{E}\left[Y_i(0,1)\right] - \mathbb{E}\left[Y_i(0,0)\right]\right) + \pi_1\left(\mathbb{E}\left[Y_i(1,1)\right] - \mathbb{E}\left[Y_i(1,0)\right]\right) \\
&= (1-\pi_1)\tau^{AIE} + \pi_1\left(\mathbb{E}\left[Y_i(1,1)\right] - \mathbb{E}\left[Y_i(0,0)\right] + \mathbb{E}\left[Y_i(0,0)\right] - \mathbb{E}\left[Y_i(1,0)\right]\right) \\
&= (1-\pi_1)\tau^{AIE} + \pi_1\tau^{AIRE} - \pi_1\tau^{AIPE}
\end{aligned}
$$

□

The proof of theorem 3.1 is as follows. The general causal parameter capturing pre-exposure is:

$$\tau^{PE}(d_0, d_0') = \mathbb{E}\left[Y_i(d_0, 1) - Y_i(t)(d_0', 0)\right] \tag{16}$$

For the first expectation in 16, we have:

$$
\begin{aligned}
\mathbb{E}\left[Y_i(t)(d,1)\right] &= \mathbb{E}\left[Y_i(d,1) \mid D_{i1} = 1\right] && (Y_i(d_0, d_1) \perp\!\!\!\perp D_{i1}) \\
&= \mathbb{E}_X\left[\mathbb{E}\left[Y_i(d,1) \mid X_i, D_{i1} = 1\right]\right] && \text{(Law of interated expectations)} \\
&= \mathbb{E}_X\left[\mathbb{E}\left[Y_i(d,1) \mid X_i, D_{i0} = d, D_{i1} = 1\right]\right] && \text{(Assumption 3.1)} \\
&= \mathbb{E}_X\left[\mathbb{E}\left[Y_i \mid X_i, D_{i0} = d, D_{i1} = 1\right]\right] && \text{(Consistency assumption)} \\
&= \int_x \int_y y_i dF_{(y_i \mid X_i = x, D_{i0} = d, D_{i1} = 1)} dF_x
\end{aligned}
$$

To prove the asymptotic bias due to misclassification shown in expression (13), denote $\theta =$

$(\tau^{AIE}, \tau^{AIPE}, \gamma, \beta_x)$. From equation (12) we have

$$\widehat{\theta}_m = (\widehat{M}^T \widehat{M})^{-1} \widehat{M}^T y = (\widehat{M}^T \widehat{M})^{-1} \widehat{M}^T (\widehat{M}\theta + \epsilon - \gamma U + \tau^{AIRE}(U^T \boldsymbol{I} D_2))$$

Taking the probability limit,

$$\operatorname{plim} \widehat{\theta}_m = \theta + \operatorname{plim} \left(\frac{\widehat{M}^T \widehat{M}}{n} \right)^{-1} \operatorname{plim} \frac{\widehat{M}^T \epsilon}{n} - \operatorname{plim} \left[(\widehat{M}^T \widehat{M})^{-1} (\tau^{AIPE} \widehat{M}^T U + \gamma \widehat{M}^T (U^T \boldsymbol{I} D_2)) \right]$$

As $\operatorname{plim} \theta = \theta$, following Greene (2012, 66–7) and by assumption 3.1, we have $\operatorname{plim} \frac{\widehat{M}^T \epsilon}{n} = 0$ and the result follows for $Asy.Bias(\widehat{\theta}_m) = \operatorname{plim} \widehat{\theta}_m - \theta$. The proof that the remaining term is non-zero is a direct result of the proof for the bias-corrected PEPS estimator presented in the online suplement.