The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2139–2164
https://doi.org/10.1214/22-AOAS1712

LEVERAGING POPULATION OUTCOMES TO IMPROVE THE GENERALIZATION OF EXPERIMENTAL RESULTS: APPLICATION TO THE JTPA STUDY

BY MELODY HUANG[1,a], NAOKI EGAMI[2,b], ERIN HARTMAN[3,c] AND LUKE MIRATRIX[4,d]

[1]*Department of Statistics, University of California, Berkeley,* [a]*melodyyhuang@berkeley.edu*
[2]*Department of Political Science, Columbia University,* [b]*naoki.egami@columbia.edu*
[3]*Departments of Political Science and Statistics, University of California, Berkeley,* [c]*ekhartman@berkeley.edu*
[4]*Graduate School of Education, Harvard University,* [d]*lmiratrix@g.harvard.edu*

Generalizing causal estimates in randomized experiments to a broader target population is essential for guiding decisions by policymakers and practitioners in the social and biomedical sciences. While recent papers have developed various weighting estimators for the population average treatment effect (PATE), many of these methods result in large variance because the experimental sample often differs substantially from the target population and estimated sampling weights are extreme. We investigate this practical problem motivated by an evaluation study of the Job Training Partnership Act (JTPA), where we examine how well we can generalize the causal effect of job training programs beyond a specific population of economically disadvantaged adults and youths. In particular, we propose post-residualized weighting in which we use the outcome measured in the observational population data to build a flexible predictive model (e.g., with machine learning) and residualize the outcome in the experimental data before using conventional weighting methods. We show that the proposed PATE estimator is consistent under the same assumptions required for existing weighting methods, importantly without assuming the correct specification of the predictive model. We demonstrate the efficiency gains from this approach through our JTPA application: we find a reduction of between 5% and 25% in variance.

1. Introduction. The Job Training Partnership Act (JTPA) was introduced by the U.S. Congress in 1982 to help provide employment and training programs to economically disadvantaged adults and youths. To assess its effectiveness, the national JTPA study evaluated the impact of the program across a diverse set of 16 experimental sites between 1987 and 1989. Eligible individuals assigned to treatment were given access to the JTPA services, while those assigned to control were told that the services were not available. Eighteen months later, researchers checked on whether these study participants were employed and measured their recent earnings (Bloom et al. (1993)). The hope is that those offered the program would be more often employed and would generally be earning higher wages.

Each site can be considered a stand-alone randomized trial. Each site has a different collection of individuals from the other sites. If a policymaker had only run their experiment in one specific population, how representative would their results have been for the other populations? This question is the essence of a current and serious critique of large-scale randomized evaluations: does a rigorous and robust finding regarding a program evaluated in a specific population actually shed light on wider questions of a program's effectiveness for a "real-world" population? Originally, the "credibility revolution" elevated the role of randomized, controlled trials (RCTs), generally praised for their strong internal validity (Baldassarri and Abascal (2017), Banerjee and Duflo (2009), Falk and Heckman (2009)). RCTs are attractive

Received October 2021; revised October 2022.
Key words and phrases. Generalizability, external validity, randomized experiments, causal inference.

in that they allow researchers to draw causal inferences about treatment effects with only minimal assumptions, but only for the experimental sample. And perhaps this last clause is too great a cost; perhaps the emphasis on causality has led researchers to overly narrow the scope of their inquiry (Deaton and Cartwright (2018), Huber (2013)). Especially with a policy finding, if one cannot generalize, what should one make of a found result? Concerns about generalizability span the social and biomedical sciences and are related to discussions about participant recruitment in pragmatic study designs (Ford and Norrie (2016)).

This critique has inspired a robust literature on methods for how to generalize experimental result to broader populations of interest. In our case, for example, one could imagine extending the results found for a specific population in one site to populations living in the other sites, adjusting the impact estimate to account for differences in populations served. The generalizability literature has provided clear outlines for the necessary assumptions for such generalization, providing tools to identify the population average treatment effect (PATE), that is, the effect of the experimental treatment in a clearly defined target population that differs from the experimental sample (Bareinboim and Pearl (2016), Cole and Stuart (2010), Egami and Hartman (2022)). In practice, the most common approaches model the experimental sample inclusion probability, with the PATE then estimated using weighting estimators (Buchanan et al. (2018), Hartman et al. (2015), Stuart et al. (2011), Tipton (2013)). Alternative estimators focus on modeling treatment effect heterogeneity (Kern et al. (2016), Nguyen et al. (2017)) or doubly robust estimation (Dahabreh et al. (2019)).

Generalizing, however, can be prohibitively costly. In practice, weighted estimators are often far more imprecise than unweighted estimators, especially when the experimental sample differs substantially from the target population. This makes it difficult for policymakers and practitioners to draw conclusions about the impact of treatment in the target population to guide their policy recommendations. Indeed, Miratrix et al. (2018) empirically found that weighted estimators often increase the mean squared error for the PATE, compared to a biased estimator that ignores sampling weights, due to paying for a smaller bias with much larger standard errors. More generally, considering the bias-variance tradeoff, the cost the large precision loss associated with conventional weighting methods makes it unclear if it is "worth weighting" and questions the applicability of these weighting methods that researchers are advocating for.

This provides a quandary: the more the target population differs from the sample, the greater the cost of generalizing, due to more extreme weights, but the greater the need to generalize to keep the findings of the original experiment relevant. In this work we seek to mitigate this tradeoff by exploiting a valuable resource commonly left on the table: the outcome data measured in the population. In particular, we aim to incorporate this additional observational population data to reduce the noise from generalizing an experimental result. Population data often have larger sample sizes and, therefore, provide an opportunity to model complex covariate-outcome relationships with more flexible modeling approaches. It is this opportunity—to incorporate large population data sets that contain outcome data to improve precision—that serves as the foundation of our method.

The multisite design of our JPTA experiment serves as an ideal test bed for our method. We generalize the results of each site individually to a target population defined by the units in the other 15 sites, allowing us to benchmark our estimates against the experimentally identified causal estimate of the excluded sites. We can then evaluate any precision gains, as compared to other generalization approaches as well as to no adjustment. We can also, for each site in turn, assess whether one should generalize, based on a diagnostic test. Ultimately, using this within study comparison approach (LaLonde (1986)), we find between a 5% to 25% reduction in variance from exploiting population data and outcomes for those sites where we determine that our methods are applicable.

Our method is post-residualized weighting, where we leverage outcome data measured in the population to improve precision in estimation of the PATE. We begin by constructing a predictive model of the outcome using the population data. We then use this to residualize the experimental outcome data, and these residuals replace the experimental outcome in the standard inverse probability weighting estimators used for generalization. Identification of the PATE proceeds under the same assumptions required for existing inverse probability weighting methods, namely, that the sampling weights are correctly specified. We show that this estimator is consistent, regardless of the residualizing model constructed in the population data. Therefore, we can safely use machine learning methods to build a predictive model. We then establish under what conditions the proposed post-residualized weighting estimator is more efficient than existing methods.

We also extend our estimator to the weighted least squares framework, which has three advantages: (1) it incorporates the well-known benefits of stabilized weighting estimators (i.e., Hàjek estimators), (2) it allows for additional precision gains from prognostic variables measured only within the experiment, and (3) it addresses concerns about scaling differences between the outcomes measured in the experiment and the population data. Importantly, we also provide a diagnostic that allows researchers to assess when the post-residualized weighting method is likely to result in efficiency gains.

As far as we know, using covariates and outcome data in this manner has not been investigated. While inverse probability weighting methods do leverage population data about pre-treatment covariates when modeling the sampling weights, use of outcome data has primarily been limited to use in placebo tests (Cole and Stuart (2010), Hartman et al. (2015)). Recently, the data fusion literature proposed using experimental data to help aid the estimation of causal effects in observational studies (e.g., see Athey, Chetty and Imbens (2020), Athey et al. (2019), Kallus and Mao (2020)) which bears some similarity to our problem.

We proceed by further introducing our empirical application. We then introduce notation and existing methods for estimating the population average treatment effect from experimental data in Section 2. In Section 3 we introduce post-residualized weighting, prove its statistical properties, and introduce a diagnostic to assess whether researchers should expect efficiency gains in their applications. We consider both a weighted estimator (a.k.a., a Hàjek estimator) and a weighted least squares estimator. We extend our results to a case in which we include the predicted outcome as a covariate in Section 4. Finally, we provide simulation evidence supporting the performance of post-residualized weighting estimators and diagnostic tools in Section 5 and apply them to the Job Training Partnership Act in Section 6.

1.1. *Background and data.* The Job Training Partnership Act (JTPA) was a large study with a 2:1 treatment to control ratio. A variety of outcomes were measured with a follow-up survey 18 months after assignment (Bloom et al. (1993)). We use the 16 experimental sites from the national JTPA study as the basis for our analysis. While the original study focused on four target groups, adult women and men (categorized formally as ages 22 and older) and female and male out-of-school youths (ages 16–21), we focus our analysis on adult women, the largest target group within the JTPA study.[1] We consider two different outcomes: employment status (binary outcome) and total earnings (zero-inflated, continuous outcome). Across the 16 sites, the average effect on earnings was \$1240 and employment was 1.63%, but point estimates across sites ranged from −\$5210 in Butte, MT, to \$3030 in Providence, RI, for earnings and −7% in Butte, MT, and Marion, OH, to 7% in Heartland, FL, and Providence, RI. Had a policymaker only run their experiment in Providence, RI,

[1]The estimated impacts of JTPA for the other target groups were not found to be statistically significant in the original study.

they may have concluded that the treatment was effective but not so in Butte, MT. Weighted estimators can adjust for demographic differences across sites, but many of the sites, such as Butte, MT, contain few units, emphasizing the need for precise estimators when generalizing results to other populations.

Using a within study comparison approach, we generalize the results of each site individually to a target population defined by the units in the other 15 sites, allowing us to benchmark our estimator against the experimentally identified causal estimate of the excluded sites and evaluate precision gains from post-residualized weighting. A summary of the JTPA experimental setup is provided in Supplementary Material Table A5 (Huang et al. (2023)).

2. Existing estimators for generalization.

2.1. *Setup*. We begin by defining the target population as an infinite super-population $\mathcal{P}$ with probability distribution F and probability density dF for which we wish to infer the effectiveness of treatment. Following Buchanan et al. (2018), suppose we observe n units as the "experimental sample," but, as with most experiments in practice, the selection into the experiment from the target population is biased. Let $\mathcal{S}$ represent the random set of n indices for the units in the experimental sample.

Units in our experimental sample are treated, or not, with treatment indicator $T_i = 1$ for units assigned to treatment and $T_i = 0$ for control. Using the potential outcomes framework (Neyman (1923), Rubin (1974)), we define $Y_i(t)$ to be the potential outcome of unit i that would realize if unit i receives treatment $T_i = t$, where $t \in \{0, 1\}$. Our primary causal quantity of interest is the population average treatment effect (PATE), which is formally defined as

$$\tau := \mathbb{E}_F\{Y_i(1) - Y_i(0)\}, \tag{1}$$

where the expectation is taken over the target population distribution F. This is in contrast to the sample average treatment effect (SATE):

$$\tau_{\mathcal{S}} := \mathbb{E}_{\tilde{F}}\{Y_i(1) - Y_i(0)\},$$

where the expectation is taken over the experimental sample distribution $\tilde{F}$, described below.

