Research Article

Johann A. Gagnon-Bartsch*#, Adam C. Sales*#, Edward Wu, Anthony F. Botelho, John A. Erickson, Luke W. Miratrix, and Neil T. Heffernan

Precise unbiased estimation in randomized experiments using auxiliary observational data

https://doi.org/10.1515/jci-2022-0011
received February 15, 2022; accepted February 11, 2023

Abstract: Randomized controlled trials (RCTs) admit unconfounded design-based inference – randomization largely justifies the assumptions underlying statistical effect estimates – but often have limited sample sizes. However, researchers may have access to big observational data on covariates and outcomes from RCT non-participants. For example, data from A/B tests conducted within an educational technology platform exist alongside historical observational data drawn from student logs. We outline a design-based approach to using such observational data for variance reduction in RCTs. First, we use the observational data to train a machine learning algorithm predicting potential outcomes using covariates and then use that algorithm to generate predictions for RCT participants. Then, we use those predictions, perhaps alongside other covariates, to adjust causal effect estimates with a flexible, design-based covariate-adjustment routine. In this way, there is no danger of biases from the observational data leaking into the experimental estimates, which are guaranteed to be exactly unbiased regardless of whether the machine learning models are "correct" in any sense or whether the observational samples closely resemble RCT samples. We demonstrate the method in analyzing 33 randomized A/B tests and show that it decreases standard errors relative to other estimators, sometimes substantially.

Keywords: education research, A/B testing, data integration

MSC 2020: 62-08, 62D20, 62P25

1 Introduction

Randomized controlled trials (RCTs) are famously free of confounding bias. Indeed, a class of estimators, often referred to as "design-based" [1] or "randomization based" [2], estimates treatment effects without assuming any statistical model other than whatever is implied by the experimental design itself. Design-based statistical estimators are typically guaranteed to be unbiased. Their associated inference – standard

# These authors are contributed equally.

*** Corresponding author: Johann A. Gagnon-Bartsch,** Department of Statistics, University of Michigan, Ann Arbor, Michigan, United Sates, e-mail: johanngb@umich.edu
*** Corresponding author: Adam C. Sales,** Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, Massachusetts, United Sates, e-mail: asales@wpi.edu
Edward Wu: Biocomplexity Institute, Social and Decision Analytics Division, University of Virginia, Charlottesville, Virginia, United Sates
Anthony F. Botelho: College of Education, University of Florida, Gainesville, Florida, United Sates
John A. Erickson: Analytics and Information Systems, Western Kentucky University, Bowling Green, KY 42101, United States
Luke W. Miratrix: Graduate School of Education, Harvard University, Cambridge, Massachusetts, United States
Neil T. Heffernan: Department of Computer Science, Worcester Polytechnic Institute, Worcester, Massachusetts, United Sates

errors, hypothesis tests, and confidence intervals – also comes with accuracy guarantees. In many cases, these apply regardless of the sample size and require only very weak regularity conditions.

While RCTs can reliably provide unbiased estimates, they are often limited in terms of precision. The statistical precision of RCT-based estimates is inherently limited by the RCT's sample size, which itself is typically subject to a number of practical constraints.

In contrast, large observational datasets can frequently be brought to bear on some of the same questions addressed by an RCT. Analysis of observational data, unlike RCTs, typically requires a number of untestable modeling assumptions, chief among them the assumption of no unmeasured confounding. Consequently, treatment effect estimates from observational data cannot boast the same guarantees to accuracy as estimates from RCTs. That said, in many cases, they boast a much larger sample – and, hence, greater precision – than equivalent RCTs.

In many cases, observational and RCT data coexist within the very same database. For instance, covariate and outcome data for a biomedical RCT may be drawn from a database of electronic health records, and that same database may contain equivalent records for patients who did not participate in the study and were not randomized. Along similar lines, covariate and outcome data for an RCT designed to evaluate the impact of an educational intervention might be drawn from a state administrative database, and that database may also contain information on hundreds of thousands of students who did not participate in the RCT. We refer to these individuals, who are nonparticipants of the RCT but who are in the same database, as the *remnant* from the study [3]. We ask, how can we use the remnant to improve power to detect effects in RCTs?