For each unit in the experiment, only one of the potential outcome variables can be observed, and the realized outcome variable for unit i is denoted by $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$. We also observe pre-treatment covariates $\mathbf{X}_i$ for units in the experiment. We use $\tilde{F}$ to represent the sampling distribution for the experimental sample, that is, $\{Y_i(1), Y_i(0), T_i, \mathbf{X}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \tilde{F}$ with density $d\tilde{F}$. Because we consider settings where the selection into the experiment from the target population $\mathcal{P}$ is biased, $F \neq \tilde{F}$.

We assume that treatment assignment is randomized within the experiment.

ASSUMPTION 1 (Randomization within Experiment).

$$d\tilde{F}(Y_i(1), Y_i(0), T_i, \mathbf{X}_i) = d\tilde{F}(Y_i(1), Y_i(0), \mathbf{X}_i) \cdot d\tilde{F}(T_i). \tag{2}$$

Under this assumption the SATE can be estimated without bias[2] using a difference-in-means estimator:

$$\widehat{\tau}_{\mathcal{S}} = \frac{1}{\sum_{i\in\mathcal{S}} T_i} \sum_{i\in\mathcal{S}} T_i Y_i - \frac{1}{\sum_{i\in\mathcal{S}}(1 - T_i)} \sum_{i\in\mathcal{S}} (1 - T_i) Y_i. \tag{3}$$

[2]Assuming that $\sum_i T_i$ is fixed; otherwise, the bias is expected to be small.

The SATE is important for evaluating the effectiveness of treatment. However, researchers often want to know to what extent the internally valid findings of an experiment are externally valid to the target population (Cole and Stuart (2010), Egami and Hartman (2022), Miratrix et al. (2018)). When the experimental sample is randomly drawn from the target population, $F = \tilde{F}$, and $\widehat{\tau}_{\mathcal{S}}$ can be used as an unbiased estimator for τ. However, in most settings experimental units are not randomly drawn from the target population with equal probability.

To estimate the PATE, we also assume we observe an i.i.d. sample of N units from the target super-population $\mathcal{P}$ as the "population data" which is separate from the experimental sample. This design is most common in the social sciences and is called the nonnested design in that the experimental sample is not a subset of the population data (Colnet et al. (2020)).[3] Typically, the size of the population data is much larger than the experimental data, that is, $N \gg n$. In the conventional setup, researchers only observe pre-treatment covariates $\mathbf{X}_i$ for each unit i in the population data. In the next subsection, we review assumptions and estimators for the PATE under this conventional setup. In Section 3 we then consider our setting in which researchers also observe an outcome measure in addition to pre-treatment covariates in the population data. Importantly, because the treatment is not randomized in the population data, we cannot identify the PATE just using the population data without further assumption.

2.2. *Assumptions*. We make the standard assumptions of no interference and that treatments are identically administered across all units (i.e., SUTVA, defined in Rubin (1980)). In order to identify the PATE using experimental data, we require additional assumptions about the sampling of the experimental units. First, we assume that, conditional on a set of pre-treatment covariates $\mathbf{X}_i$, the sample selection mechanism is ignorable. Below we present this more formally.

ASSUMPTION 2 (Ignorability of Sampling and Potential Outcomes).

$$dF(Y_i(1), Y_i(0) \mid \mathbf{X}_i = \mathbf{x}) = d\tilde{F}(Y_i(1), Y_i(0) \mid \mathbf{X}_i = \mathbf{x}). \tag{4}$$

Assumption 2 states that, conditional on $\mathbf{X}_i$, the distribution of the potential outcomes $\{Y_i(1), Y_i(0)\}$ is the same across the experimental sample and the target population (Kern et al. (2016), Pearl and Bareinboim (2014), Stuart et al. (2011)).[4] We also assume that, for any pre-treatment covariate profile $\mathbf{X}_i = \mathbf{x}$ we might see in the population, we have a nonzero chance of seeing it in the sample as well (Westreich and Cole (2010)):

ASSUMPTION 3 (Positivity). For all $\mathbf{x}$, we have

$$dF(\mathbf{X}_i = \mathbf{x}) > 0 \quad \Rightarrow \quad d\tilde{F}(\mathbf{X}_i = \mathbf{x}) > 0. \tag{5}$$

2.3. *Estimation of PATE*. There is a robust and growing literature on methods for estimating the PATE. The most common approach is the inverse probability weighting estimator

[3]While we focus on the nonnested design in this paper, the same proposed approach is useful for the nested design where the experimental sample is a subset of the population data. The main difference arises in the analytical expressions of the efficiency gain from our proposed approach.

[4]For identification of the PATE, a weaker assumption of conditional ignorability of sampling and treatment effect heterogeneity may be invoked instead. However, our variance derivations rely on the conditional ignorability of sampling and potential outcomes.

(IPW) (Cole and Stuart (2010)). The IPW estimator relies on sampling weights, usually defined as an inverse of the probability of being sampled into the experiment. In our case, given the infinite superpopulation defined by F, this translates to, for each unit i,

$$w_i \propto \frac{1}{\pi(\mathbf{X}_i)},$$

with $\pi(\mathbf{X}_i)$ the relative density of

$$\pi(\mathbf{X}_i) = \frac{d\tilde{F}(\mathbf{X}_i)}{dF(\mathbf{X}_i)}. \tag{6}$$

Weights are typically estimated using a binary outcome model, such as logistic regression, by exploiting the fact that weights are proportional to the relative probability of being in the observed population data to the probability of being in the experimental sample, conditional on being in either set:

$$w_i \propto \frac{\Pr(S_i = 0 \mid \mathbf{X}_i)}{\Pr(S_i = 1 \mid \mathbf{X}_i)},$$

where S_i takes on a value of 1, if the unit belongs to the experimental sample, and 0 if the unit belongs to the observed population data.

Researchers can estimate $\Pr(S_i = s \mid \mathbf{X}_i)$ using a binary outcome model, regressing S_i on $\mathbf{X}_i$ using the stacked dataset of both the experimental and population data (Buchanan et al. (2018), Egami and Hartman (2021), O'Muircheartaigh and Hedges (2014), Stuart et al. (2011)). Alternatively, researchers can use balancing methods, such as entropy balancing, which estimates weights such that weighted moments (e.g., means of each pre-treatment covariate $\mathbf{X}_i$) of the experimental data equal the corresponding moments of the observed population data (Deville and Särndal (1992), Hainmueller (2012), Hartman et al. (2015)).

Once researchers have estimated the sampling weights, the PATE can be estimated using a weighted estimator, also known as the Hàjek estimator,

$$\hat{\tau}_W := \frac{\sum_{i\in\mathcal{S}} \hat{w}_i T_i Y_i}{\sum_{i\in\mathcal{S}} \hat{w}_i T_i} - \frac{\sum_{i\in\mathcal{S}} \hat{w}_i (1-T_i) Y_i}{\sum_{i\in\mathcal{S}} \hat{w}_i (1-T_i)}. \tag{7}$$

As with estimation of the SATE, researchers can also include covariate adjustment to increase efficiency. This approach is popular because, while the estimation of the weights requires covariates to be measured across both the population and the experimental data, covariate adjustment can leverage covariates that are only measured in the experimental data (Stuart and Rhodes (2017)).

The weighted least squares estimator $\hat{\tau}_{\text{wLS}}$ for the PATE can be computed via a weighted regression of the outcome on an intercept, the treatment indicator and pre-treatment covariates using the estimated weights. Formally,

$$(\hat{\tau}_{\text{wLS}}, \hat{\alpha}, \hat{\gamma}) = \underset{\tau,\alpha,\gamma}{\operatorname{argmin}} \frac{1}{n} \sum_{i\in\mathcal{S}} \hat{w}_i \left(Y_i - (\tau T_i + \alpha + \widetilde{\mathbf{X}}_i^\top \gamma)\right)^2, \tag{8}$$

where $\widetilde{\mathbf{X}}_i$ are experimental pre-treatment covariates included in the covariate adjustment. Covariates $\widetilde{\mathbf{X}}_i$ can differ from the $\mathbf{X}_i$ required for Assumptions 2–3. The weighted estimator (equation (7)) is a special case of this weighted least squares estimator (equation (8)) because it is numerically equivalent to the estimated coefficient of the treatment indicator when no covariate is included, that is, $\widetilde{\mathbf{X}}_i = \varnothing$. Because the weighted estimator is a special case of the weighted least squares estimator, we focus on the weighted least squares estimator in this paper, but use the simpler weighted estimator to illustrate intuitions when appropriate. Under Assumptions 1–3 and the consistent estimation of the sampling weights, the weighted

estimator $\hat{\tau}_W$ and the weighted least squares estimator $\hat{\tau}_{\text{wLS}}$ are both consistent for the PATE, regardless of what covariates $\widetilde{\mathbf{X}}$ we include as covariate adjustment (Buchanan et al. (2018), Dahabreh et al. (2019)).

In practice, weighted estimators can suffer from large variance due to extreme weights, which in this case depends on how much the individual unit-level probabilities of inclusion in the experimental sample varies relative to their average probability of inclusion. This problem has been highlighted in the observational causal inference literature with respect to inverse propensity score weighted estimators, in which large imbalances between treatment and control groups can result in extreme weights (Kang and Schafer (2007), Stuart (2010)). This issue is often exacerbated in the generalization setting, where imbalances between a convenience experimental sample and target population can be relatively large. As a result, losses in precision from weighting can be challenging to overcome when generalizing from the SATE to the PATE (Miratrix et al. (2018)).

3. Post-residualized Weighting. Existing methods, such as the weighted estimator and weighted least squares estimator described above, require pre-treatment covariate data, measured in both the experimental sample and target population, for estimating the sampling weights. However, researchers often have access to an outcome variable in the observational population data as well. Our proposed method, *post-residualized weighting*, aims to improve precision in estimation of the PATE by leveraging this outcome variable measured in the observational population data; see Figure 1 for a visualization of the difference in settings from conventional methods.

In addition to our JTPA application, which inspires our method, we next describe two canonical social science examples below that motivate the data settings that underpin our method. We return to these examples, in addition to the JTPA application, for conceptual clarity. We describe our benchmark analysis of the JTPA data in Section 6.

EXAMPLE (Get-Out-the-Vote (GOTV) experiments). Political scientists have conducted a number of field experiments to evaluate the impact of canvassing efforts, including door-to-door, phone, and mail, on voter turnout. Such GOTV experiments typically rely on administrative data to measure the outcome, namely, voter turnout data from the secretary of state. These experiments are often conducted in a small geographic region (e.g., New Haven, Connecticut, in Gerber and Green (2000)), but scholars are often interested in generalizing

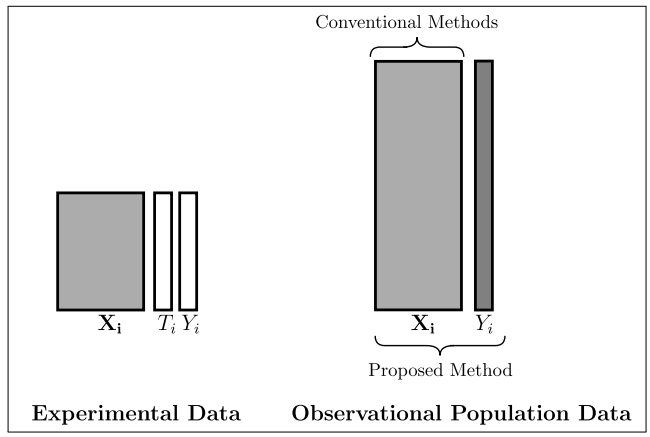

FIG. 1. *Data requirements. Conventional estimation methods only use the covariate data* $\mathbf{X}_i$ *(in light gray). Our proposed approach leverages the outcome data in addition to the covariate data at the population level (as highlighted in dark gray).*

the effect to broader populations, such as for a statewide election. Importantly, when considering generalization, the outcome variable of voter turnout is available not only for the experimental data but also for the broader target population of interest. In our framework we could use this information about voter turnout measured in the observational population data to improve precision in the estimation of the PATE.

EXAMPLE (Education experiments). Education research relies on experiments to evaluate the performance of classroom interventions, such as the impact of smaller class size on curriculum-based and standardized tests (e.g., Word et al. (1990)). These experiments are often done in partnership with school systems. For example, the Tennessee STAR experiment was conducted in classrooms across Tennessee. However, researchers are interested in the broader impact of such interventions. For example, a researcher may ask what the long term impact of small class sizes in primary school is on standardized test scores, such as the SAT, for all public schools in the United States. To estimate the PATE, existing methods use demographic variables from a random sample of public school students to construct sampling weights. In our framework we can additionally use SAT scores measured for such a sample to improve estimation accuracy.

REMARK. We emphasize that the outcome variable available in the population data can be either the potential outcomes under treatment $Y_i(1)$, the potential outcomes under control $Y_i(0)$, or their mix. Indeed, researchers do not need to know the treatment condition of units in the target population. This is because consistency of our proposed approach does not depend on the correct specification of a predictive model we will build with the outcome variable available in the population data (see Theorem 1 below). More generally, the outcome variable available in the population data can even be a proxy of the outcome variable in the experimental data (i.e., not equal to either the potential outcomes under treatment or control), and we consider this case in Section 4.

3.1. *Post-residualized weighted estimators*. The key idea of our proposed post-residualized weighting approach is that we estimate a predictive model with the outcome measured in the population data and then use this estimated predictive model to *residualize* outcomes in the experimental data, before using conventional weighting estimators to estimate the PATE. For example, in our JTPA application we predict earnings or employment across the target sites (i.e., the "population"), which we use to residualize the outcomes in the experimental site.

In total, post-residualized weighting has four steps. The first step is to estimate sampling weights w_i which is the same as the conventional weighting approach. In the second step, we fit a flexible model in the population data to predict the outcome variable Y_i using pretreatment $\mathbf{X}_i$. We refer to this predictive model fit in the population data as a *residualizing model* and formally denote it as $g(\mathbf{X}_i)$: $\mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is the support of $\mathbf{X}_i$. In the third step, we use the estimated residualizing model to predict outcomes $\widehat{Y}_i$ in the experimental data which is separate from the population data used to estimate the residualizing model. In the fourth and final step, we apply the weighted least squares estimator (equation (8)), using the residuals from this prediction (denoted by $\hat{e}_i = Y_i - \hat{Y}_i$) as outcomes (instead of Y_i used in the conventional weighted least squares estimator).

We summarize our proposed approach in Table 1. In the following section, we directly extend the weighted estimator and the weighted least squares estimator discussed in Section 2.

DEFINITION 1 (Post-residualized Weighted Least Squares Estimator). Given a residualizing model estimated as $\hat{g}(\cdot)$, the post-residualized weighted least squares estimator $\hat{\tau}^{\text{res}}_{\text{wLS}}$ for

TABLE 1
Steps of post-residualized weighting

Post-residualized weighting for the PATE estimation:	
Step 1:	Estimate sampling weights, w_i, for units in the experimental sample.
Step 2:	Choose a residualizing model $g(\mathbf{X}_i)$: $\mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is the support of $\mathbf{X}_i$. Using the population data, estimate $\hat{g}(\mathbf{X}_i)$ that predict the population outcomes using pre-treatment covariates $\mathbf{X}_i$.
Step 3:	Predict $\hat{Y}_i = \hat{g}(\mathbf{X}_i)$ for each unit in the experimental data, and compute residual $\hat{e}_i = Y_i - \hat{Y}_i$ for units in the experimental sample.
Step 4:	Estimate the PATE using residuals $\hat{e}_i$ and estimated sampling weights $\hat{w}_i$.
	No covariate adjustment within the experimental data ↳ See post-residualized weighted estimator $\hat{\tau}^{\text{res}}_W$ in equation (10).
	With covariate adjustment within the experimental data ↳ See post-residualized weighted least squares estimator $\hat{\tau}^{\text{res}}_{\text{wLS}}$ (Definition 1).

the PATE is defined as

$$(\hat{\tau}^{\text{res}}_{\text{wLS}}, \hat{\alpha}^{\text{res}}, \hat{\gamma}^{\text{res}}) = \underset{\tau,\alpha^{\text{res}},\gamma^{\text{res}}}{\operatorname{argmin}} \frac{1}{n} \sum_{i \in \mathcal{S}} \hat{w}_i \left(\hat{e}_i - (\tau T_i + \alpha^{\text{res}} + \widetilde{\mathbf{X}}_i^\top \gamma^{\text{res}}) \right)^2, \tag{9}$$

where $\hat{e}_i = Y_i - \hat{g}(\mathbf{X}_i)$ and $\widetilde{\mathbf{X}}_i$ are experimental pre-treatment covariates included in the covariate adjustment. We allow $\widetilde{\mathbf{X}}_i$ to differ from the $\mathbf{X}_i$ used to calculate $\hat{g}(\mathbf{X}_i)$.

In practice, the post-residualized weighted least squares estimator can be estimated by running a weighted regression, where the estimated residualized values $\hat{e}_i$ is regressed on an intercept, the treatment indicator T_i and covariates $\widetilde{\mathbf{X}}_i$, and using the sampling weights $\hat{w}_i$ as the weights. The coefficient of the treatment indicator is the post-residualized weighted least squares estimate for the PATE.

In a special case where no pre-treatment covariates are included, the post-residualized weighted least squares estimator is equivalent to the following post-residualized weighted estimator:

$$\hat{\tau}^{\text{res}}_W := \frac{\sum_{i \in \mathcal{S}} \hat{w}_i T_i \hat{e}_i}{\sum_{i \in \mathcal{S}} \hat{w}_i T_i} - \frac{\sum_{i \in \mathcal{S}} \hat{w}_i (1 - T_i) \hat{e}_i}{\sum_{i \in \mathcal{S}} \hat{w}_i (1 - T_i)}. \tag{10}$$

We summarize several key aspects of the post-residualized weighted least squares estimator here and formally discuss each point in the subsequent sections. First, the identification of the PATE is obtained under the same assumptions required for existing weighted and weighted least squares estimators, and we do not make any additional assumptions (Section 3.2). Most importantly, our proposed estimators are consistent for the PATE, regardless of the choice of the residualizing model. That is, we do not require the correct specification of the residualizing model $g(\mathbf{X}_i)$ to guarantee consistency of the proposed estimators. Therefore, akin with Rosenbaum (2002) and Sales, Hansen and Rowan (2018), the residualizing model $g(\mathbf{X}_i)$ can be seen as an "algorithmic model" in that the goal is to predict outcomes, rather than substantively explain an underlying probabilistic process.

Second, the proposed post-residualized weighted least squares estimator, $\hat{\tau}^{\text{res}}_{\text{wLS}}$, can achieve significant improvements in precision over the traditional weighted least squares estimator (equation (8)) when the residualizing model can predict outcomes in the experiment well (Section 3.3). We will show in Section 3.3 that, while we maintain consistency regardless, how much efficiency gain we achieve depends on the predictive performance of the fitted

residualizing model $\widehat{g}(\mathbf{X}_i)$. As such, researchers should, when possible, use not only simple models, such as ordinary least squares, but also more flexible machine learning models, such as random forests or other ensemble learning methods (Breiman (2001), van der Laan, Polley and Hubbard (2007)) as the residualizing models to improve precision of the PATE estimation.

Finally, we derive a diagnostic measure that researchers can use to determine whether residualizing will likely lead to precision gains when estimating the PATE (Section 3.4). As emphasized in the second point above, when the residualizing model can predict outcomes in the experiment well, we can expect efficiency gains. However, when the residualizing model fails to predict outcome measures in the experimental data, it is possible for post-residualizing to increase uncertainty of the PATE estimation. Our diagnostic measure helps researchers to estimate the expected efficiency gain, thereby deciding whether residualizing is beneficial in their applications.

REMARK. Our proposed post-residualized weighted least squares estimator is closely connected to the augmented inverse probability weighted estimators (AIPW) (Robins, Rotnitzky and Zhao (1994)) developed for the PATE (Dahabreh et al. (2019)) in that both estimators combine weighting and outcome-modeling. The process of estimating weights for both the post-residualized weighting estimators and AIPW is the same. However, the key difference between two approaches is that the AIPW estimates the outcome model using only the experimental data, thereby not exploiting the outcome variable available in the population data. In contrast, our post-residualized weighting estimator explicitly uses the outcome information available in the population data to estimate the residualizing model and improve precision. Furthermore, post-residualized weighting does not attempt to model both the treatment and control outcomes separately and, therefore, does not have the double robustness that the AIPW has.

REMARK. Compared to the simpler post-residualized weighted estimator (equation (10)), there are two advantages to a more general, post-residualized weighted least squares estimator (equation (9)). First, it can leverage precision gains from pre-treatment covariates that are measured in the experimental data but not in the population data. That is, $\widetilde{\mathbf{X}}_i$ can include more covariates than $\mathbf{X}_i$. Second, $\hat{\tau}^{\text{res}}_{\text{wLS}}$ provides additional robustness over the post-residualized weighted estimator $\hat{\tau}^{\text{res}}_{W}$. More specifically, without further covariate adjustment, residualizing can be sensitive to differences between the population and experimental units in the covariate-outcome relationships. For example, considering JTPA, if earnings and employment depend heavily on local economic conditions, and thus the covariate relationships differ across sites, then residualizing may not provide efficiency gains. As illustrated in Section 3.3, when this difference is large, residualizing can result in efficiency loss. However, by performing covariate adjustment on the residualized outcomes in the experimental data, we have an opportunity to correct for the difference in the covariate-outcome relationships between the experimental data and the population data. In other words, the post-residualized weighted least squares estimator, $\hat{\tau}^{\text{res}}_{\text{wLS}}$, gives researchers two opportunities to combat the precision loss of weighting: once from using the population data in the residualizing process and a second from adjusting for covariates in the experimental data.

3.2. *Consistency.* In this section we show that the post-residualized weighted least squares estimator is a consistent estimator of the PATE, regardless of the choice of the residualizing model $g(\mathbf{X}_i)$ and pre-treatment covariates $\widetilde{\mathbf{X}}_i$ that researchers adjust for in the weighted least squares estimator. This emphasizes the point that $g(\mathbf{X}_i)$ need not be a correct specification of the underlying data-generating process but merely a function that predicts outcomes measured in the population.

THEOREM 1 (Consistency of Post-residualized Weighted Least Squares Estimators). *Assume that sampling weights $\hat{w}_i$ are consistently estimated and Assumptions 1–3 hold with pre-treatment covariates $\mathbf{X}_i$. Then the post-residualized weighted least squares estimator that adjusts for pre-treatment covariates $\widetilde{\mathbf{X}}_i$ (equation (9)) is a consistent estimator*

$$\hat{\tau}^{\text{res}}_{\text{wLS}} \xrightarrow{p} \tau$$

with any residualizing model $g(\mathbf{X}_i)$ and any pre-treatment covariates $\widetilde{\mathbf{X}}_i$. The post-residualized weighted estimator (equation (10)) is also consistent, as it is a special case when no covariate is included.

The proof of Theorem 1 can be found in Supplementary Material Section 1. This property allows for a large degree of flexibility in building the residualizing model, since consistency is guaranteed *regardless* of model specification or performance of $g(\mathbf{X}_i)$. We obtain consistency, even for a misspecified residualizing model $g(\mathbf{X}_i)$, because the predicted experimental outcome $\hat{Y}_i = \hat{g}(\mathbf{X}_i)$ is only a function of the pre-treatment covariates $\mathbf{X}_i$, and thus, with randomized treatments (Assumption 1) its distribution is the same across treatment and control units on average for any sample size. As such, residualizing preserves the consistency of the original weighted estimator without requiring any additional assumptions.

A potential concern with covariate adjustment is that performing covariate adjustment within the experimental data can result in worsened asymptotic precision and invalid measures of uncertainty (Freedman (2008)). An alternative approach is to include interaction terms between the treatment indicator and covariates (Lin (2013)). Regardless, because the proposed post-residualized weighted least squares estimator is an extension of a weighted least squares estimator, we can compute valid standard errors with the standard Huber–White sandwich estimator.

While consistency is guaranteed, efficiency gains from residualizing *do* depend on the ability of the residualizing model to predict outcome measures in the experimental data. Theorem 1 allows for researchers to leverage complex, "black box" approaches (such as ensemble methods) to maximize the predictive accuracy, as interpretability of the residualizing model is secondary to being able to fit the data well. In the next section, we will formalize the criteria for variance reduction from residualizing.

3.3. *Efficiency gains*. The post-residualized weighted estimator allows researchers to include information from the observational population data about the relationship between the pre-treatment covariates and the population outcomes into the estimation process. Whether or not we obtain precision gains, and the magnitude of these precision gains, will depend on the nature of the residualizing model. In general, the better researchers are able to explain the outcomes measured in the experiment using the residualizing model, the greater the efficiency gains. For example, as shown in Section 6, we see greater gains from post-residualized weighting for earnings, where our predictive model performs better, than we do for employment, which is more difficult to predict with the auxiliary covariates.

To make these gains more explicit, we first define the *weighted variance* and *weighted covariance* as follows:

$$\text{Var}_w(A_i) = \int \frac{1}{\pi(\mathbf{X}_i)^2} \cdot (A_i - \bar{A})^2 \, d\tilde{F}(\mathbf{X}_i, A_i), \tag{11}$$

$$\text{Cov}_w(A_i, B_i) = \int \frac{1}{\pi(\mathbf{X}_i)^2} \cdot (A_i - \bar{A})(B_i - \bar{B}) \, d\tilde{F}(\mathbf{X}_i, A_i, B_i), \tag{12}$$

where $\bar{A} = \mathbb{E}_F(A_i)$ and $\bar{B} = \mathbb{E}_F(B_i)$.

For simplicity, we first describe the efficiency gain for the post-residualized weighted estimator (equation (10)).

THEOREM 2 (Efficiency Gain for Post-residualized Weighted Estimators). *The difference between the asymptotic variance of* $\hat{\tau}_W^{\text{res}}$ *and that of* $\hat{\tau}_W$ *is*

$$\begin{aligned}&\text{AVar}_{\tilde{F}}(\hat{\tau}_W) - \text{AVar}_{\tilde{F}}(\hat{\tau}_W^{\text{res}})\\ &\quad= -\frac{1}{p(1-p)}\text{Var}_w(\hat{Y}_i) + \frac{2}{p}\text{Cov}_w(Y_i(1), \hat{Y}_i) + \frac{2}{1-p}\text{Cov}_w(Y_i(0), \hat{Y}_i),\end{aligned} \tag{13}$$

where $\text{AVar}_{\tilde{F}}(Z)$ *denotes the scaled asymptotic variance of random variable* Z *over the sampling distribution* $\tilde{F}$*, that is,* $\text{AVar}_{\tilde{F}}(Z) = \lim_{n\to\infty}\text{Var}_{\tilde{F}}(\sqrt{n}Z)$*.* p *is the probability of being treated within the experiment, that is,* $p = \Pr_{\tilde{F}}(T_i = 1)$*.*

The proof of Theorem 2 can be found in Supplementary Material Section 1. Theorem 2 decomposes the efficiency gain from post-residualized weighting into two components: (1) the variance of the predicted experimental outcomes $\text{Var}_w(\hat{Y}_i)$ and (2) how related the predicted outcomes are to the actual outcomes in the experimental samples (represented by $\text{Cov}_w(Y_i(1), \hat{Y}_i)$ and $\text{Cov}_w(Y_i(0), \hat{Y}_i)$). If the covariance between the predicted outcomes and actual outcomes in the experimental sample is greater than the variance of the predicted outcomes, we expect precision gains. In other words, the gains to precision from residualizing depend on how well outcome measures in the experiment are explained by the residualizing model fitted to the population data.[5] As such, researchers should leverage the large amounts of data available at the population level to apply flexible modeling strategies in order to maximize the variation explained by the residualizing model.

More generally, we can formally write the efficiency gain for the post-residualized weighted least squared estimator (equation (9)) as follows.

THEOREM 3 (Efficiency Gain for Post-residualized Weighted Least Squares Estimators). *The difference between the asymptotic variance of* $\hat{\tau}_{\text{wLS}}$ *and that of* $\hat{\tau}_{\text{wLS}}^{\text{res}}$ *is*

$$\begin{aligned}&\text{AVar}_{\tilde{F}}(\hat{\tau}_{\text{wLS}}) - \text{AVar}_{\tilde{F}}(\hat{\tau}_{\text{wLS}}^{\text{res}})\\ &\quad= \frac{1}{p}\left\{\text{Var}_w(Y_i(1) - \tilde{\mathbf{X}}_i^\top\gamma_*) - \text{Var}_w(Y_i(1) - \hat{g}(\mathbf{X}_i))\right\}\\ &\quad\quad\underbrace{+\frac{1}{1-p}\left\{\text{Var}_w(Y_i(0) - \tilde{\mathbf{X}}_i^\top\gamma_*) - \text{Var}_w(Y_i(0) - \hat{g}(\mathbf{X}_i))\right\}}_{(a)\text{ Explanatory power of residualizing model over linear regression}}\\ &\quad\quad\underbrace{+\frac{2}{p}\text{Cov}_w(\hat{e}_i(1), \tilde{\mathbf{X}}_i^\top\gamma_*^{\text{res}}) + \frac{2}{1-p}\text{Cov}_w(\hat{e}_i(0), \tilde{\mathbf{X}}_i^\top\gamma_*^{\text{res}}) - \frac{1}{p(1-p)}\text{Var}_w(\tilde{\mathbf{X}}_i^\top\gamma_*^{\text{res}})}_{(b)\text{ Remaining variation in residualized outcomes explained by linear regression on }\tilde{\mathbf{X}}_i},\end{aligned} \tag{14}$$

where γ_* *and* γ_*^{res} *are the true coefficients*[6] *associated with the pre-treatment covariates,* $\tilde{\mathbf{X}}_i$ *defined in the weighted least squares regression* (*equation* (8))*, and the post-residualized weighted least squares regression* (*equation* (9))*, respectively.*

[5] We note that the efficiency gain expression does not include uncertainty associated with estimating the residualizing model. This is because the chosen $\hat{g}(\mathbf{X}_i)$ is a dimension reducing function of the fixed pre-treatment covariates.

[6] We define the true coefficients as the coefficients that would be estimated as the experimental sample size $n \to \infty$; see Supplementary Material for more information.

When we include covariate adjustment to the experimental data, the gains to precision depend on two factors. The first factor, (a), compares the explanatory power of the residualizing model with the linear regression. More specifically, if $\hat{g}(\mathbf{X}_i)$ is able to explain more variation than the linear combination of $\tilde{\mathbf{X}}_i$, then we expect the first term to be positive. The second term, (b), represents the amount of variation in the residualized outcomes that can be explained by the pre-treatment covariates $\tilde{\mathbf{X}}_i$.

A natural question is, "Why not directly adjust for covariates within the experimental sample instead of using a residualizing model?" One advantage to using the post-residualized weighting over directly adjusting for covariates within the experimental sample arises from the fact that there is typically a larger amount of data available in the population data (i.e., $N \gg n$). While researchers could choose to use a flexible model within the experimental data to perform covariate adjustment, there is a greater restriction with respect to degrees-of-freedom to what type of model can be fit. The availability of large amounts of population data can be leveraged in the residualizing process to better estimate covariate-outcome relationships. Additionally, by using population data to build and tune the residualizing model, we protect the fidelity of inferences using the experimental data since it is only used for estimation of the PATE.

In the following subsection, we will describe a diagnostic measure that can help researchers determine whether or not they should expect precision gains from residualizing.

3.4. *Diagnostics*. As discussed above, while post-residualized weighting stands to greatly improve precision in estimation of the PATE, this is not guaranteed. To address this concern, we derive a diagnostic that evaluates when researchers should expect precision gains from residualizing.

Again to simplify our presentation, we first start with the post-residualized weighted estimator (equation (10)). We can define a pseudo-R^2 measure as

$$R_0^2 := 1 - \frac{\mathrm{Var}_w(\hat{e}_i(0))}{\mathrm{Var}_w(Y_i(0))}, \tag{15}$$

where we define $\hat{e}_i(t) = Y_i(t) - \hat{Y}_i$ for $t \in \{0, 1\}$.

R_0^2 can be interpreted as the weighted goodness-of-fit of the residualizing model for the potential outcomes under control for units in the experiment. Researchers can estimate R_0^2 using the estimated residuals across the control units in the experiment. When $R_0^2 > 0$, we expect an improvement in precision across the control units from residualizing.

More generally, for the post-residualized weighted least squares estimator (equation (9)) we can define R_0^2 as

$$R_0^2 = 1 - \frac{\mathrm{Var}_w(\hat{e}_i(0) - \widetilde{\mathbf{X}}_i^\top \gamma_*^{\mathrm{res}})}{\mathrm{Var}_w(Y_i(0) - \widetilde{\mathbf{X}}_i^\top \gamma_*)}, \tag{16}$$

where we now include covariate adjustments from weighted least squares regression in our diagnostic. The $\hat{e}_i(0) - \widetilde{\mathbf{X}}_i^\top \gamma_*^{\mathrm{res}}$ are the residuals that arise from regressing the residualized outcomes under control on the pre-treatment covariates in the weighted regression. Similarly, the quantity $Y_i(0) - \widetilde{\mathbf{X}}_i^\top \gamma_*$ are the residuals from regressing the raw outcomes under control on the pre-treatment covariates. In this way we are directly comparing the variance of the outcomes, following covariate adjustment, across the control units. The interpretation of this value is identical to that of the pseudo-R^2 value in the weighted estimator case. It is easy to see that R_0^2 in equation (15) is a special case of R_0^2 in equation (16) when $\widetilde{\mathbf{X}}_i = \varnothing$.

In line with Rubin's "locked box" approach (Rubin (2008)), we do not suggest estimating the analogous R_0^2 among treated units. However, if the variation in the control outcomes is

greater than the overall treatment effect heterogeneity, then checking if R_0^2 is greater or less than zero is an effective diagnostic for whether or not we expect precision gains from residualizing. We formalize this in the following corollary, where we write the relative reduction from residualizing as a function of this proposed R_0^2 measure.

COROLLARY 1 (Relative Reduction from Residualizing). *With R_0^2 defined as in equation (15), define R_1^2 as the analogous weighted goodness-of-fit of the residualizing model for the potential outcomes under treatment*:

$$R_1^2 := 1 - \frac{\mathrm{Var}_w(\hat{e}_i(1) - \widetilde{\mathbf{X}}_i^\top \gamma_*^{\mathrm{res}})}{\mathrm{Var}_w(Y_i(1) - \widetilde{\mathbf{X}}_i^\top \gamma_*)} = R_0^2 - \xi, \text{ where } \xi = R_0^2 - R_1^2.$$

Furthermore, define the ratio $f = p\,\mathrm{Var}_w(Y_i(0) - \tilde{\mathbf{X}}_i^\top \gamma_*)/(1-p)\,\mathrm{Var}_w(Y_i(1) - \tilde{\mathbf{X}}_i^\top \gamma_*)$. Then the relative reduction in variance from residualizing is given by

$$\textit{Relative Reduction} := \frac{\mathrm{AVar}_{\tilde{F}}(\hat{\tau}_{\mathrm{wLS}}) - \mathrm{AVar}_{\tilde{F}}(\hat{\tau}_{\mathrm{wLS}}^{\mathrm{res}})}{\mathrm{AVar}_{\tilde{F}}(\hat{\tau}_{\mathrm{wLS}})} = R_0^2 - \frac{1}{1+f} \cdot \xi.$$

Corollary 1, proof available in Supplementary Material Section 1, decomposes the overall relative reduction in variance of the weighted least squares estimator from residualizing into two components: (1) our proposed diagnostic measure R_0^2 and (2) a factor, represented by ξ, that measures the difference in prediction error between the experimental control and experimental treated potential outcomes. If the residualizing model explains similar amounts of variation across both the treated and control potential outcomes, then $R_1^2 \approx R_0^2$ and $\xi \approx 0$. In that scenario R_0^2 will be roughly indicative of the expected relative reduction. When R_0^2 takes on a negative value, this is a strong indication that residualizing is unlikely to result in precision gains, since it is unlikely the prediction error will be significantly lower for treated units.

To summarize, R_0^2 can diagnose when one should expect improvements in precision from residualizing. When R_0^2 takes on negative values, researchers should not proceed with residualizing, as it is likely to result in precision loss.

4. Extension: Using the predicted outcomes as a covariate. Thus far, we have discussed residualizing or directly subtracting the predicted outcome values from the outcomes measured in the experimental sample. An alternative approach is to regress the outcomes measured in the experimental sample on the predicted outcomes $\hat{Y}_i$ from our residualizing model. In particular, we include $\hat{Y}_i$ as a covariate in a weighted linear regression,

$$\left(\hat{\tau}_W^{\mathrm{cov}}, \hat{\beta}, \hat{\alpha}\right) = \underset{\tau,\beta,\alpha}{\mathrm{argmin}} \frac{1}{n} \sum_{i \in \mathcal{S}} \hat{w}_i \left(Y_i - (\tau T_i + \beta \hat{Y}_i + \alpha)\right)^2.$$

We can extend this approach to also include pre-treatment covariates,

$$\left(\hat{\tau}_{\mathrm{wLS}}^{\mathrm{cov}}, \hat{\beta}, \hat{\alpha}, \hat{\gamma}\right) = \underset{\tau,\beta,\gamma,\alpha}{\mathrm{argmin}} \frac{1}{n} \sum_{i \in \mathcal{S}} \hat{w}_i \left(Y_i - (\tau T_i + \beta \hat{Y}_i + \alpha + \widetilde{\mathbf{X}}_i^\top \gamma)\right)^2.$$

The residualizing methods we discussed in Section 3 can be seen as special cases of these methods where we set $\beta = 1$.

Residualizing by directly including $\hat{Y}_i$ as a covariate in the weighted least squares has many advantages. The primary advantage is that this approach allows researchers to flexibly use proxy outcomes measured in the target population. When the outcome of interest is not measured at the population level or if the outcomes are measured in different ways across the

experimental sample and the observed population data, researchers can estimate the residualizing model $g(\mathbf{X}_i)$ using alternative proxy outcomes $\tilde{Y}_i$ related to the outcome of interest. However, use of these proxies can lead to scaling issues that limit the ability of the weighted and weighted least squares methods for post-residualizing to achieve efficiency gains. We show how including $\hat{Y}_i$ as a covariate addresses these concerns.

As with our post-residualized estimators $\hat{\tau}^{\text{res}}_W$ and $\hat{\tau}^{\text{res}}_{\text{wLS}}$ discussed in Section 3, both $\hat{\tau}^{\text{cov}}_W$ and $\hat{\tau}^{\text{cov}}_{\text{wLS}}$ are consistent estimators of the PATE (Section 4.2). Finally, including the predicted outcome $\hat{Y}_i$ as a covariate protects against efficiency loss, unlike $\hat{\tau}^{\text{res}}_W$ and $\hat{\tau}^{\text{res}}_{\text{wLS}}$ in the previous sections. This is true whether researchers rely on a proxy outcome $\tilde{Y}_i$ or if they build the residualizing model on Y_i.

4.1. *Proxy outcomes in the population data.* There are many settings in which researchers may rely on a proxy outcome $\tilde{Y}_i$. First, an outcome measure used to estimate the residualizing model in the population data may differ from the outcome measure in the experiment. Second, even when the outcome measure used to estimate the residualizing model in the population data is, in principle, the same measure as the outcome of interest in the experimental data, there can be differences between $\tilde{Y}_i$ and Y_i that may arise due to differences in how the outcomes are measured or operationalized across the experimental sample and the population or when the potential outcomes depend on context. For example, this might occur if the population is a mix of both treatment and control conditions with nonrandom treatment selection.

EXAMPLE (JTPA). Assume that we wish to generalize the impact of JTPA on employment in an experimental site to a new target site. However, in this target site, instead of current employment, we only have access to total weeks worked in the past year or whether an individual is collecting unemployment benefits which differ from the employment indicator collected at the end-point in the experiment. These could serve as proxy measures for employment when using post-residualized weighting for generalizing the impact of JTPA to a target site. In Section 6 we use our two primary outcomes, earnings and an employment indicator, as proxies for one another.

EXAMPLE (Get-Out-the-Vote (GOTV) experiments). Consider Get-Out-the-Vote experiments, again, where we are interested in the causal effect of a randomized GOTV message on voter turnout which is measured by administrative voter files in the United States (e.g., Gerber and Green (2000)). Imagine, however, that we do not have administrative data available on our population, such as for all voters in the United States, but rather we have a nationally representative survey. For many nationally representative surveys, it is infeasible to link administrative individual-level voting history data, due to privacy issues and data constraints; as such, we do not have access to voter turnout. Instead, surveys often ask voters an "intent-to-vote" question which can proxy for actual voter turnout. Our proposed method can use this "intent-to-vote" variable to build a residualizing model.

EXAMPLE (Education experiments). Imagine that researchers are primarily interested in the causal effect of small class sizes not on standardized outcomes, such as the SAT, but rather on a curriculum-based test score specific to a state collected during a given academic year. In this case researchers may not have access to this curriculum-based measure in the state-level population data but may have access to related standardized testing scores. These standardized test scores may be used as a proxy to the curriculum-based test score of interest that is measured in the experimental data when constructing the residualizing model.

When using proxy outcomes to estimate the residualizing model, the efficiency gain will be impacted by how similar the proxy outcomes are to the actual outcomes of interest. More formally, consider the following decomposition of the residuals $\hat{e}_i$:

$$\hat{e}_i = \underbrace{Y_i - \tilde{Y}_i}_{\substack{\text{(a)}\\ \text{Difference between outcomes in experiment and proxy outcome}}} + \underbrace{\tilde{Y}_i - \hat{Y}_i}_{\substack{\text{(b)}\\ \text{Prediction Error for proxy outcome}}}, \tag{17}$$

where we define $\tilde{Y}_i$ as the proxy outcome. Conceptually, $\tilde{Y}_i$ represents the proxy outcome, had it been measured for the experimental data. For example, in the JTPA experiment, $\tilde{Y}_i$ could represent the variable for collecting unemployment, had it been measured for the experimental sample.

Equation (17) decomposes the residual term into two components. The second component (b) is the model prediction error. This is driven by how well the chosen residualizing model $g(\mathbf{X}_i)$ fits proxy outcomes measured in the population data. The first component (a) is how similar the proxy outcomes, measured in the population data, are to the outcome measures used in the experimental data. If the proxy outcomes differ substantially from the outcomes measured in the experimental data, while the post-residualized weighted estimators will still be consistent (see Theorem 1), there may be losses in efficiency from residualizing, regardless of how much we are able to minimize the prediction error in the second term (b).

4.2. *Consistency.* Like the previously proposed post-residualized weighted estimators $\hat{\tau}^{\text{res}}_W$ and $\hat{\tau}^{\text{res}}_{\text{wLS}}$, both $\hat{\tau}^{\text{cov}}_W$ and $\hat{\tau}^{\text{cov}}_{\text{wLS}}$ will be consistent estimators of the PATE. This follows from the fact that $\hat{Y}_i = \widehat{g}(\mathbf{X}_i)$ is just a function of pre-treatment covariates $\mathbf{X}_i$. In this sense we can think of $\hat{\tau}^{\text{cov}}_W$ and $\hat{\tau}^{\text{cov}}_{\text{wLS}}$ as extensions of the weighted least squares estimator, where $\hat{Y}_i$ is an additional pre-treatment covariate included in the weighted linear regression. Thus, as shown in Section 3, both $\hat{\tau}^{\text{cov}}_W$ and $\hat{\tau}^{\text{cov}}_{\text{wLS}}$ are consistent estimators of the PATE.

4.3. *Efficiency gain and diagnostics*. There are two advantages to using $\hat{Y}_i$ as an additional covariate. First, because $\hat{Y}_i$ is treated as a covariate in a weighted regression, the estimated coefficient (i.e., $\hat{\beta}$) can capture any potential scaling differences between the proxy outcomes and the actual outcomes of interest. While the standard post-residualized weighted estimator can account for additive differences between the proxy outcome and actual outcome, including $\hat{Y}_i$ as a covariate in a weighted regression allows for our method to additionally account for scale differences between the proxy and actual outcomes. For example, returning to the Get-Out-the-Vote experiments, intent-to-vote is often measured on a Likert scale, while voter turnout is simply a binary variable of whether the individual voted or not. In such a scenario, residualizing directly on $\hat{Y}_i$ can lead to efficiency loss, despite the fact that intent-to-vote is correlated to voter turnout.

Second, treating $\hat{Y}_i$ as a covariate protects against precision loss when the proxy outcomes are significantly different from the outcomes of interest. At worst, $\hat{Y}_i$ is unrelated to Y_i, and we expect the coefficient in front of $\hat{Y}_i$ to be near zero. When this occurs, we expect the variance of the post-residualized weighted estimator, when using $\hat{Y}_i$ as a covariate, to be similar to the variance of a conventional estimator that does not include population-level outcome information. Below we demonstrate this more formally.

COROLLARY 2. *The post-residualized weighted estimators using $\hat{Y}_i$ as a covariate will be at least as asymptotically efficient as the standard weighted estimators*,

$$\text{AVar}(\hat{\tau}_W) - \text{AVar}(\hat{\tau}^{\text{cov}}_W) \geq 0,$$
$$\text{AVar}(\hat{\tau}_{\text{wLS}}) - \text{AVar}(\hat{\tau}^{\text{cov}}_{\text{wLS}}) \geq 0.$$

This result follows from Ding (2021), who shows that the variance of an estimator that accounts for pre-treatment covariates will be asymptotically less than or equal to the variance of an estimator that does not account for pre-treatment covariates.

To account for whether or not the predicted outcomes sufficiently explain enough of the variation in the experimental sample, we extend our previously proposed diagnostic measures to the proxy outcome setting. To do so, we propose using sample splitting across the control units in the experimental sample. We regress $\hat{Y}_i$ on the control outcomes Y_i across one subset of the sample. This allows us to estimate $\hat{\beta}$. Then, using $\hat{\beta}$, we can estimate residuals, accounting for the scaling factor (i.e., $Y_i - \hat{\beta}\hat{Y}_i$), across the held out sample, and calculate the $\hat{R}^2_0$ and $\hat{R}^2_{0,\text{wLS}}$ diagnostics from before. Finally, we conduct cross-fitting, that is, repeating the same procedure by flipping the role of training and test data and then averaging diagnostics from both sample splits.

4.4. *When to worry about external validity.* When diagnostic measures indicate that post-residualized weighting is unsuitable for the data at hand, it is important to understand why. In particular, equation (17) shows that efficiency loss could occur from: (1) the residualizing model's prediction error and (2) the difference between the outcomes in the population and the outcomes measured in the experimental sample. Low diagnostic values indicate that post-residualizing methods may not provide efficiency gains; however, it may also be indicative of contextual differences in the potential outcomes which affect the validity of the PATE estimate.

The residualizing model's prediction error, from equation (17)-(b), can be estimated through cross-validation using the population-level data. Researchers can hold out random subsets of the population-level data when estimating the residualizing model and calculate the prediction error across the held out sample. If the cross-validated error is large, there will likely be little to no efficiency gains from using post-residualized weighting, due to poor prediction, even if the true outcome Y_i were used to estimate $\hat{g}$. The difference between the outcomes Y_i and the proxy outcome $\tilde{Y}_i$, from equation (17)-(a), can be estimated when the proxy outcome is also measured in the experimental sample. For example, in the Get-Out-the-Vote experiments, researchers may have voters' intent-to-vote in the experimental sample. Alternatively, in the education experiments researchers could measure both the curriculum-based test score and the standardized test score in the experimental sample. In JTPA, employment outcomes may be operationalized differently across sites.

In settings where $\tilde{Y}_i$ is not measured in the experimental data, researchers can still use the proposed diagnostic measures to determine if there are concerns about generalizability. For example, if the cross-validated prediction error is low but the diagnostics indicate that post-residualized weighting will not improve efficiency, then this indicates that the residualizing model predicts the population outcomes well but does not predict outcomes measured in the experiment well. This could be due to two problems. First, if the population outcome is a proxy measure of the outcome measured in the experimental sample, then it could be that the measure used in the population data is not a good proxy for the experimental outcome. Alternatively, if researchers believe that the experimental and population outcomes are measured in the same way, then a low or negative R^2_0 measure, in conjunction with low cross-validated prediction error, would indicate that the outcome-covariate relationships in the population are considerably different from the outcome-covariate relationships in the experimental sample. In this case there may be limited external validity of the experiment, due to a failure of the consistency of parallel studies assumption, since the potential outcomes may depend on context (see Egami and Hartman (2022) for more discussion).

5. Simulation. We now run a series of simulations to empirically examine the proposed post-residualizing method. In total, we consider four different data-generating scenarios, based on the following model for the potential outcomes under control:

$$Y_i(0) = \beta_1 X_{1i} + \beta_2 X_{2i} + \gamma_1 X_{1i}^2 + \gamma_2\sqrt{|X_{2i}|} + \gamma_3(X_{1i} \cdot X_{2i}) + \beta_S \cdot (1 - S_i) \cdot (\alpha + \beta_3 X_{1i} + \gamma_4 X_{1i} \cdot X_{2i}) + \varepsilon_i,$$

where (X_{1i}, X_{2i}) are observed pre-treatment covariates and $S_i \in \{0, 1\}$ is a binary indicator variable, taking the value of one when unit i is in the experimental data and taking the value of zero when unit i is in the population data. β_S controls for differences between the experimental sample and population data outcomes, and the γ terms dictate the nonlinearity of the data-generating processes.

We then define the treatment effect model as follows:

$$\tau_i = \alpha_\tau + X_{\tau,i},$$

where $X_{\tau,i}$ is an observed pre-treatment covariate that governs treatment effect heterogeneity. Therefore, the observed outcomes take on the following form: $Y_i = Y_i(0) + \tau_i \cdot T_i$. We provide additional details, including the sampling model and distributions of observed covariates, in Supplementary Material Section 2.

The first two scenarios test the method when the outcome measures for both the experimental sample and the population data are drawn from the same underlying data-generating process to explore a setting where the outcome is measured identically across the experiment and target population (i.e., $\beta_S = 0$). The third and fourth scenarios use different data-generating processes to simulate a context where the outcome measure differs between the experimental sample and the population (i.e., $\beta_S \neq 0$). This represents real-world settings in which the outcomes in the experimental sample and population are measured differently or are situated in different contexts, which can result in differences in the outcome-covariate relationships. This setting also mimics the case in which researchers use a proxy outcome. For each of these settings, we consider a version of the data generating processes that is linear in the included covariates ($\gamma_\circ = 0$–i.e., all γ coefficients are set to zero) and a second version that contains nonlinearities ($\gamma_\circ \neq 0$). Table 2 provides a summary of the different scenarios.

We compare conventional and post-residualized versions of two sets of estimators in each simulation. We perform post-residualizing in two different ways: the first directly residualizes the outcomes in the experimental sample by subtracting the predicted outcomes, and the second treats the predicted outcomes as a covariate in a weighted regression. Therefore, we compare a total of six different estimators: (1) the weighted estimators $\hat{\tau}_W$, $\hat{\tau}_W^{\text{res}}$, $\hat{\tau}_W^{\text{cov}}$, and (2) weighted least squares (wLS) $\hat{\tau}_{\text{wLS}}$, $\tau_{\text{wLS}}^{\text{res}}$, and $\hat{\tau}_{\text{wLS}}^{\text{cov}}$. The difference-in-means estimator (DiM) is also provided as a baseline with no weighting adjustment.

The underlying sampling process is governed by a logit model. At each iteration of the simulation, we draw both a biased experimental sample and a random sample of a larger

TABLE 2
Summary of different simulation scenarios

	Proxy and experimental sample outcomes	DGP type
Scenario 1	Identical DGP ($\beta_S = 0$)	Linear ($\gamma_\circ = 0$)
Scenario 2	Identical DGP ($\beta_S = 0$)	Nonlinear ($\gamma_\circ \neq 0$)
Scenario 3	Different DGP ($\beta_S \neq 0$)	Linear ($\gamma_\circ = 0$)
Scenario 4	Different DGP ($\beta_S \neq 0$)	Nonlinear ($\gamma_\circ \neq 0$)

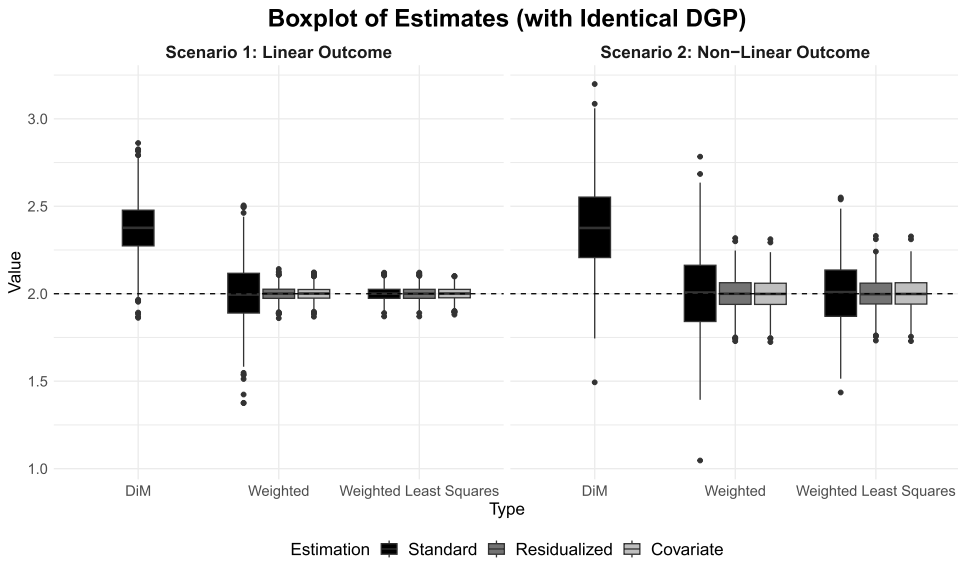

FIG. 2. *Summary of estimates across* 1000 *simulations for Scenarios* 1 *and* 2 *in which the experimental sample and population outcomes are drawn from the same data-generating process. The dashed line represents the super-population PATE.*

target population as the population data. The population data is used to estimate the residualizing model and sampling weights. We use entropy balancing to estimate the sampling weights $\hat{w}_i$ for each simulation. Our residualizing model is a regression that contains all the pairwise interactions of the included covariates. The weighted least squares regression includes covariates additively without any interactions.[7] Our results follow:

Overall, we find that when the underlying outcome model is complex and contains nonlinear terms, our post-residualizing method exhibits large precision gains compared to conventional methods. When there is no difference between the population-level outcomes and the outcomes in the experimental sample, seen in Figure 2, direct residualizing and including $\hat{Y}_i$ as a covariate performs identically.

Scenario 1. When we consider a linear DGP, residualizing results in substantial precision gains for the weighted estimator. However, for the weighted least squares estimator, residualizing does not result in precision gains, because the covariate adjustment taking place in the weighted regression already includes the linear terms in the data-generating process, and thus, the residualizing step does not model anything in the outcomes that is not already accounted for in the wLS regression.

Scenario 2. When we include nonlinear terms into the data-generating process, residualizing results in precision gains for all of the estimators, because the residualizing model is able to account for some of the nonlinearities that the wLS regression does not account for. It is worth noting that the estimated residualizing model is not a correct specification of the underlying outcome model for the population data. However, because we have included the pairwise interactions between the covariates, the residualizing model is able to significantly reduce the variance for both estimators, even without accounting for all of the nonlinear terms in the underlying data-generating process.

[7]It is possible, in practice, to include nonlinear transformation of pre-treatment covariates in the regression adjustment step. However, we have omitted it to illustrate the efficiency gains that can be obtained from accounting for nonlinearities through the residualizing step. This mimics how, in practice, we are able to fit more complex models to more data.

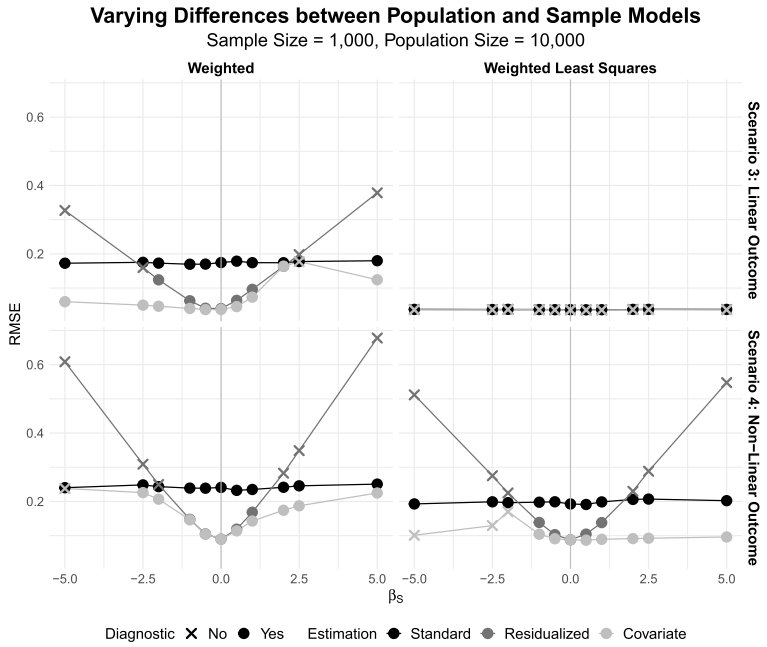

FIG. 3. *Plot of RMSE of the different estimators for Scenarios* 3 *and* 4, *in which the experimental sample and population outcomes are drawn from different data generating processes.* β_S *controls for how different the two processes are* (*i.e., the larger* $|\beta_S|$ *is, the larger the difference between the two processes*). *The standard estimators are presented in black and the residualized estimators in gray and light gray. We label all the points for which the diagnostic measure estimates a loss* (×) *or gain* (•) *in efficiency from residualizing more than* 50% *of the time in the* 1000 *iterations.*

Scenarios 3 *and* 4. Next, we consider a difference in the underlying data-generating process between the experimental and population outcomes, presented in Figure 3. We operationalize this by including an interaction between treatment, the sampling indicator, and covariates. The degree to which the two processes differ is varied across different simulations using a single parameter, β_S. When the difference is relatively small (i.e., small $|\beta_S|$), the two methods used to residualize the experimental sample outcomes perform identically. This is evident by a lower RMSE when $|\beta_S| < 2$ for the post-residualized weighted estimators. When the difference in the DGP are large (i.e., $|\beta_S| > 2$), residualizing by directly subtracting the outcomes from the predicted outcomes results in precision loss, evident by a larger RMSE for the post-residualized weighted estimator $\hat{\tau}_W^{\text{res}}$ and for the post-residualized weighted least square estimator $\hat{\tau}_{\text{wLS}}^{\text{res}}$ when the true DGP is nonlinear. However, treating the predicted outcomes as a covariate in a weighted linear regression $\hat{\tau}_W^{\text{cov}}$ and $\hat{\tau}_{\text{wLS}}^{\text{cov}}$ allows for precision gain, even in these settings. We see that, at worst, the covariate-based residualizing approach performs equivalently to the conventional estimators.

It is important to highlight that, regardless of the degree of divergence between the population and experimental sample DGP's, post-residualized weighting is able to maintain nominal coverage. Furthermore, our proposed diagnostic measures adequately capture when we expect to gain or lose precision from residualizing. We provide coverage results and a summary of the diagnostic performance in Supplementary Material Section 2.

6. Empirical evaluation: Job Training Partnership Act. To evaluate and benchmark how our proposed post-residualizing method may work in practice, we now turn to an empirical application. Recall that, while the original study evaluated the overall impact of JTPA, our focus is on generalizing the effect of each site individually to the other 15 sites. More specifically, in our leave-one-out analysis for each site we define the PATE as the average

treatment effect among units in the remaining 15 sites. We then generalize the experimental results from one site to the population defined by the pooled remaining sites. This allows us to validate our method's performance by comparing our PATE estimators to the pooled experimental benchmark in the remaining sites. We evaluate generalizability for two outcomes: employment status (binary outcome) and total earnings (zero-inflated, continuous outcome).

6.1. *Post-residualized Weighting*.

6.1.1. *Residualizing model*. We include baseline covariates measured at the interview stage of the JTPA study. The covariates include measures of age, previous earnings, marital status, household composition, public assistance history, education and employment history, access to transportation, and ethnicity. More details about the pre-treatment covariates can be found in Supplementary Material Section 3.

We construct our residualizing model using an ensemble method, the *SuperLearner* (van der Laan, Polley and Hubbard (2007)). The ensemble model contains the random forest, with varying hyperparameters, and the LASSO, with hyperparameters chosen using cross validation. This allows us to capture nonlinearities in the data through the random forest as well as linear relationships using the LASSO (van der Laan, Polley and Hubbard (2007)). We build separate models for the probability of employment and total earnings. We fit our residualizing model on the control units from the target population. Details can be found in Supplementary Material Section 3.

6.1.2. *Estimators*. We estimate the PATE using two different estimators: the weighted estimator and the weighted least squares estimator (wLS). For each estimator we consider the conventional estimators ($\widehat{\tau}_W$ and $\widehat{\tau}_{\text{wLS}}$), the post-residualized estimators directly subtracting the predicted outcomes from the outcomes in the experimental sample ($\widehat{\tau}_W^{\text{res}}$ and $\widehat{\tau}_{\text{wLS}}^{\text{res}}$), and the post-residualized estimators using the predicted outcomes as a covariate ($\widehat{\tau}_W^{\text{cov}}$ and $\widehat{\tau}_{\text{wLS}}^{\text{cov}}$). Sampling weights are estimated using entropy balancing in which we match main margins for age, education, previous earnings, race, and marital status (Hainmueller (2012)). Our weighted least squares (wLS) estimators include age, education level, and marital status as controls. Standard errors are estimated using heteroskedastic-consistent standard errors (HC2).

6.1.3. *Diagnostics*. For each site we compute the pseudo-R^2 diagnostics. This can be done directly for the post-residualized weighted and weighted least squares estimators. When treating $\hat{Y}_i$ as a covariate, we use sample splitting to estimate the pseudo-R^2 values. Because some of the experimental sites comprise relative few units (i.e., the experimental site of Montana contains only 38 units total), we perform repeated sample splitting, taking the average of the diagnostic across the repeated splits (Chernozhukov et al. (2018), Jacob (2020)).

6.2. *Results*.

6.2.1. *Bias*. Because the conventional estimators and our proposed approach rely on the same identification assumptions, we first want to verify that the overall bias in the PATE estimation is not affected by the post-residualized weighting step. Across all 16 sites, the point estimates from post-residualized weighting do not change substantially from standard estimation approaches. Even in experimental sites in which it may not be advantageous to perform post-residualized weighting for efficiency gains, point estimates from post-residualized weighting methods are close to those from the conventional weighting estimators. We report the mean absolute error for all 16 sites in Supplementary Material Table A7.

6.2.2. *Diagnostics*. To evaluate whether the post-residualized weighting estimators provide efficiency gains over conventional approaches, we estimate our diagnostics. Supplementary Material Table A9 summarizes the performance of the diagnostic measures across all 16 sites for both earnings and employment.

On average, we see that the proposed diagnostic measures are able to adequately capture when researchers should expect precision gains from residualizing. The $\hat{R}_0^2$ diagnostic has a high true positive rate for both directly residualizing and using $\hat{Y}_i$ as a covariate. As such, when the diagnostic measures indicate that researchers should residualize, residualizing results in precision gains. In the case when we are directly residualizing, the diagnostic measure also has a relatively high true negative rate which implies that, when $\hat{R}_0^2 < 0$, there is a loss in precision from directly residualizing. In the case of including $\hat{Y}_i$ as a covariate, there is a greater false negative rate, as the diagnostic tends to be more conservative in this setting. This is especially noticeable when employment is the outcome. Many of the false negatives here correspond to estimated $\hat{R}_0^2$ values that are negative but very close to zero.

6.2.3. *Efficiency gain*. Results on the efficiency gains to post-residualized weighting are summarized in Table 3 and graphically displayed in Figure 4. Restricting our attention to the sites for which the $\hat{R}_0^2$ values are greater than zero, there is a large reduction in variance overall from residualizing. When directly residualizing, for earnings, residualizing results in a 21% reduction in estimated variance for the weighted estimator and a 12% reduction for the weighted least squares estimator. For employment, directly residualizing leads to a 10% reduction in estimated variance for the weighted estimator and a 5% reduction for the weighted least squares estimator.

When using $\hat{Y}_i$ as a covariate, we see that including the predicted outcomes as a covariate results in a 25% reduction in variance for the weighted estimator and 16% reduction for weighted least squares when earnings is the outcome. For employment, adjusting for the predicted outcomes results in a 9% reduction in variance for the weighted estimator and a 4% reduction for the weighted least squares.

There are several takeaways to highlight. First, we see that directly residualizing the outcomes can result in significant precision gain. In particular, the reduction in variance in the post-residualized weighted least squares demonstrates the advantage residualizing has over

TABLE 3

Summary of gains to post-residualized weighting. Columns 1 *and* 4 *give the number of sites for which the diagnostic measure indicates gains to post-residualized weighting. The average standard error among selected sites are presented for the conventional estimators* (*columns* 2 *and* 5) *and post-residualized estimators* (*columns* 3 *and* 6)

Summary of standard errors across experimental sites subset by diagnostic

	Earnings			Employment		
	Number of sites	Conventional	Post-resid. weighting	Number of sites	Conventional	Post-resid. weighting
Weighted						
Direct Residualizing	10	2.42	2.13	11	8.33	7.81
$\hat{Y}_i$ as Covariate	7	2.17	1.86	1	5.58	5.01
Weighted Least Squares						
Direct Residualizing	12	2.71	2.56	11	7.88	7.64
$\hat{Y}_i$ as Covariate	7	1.87	1.71	1	5.56	5.45

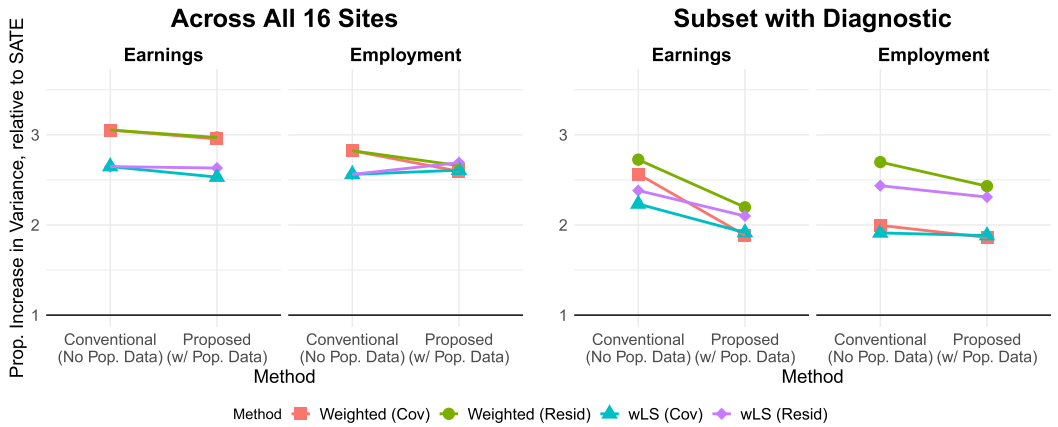

FIG. 4. *Reduction in variance from using post-residualized weighting. We calculate the variance of the estimators, relative to the variance of the difference-in-means (DiM) estimator. We can interpret the y-axis as the amount of variance inflation that is incurred from generalization and see that using the proposed method of incorporating population data can allow us to offset some of the precision loss incurred from reweighting. We see that, when using the proposed diagnostic measure, post-residualized weighting results in substantial precision gains across all four estimators for identified sites.*

just using regression adjustment. Second, the larger reduction in variance from using $\hat{Y}_i$ as a covariate underscores the value of being able to capture the scaled relationships between the outcomes in the population data and in the experimental sample.

Figure 4 shows the relative variance of the PATE estimators to the unweighted SATE. It is well known that PATE estimators typically have higher variance than the SATE (Miratrix et al. (2018)); however, we see that, with the post-residualized method, some of the precision loss incurred from the weighted PATE estimators can be offset. Table 3 provides a summary of the standard errors of the PATE estimators, relative to the difference-in-means estimators.

In the left panel of Figure 4, we also report the results when pooling all 16 sites together, which represents the setting in which researchers do not use the diagnostic and naively perform post-residualized weighting across all settings. We still generally see some improvements in precision from using post-residualized weighting. However, the improvements are much smaller than in the setting in which we subset to sites using the diagnostic measure. As such, we recommend that, when possible, researchers should use the proposed diagnostic measures.

7. Conclusion. In this paper we introduce post-residualized weighting as a method for mitigating the precision cost of generalizing experiments to larger populations. Existing estimators for population effects typically have high variance, especially if some sampling weights are extreme (Miratrix et al. (2018)), making it difficult for policymakers and practitioners to draw conclusions about the impact of treatment in the target population. For example, in our stylized example a single site from the JTPA might not be representative of the full experiment, so a generalized estimate based on it would potentially be too lacking in precision to inform any policy decision. Our precision gains come from leveraging a valuable type of data that has been typically unused in the generalizability literature so far: outcome data measured in the target population.

To assess the benefits of our approach in practice, we reevaluate the impact of the Job Training Partnership Act (JTPA), using the multisite nature of the experiment to benchmark the performance of our estimators, relative to common methods using a within study comparison approach. We evaluate two outcomes, employment and earnings. We find that the post-residualized methods result in a 5–25% average reduction in variance and that confidence

intervals maintain nominal coverage. We achieve the most significant gains from including the predicted outcomes as a covariate, underscoring the value of this method when scaling issues may be present in the relationship between the outcomes in the population data and in the experimental sample. Finally, our diagnostic measures accurately capture when the post-residualized estimators result in precision gains in estimation of the PATE.

In short, our proposed method first builds a flexible model using population outcome and covariate data which is then used to residualize the experimental outcome data. We show that post-residualized weighting estimators, which rely on residualized outcomes, are consistent for the PATE under the same identifying assumptions as current methods. However, by utilizing residualized outcomes, the post-residualized weighting estimators can obtain large precision gains over conventional approaches. We propose three classes of post-residualized weighting estimators: a weighting estimator using the residualized experimental outcomes, a weighted least squares estimator based on the residualized experimental outcomes, and an extension of weighted least squares in which the predicted values of the residualizing model are included as a covariate.

Our proposed framework has many advantages. As discussed in Section 3.1, the residualizing model, $g(\mathbf{X}_i)$, is an "algorithmic model" that merely needs to adequately predict the outcomes measured in the experiment but does not need to be correctly specified. This allows researchers a great deal of flexibility in constructing it. In Section 4 we discuss how researchers can leverage proxy outcomes that are correlated with, but different from, the outcome measured in the experimental setting. Finally, we provide diagnostic measures, based on the outcomes measured among experimental controls, that allow researchers to determine whether post-residualized weighting will likely improve precision in estimating the PATE.

We evaluate our three post-residualized estimators through simulation studies and an empirical application. Our simulations and JTPA application show significant precision gains from post-residualized weighting and confirm the performance of the diagnostic measure to differentiate when researchers should expect precision gains from post-residualized weighting. We also find that including the predicted outcomes as a covariate ensures that post-residualized weighting does not hurt precision.

Acknowledgments. The authors would like to thank Nicole Pashley, Dustin Tingley, Tara Slough, the Miratrix CARES Lab, and the UCLA Causal Inference reading group.

Funding. Melody Huang is supported by the National Science Foundation Graduate Research Fellowship under Grant No. 2146752. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation.

SUPPLEMENTARY MATERIAL

Supplementary materials: Proofs and additional analyses (DOI: 10.1214/22-AOAS1712SUPP; .pdf). This file contains proofs and derivations, detailed simulation results, and additional information and analyses for the empirical application.

REFERENCES

ATHEY, S., CHETTY, R. and IMBENS, G. (2020). Combining experimental and observational data to estimate treatment effects on long term outcomes. arXiv preprint arXiv:2006.09676.

ATHEY, S., CHETTY, R., IMBENS, G. W. and KANG, H. (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical Report, National Bureau of Economic Research.

BALDASSARRI, D. and ABASCAL, M. (2017). Field experiments across the social sciences. *Annu. Rev. Sociol.* **43** 41–73.

BANERJEE, A. V. and DUFLO, E. (2009). The experimental approach to development economics. *Ann. Rev. Econ.* **1** 151–178.

BAREINBOIM, E. and PEARL, J. (2016). Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci. USA* **113** 7345–7352.

BLOOM, H. S. ORR, L. CAVE, G. BELL, S. and DOOLITTLE, F. (1993). The national JTPA study. Title II-a impacts on earnings and employment at 18 months. Bethesda, MD: Abt Associates.

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

BUCHANAN, A. L., HUDGENS, M. G., COLE, S. R., MOLLAN, K. R., SAX, P. E., DAAR, E. S., ADIMORA, A. A., ERON, J. J. and MUGAVERO, M. J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *J. Roy. Statist. Soc. Ser. A* **181** 1193–1209. MR3876388 https://doi.org/10.1111/rssa.12357

CHERNOZHUKOV, V., DEMIRER, M., DUFLO, E. and FERNANDEZ-VAL, I. (2018). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India. Technical Report, National Bureau of Economic Research.

COLE, S. R. and STUART, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *Am. J. Epidemiol.* **172** 107–115. https://doi.org/10.1093/aje/kwq084

COLNET, B., MAYER, I., CHEN, G., DIENG, A., LI, R., VAROQUAUX, G., VERT, J.-P., JOSSE, J. and YANG, S. (2020). Causal inference methods for combining randomized trials and observational studies: A review. arXiv preprint arXiv:2011.08047.

DAHABREH, I. J., ROBERTSON, S. E., TCHETGEN, E. J., STUART, E. A. and HERNÁN, M. A. (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics* **75** 685–694. MR3999190 https://doi.org/10.1111/biom.13009

DEATON, A. and CARTWRIGHT, N. (2018). Understanding and misunderstanding randomized controlled trials. *Soc. Sci. Med.* **210** 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

DEVILLE, J.-C. and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.* **87** 376–382. MR1173804

DING, P. (2021). Two seemingly paradoxical results in linear models: The variance inflation factor and the analysis of covariance. *J. Causal Inference* **9** 1–8. MR4289523 https://doi.org/10.1515/jci-2019-0023

EGAMI, N. and HARTMAN, E. (2021). Covariate selection for generalizing experimental results: Application to a large-scale development program in Uganda. *J. Roy. Statist. Soc. Ser. A* **184** 1524–1548. MR4344647 https://doi.org/10.1111/rssa.12734

EGAMI, N. and HARTMAN, E. (2022). Elements of external validity: Framework, design, and analysis *Amer. Polit. Sci. Rev.* **First View**, 1–19. https://doi.org/10.1017/S0003055422000880

FALK, A. and HECKMAN, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science* **326** 535–538.

FORD, I. and NORRIE, J. (2016). Pragmatic trials. *N. Engl. J. Med.* **375** 454–463. PMID: 27518663. https://doi.org/10.1056/NEJMra1510059

FREEDMAN, D. A. (2008). On regression adjustments in experiments with several treatments. *Ann. Appl. Stat.* **2** 176–196. MR2415599 https://doi.org/10.1214/07-AOAS143

GERBER, A. S. and GREEN, D. P. (2000). The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *Amer. Polit. Sci. Rev.* **94** 653–663.

HAINMUELLER, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* 25–46.

HARTMAN, E., GRIEVE, R., RAMSAHAI, R. and SEKHON, J. S. (2015). From sample average treatment effect to population average treatment effect on the treated: Combining experimental with observational studies to estimate population treatment effects. *J. Roy. Statist. Soc. Ser. A* **178** 757–778. MR3348358 https://doi.org/10.1111/rssa.12094

HUANG, M., EGAMI, N., HARTMAN, E. and MIRATRIX, L. (2023). Supplement to "Leveraging population outcomes to improve the generalization of experimental results: Application to the JTPA study." https://doi.org/10.1214/22-AOAS1712SUPP

HUBER, J. (2013). Is theory getting lost in the "identification revolution"? The Monkey Cage.

JACOB, D. (2020). Cross-fitting and averaging for machine learning estimation of heterogeneous treatment effects. arXiv preprint arXiv:2007.02852.

KALLUS, N. and MAO, X. (2020). On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. arXiv preprint arXiv:2003.12408.

KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458 https://doi.org/10.1214/07-STS227

KERN, H. L., STUART, E. A., HILL, J. and GREEN, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *J. Res. Educ. Eff.* **9** 103–127. https://doi.org/10.1080/19345747.2015.1060282

LALONDE, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *Amer. Econ. Rev.* 604–620. https://doi.org/10.2307/1806062

LIN, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Ann. Appl. Stat.* **7** 295–318. MR3086420 https://doi.org/10.1214/12-AOAS583

MIRATRIX, L. W., SEKHON, J. S., THEODORIDIS, A. G. and CAMPOS, L. F. (2018). Worth weighting? How to think about and use weights in survey experiments. *Polit. Anal.* **26** 275–291. https://doi.org/10.1017/pan.2018.1

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles (with discussion). Section 9 (translated). *Statist. Sci.* **5** 465–472. MR1092986

NGUYEN, T. Q., EBNESAJJAD, C., COLE, S. R. and STUART, E. A. (2017). Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *Ann. Appl. Stat.* **11** 225–247. MR3634322 https://doi.org/10.1214/16-AOAS1001

O'MUIRCHEARTAIGH, C. and HEDGES, L. V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 195–210. MR3234340 https://doi.org/10.1111/rssc.12037

PEARL, J. and BAREINBOIM, E. (2014). External validity: From do-calculus to transportability across populations. *Statist. Sci.* **29** 579–595. MR3300360 https://doi.org/10.1214/14-STS486

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

ROSENBAUM, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286–327. With comments and a rejoinder by the author. MR1962487 https://doi.org/10.1214/ss/1042727942

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688. https://doi.org/10.1037/h0037350

RUBIN, D. B. (1980). Discussion of 'Randomization analysis of experimental data: The Fisher randomization test comment' by Basu. *J. Amer. Statist. Assoc.* **75** 591–593. https://doi.org/10.2307/2287653

RUBIN, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* **2** 808–804. MR2516795 https://doi.org/10.1214/08-AOAS187

SALES, A. C., HANSEN, B. B. and ROWAN, B. (2018). Rebar: Reinforcing a matching estimator with predictions from high-dimensional covariates. *J. Educ. Behav. Stat.* **43** 3–31. https://doi.org/10.3102/1076998617731518

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812 https://doi.org/10.1214/09-STS313

STUART, E. A. and RHODES, A. (2017). Generalizing treatment effect estimates from sample to population: A case study in the difficulties of finding sufficient data. *Eval. Rev.* **41** 357–388. https://doi.org/10.1177/0193841X16660663

STUART, E. A., COLE, S. R., BRADSHAW, C. P. and LEAF, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *J. Roy. Statist. Soc. Ser. A* **174** 369–386. MR2898850 https://doi.org/10.1111/j.1467-985X.2010.00673.x

TIPTON, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *J. Educ. Behav. Stat.* **38** 239–266.

VAN DER LAAN, M. J., POLLEY, E. C. and HUBBARD, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* **6** Art. 25, 23. MR2349918 https://doi.org/10.2202/1544-6115.1309

WESTREICH, D. and COLE, S. R. (2010). Invited commentary: Positivity in practice. *Amer. J. Epidemiol.* **171** 674–677. https://doi.org/10.1093/aje/kwp436

WORD, E. R. et al. (1990). The state of Tennessee's student/teacher achievement ratio (STAR) project: Technical report (1985–1990). Final summary report. Nashville: Tennessee Dept. Education.