An example from the field of education is www.ETRIALStestbed.org (formerly the ASSISTments TestBed [4,5]). The TestBed is an A/B testing program designed for conducting education research that runs within ASSISTments and has been made accessible to third-party education researchers. Using the TestBed, a researcher can propose A/B tests to run within ASSISTments. That is, a researcher may specify two contrasting conditions, such as video- or text-based instructional feedback, and a particular homework topic, such as "Adding Whole Numbers," or "Factoring Quadratic Equations." Then, students working on that topic are individually randomized between the two conditions. The researcher could then compare the relative impact of video- vs text-based feedback on an outcome variable of interest such as homework completion. The anonymized data associated with the study, consisting of several levels of granularity and rich covariates describing both historical prestudy and within-study student interaction, are made available to the researcher. The TestBed currently hosts over 100 such RCTs, and several of these RCTs have recently been analyzed, e.g., refs [6–11].

In the ASSISTments TestBed example, a given RCT is likely to consist of just a few hundred students assigned to a specific homework assignment, limiting statistical power and precision. For instance, in one typical ASSISTments TestBed A/B test, a total of 294 students were randomized between two conditions, leading to a standard error of roughly four percentage points when estimating the effect on homework completion. This standard error is too large to either determine the direction of a treatment effect or rule out clinically meaningful effect sizes. But the ASSISTments database contains data on hundreds of thousands of other ASSISTments users, many of whom may have completed similar homework assignments, or who may have even completed an identical homework assignment but in a previous time period.

This article outlines an approach to estimate treatment effects in an RCT while incorporating high-dimensional covariate data, large observational remnant data, and machine learning prediction algorithms to improve precision. It does so without compromising the accuracy guarantees of traditional design-based RCT estimators, yielding unbiased point estimates and sampling variance estimates that are conservative in expectation; the approach is design-based, relying only on the randomization within the RCT to make these guarantees. In particular, the method prevents "bias leakage": bias that might have occurred due to differences between the remnant and the experimental sample, biased or incorrect modeling of covariates, or other data analysis flaws, does not leak into the RCT estimator. We combine recent causal methods for within-RCT covariate adjustment with other methods that have sought to incorporate high-dimensional remnant data into RCT estimators. In particular, we focus on the challenge of precisely estimating treatment

effects from a set of 33 TestBed experiments [12], using prior log data from experimental participants and nonparticipants in the ASSISTments system.

The nexus of machine learning and causal inference has recently experienced rapid and exciting development. This has included novel methods to analyze observational studies, e.g., ref. [13], to estimate subgroup effects, e.g., ref. [14], or to optimally allocate treatment, e.g., ref. [15]. Other developments share our goal, i.e., improving the precision of average treatment effect estimates from RCTs. These include the flexible approaches of refs [16–18], all of which can incorporate arbitrary prediction methods; [19], which uses the Lasso regression estimator to analyze experiments; and the targeted learning framework [20,21], which combines ensemble machine learning with semiparametric maximum likelihood estimation.

A large literature has explored the possibility of improving precision in RCTs by pooling the controls in the RCT with historical controls from observational datasets or from other similar RCTs. This literature dates back at least to [22]; for a review, see ref. [23]. Much of this work uses a Bayesian framework, although frequentist approaches exist as well [24]. In many of these methods, biases can be arbitrary large depending on the choice of historical controls. Other recent efforts have sought to improve precision in RCT estimates by using the results of separate models fit on observational data. These include ref. [25], which fits a covariate model to preexperimental data and then uses it to reduce standard errors of online A/B tests; Ref. [26], which uses the RCT to de-bias a broken IV estimate obtained from observational data and then further combines this with an independent RCT-based estimate; and [27], which develops a variant of the sample-splitting estimator that we review below and suggests a role for auxiliary data as well.

Other literature has sought to combine effect estimates from experimental and observational studies, often under the framework of "data fusion" [28]; these methods require observational data on both treated and untreated subjects. In addition to variance reduction, these methods may also seek to generalize the results of RCTs to other populations or other outcome variables, improve the design of RCTs, detect problems in observational studies, or accomplish other goals [29–35]. For recent reviews, see refs [36,37].

A parallel literature in survey methodology discusses the possibility of combining probability and nonprobability samples in order to increase precision, especially for small area estimation [38–41].

In this article, our goal is to estimate the average treatment effect within the RCT, and our focus is on using observational data – nonrandomized subjects in the control or treatment conditions, or both, or neither – to improve the precision of the estimate. The main idea is to use observational data to train an algorithm that predicts RCT outcomes and to use the resulting predictions in the randomized sample as a new covariate. While this approach will work with any covariate adjustment technique, we suggest an approach based on the principal of "first, do no harm," meaning that we prioritize retaining the advantages of randomized experiments highlighted earlier. In particular, we seek to ensure that our method (1) does not introduce any bias, (2) will not harm precision and ideally will improve precision, and (3) does not require any additional statistical assumptions beyond those typically made in design-based analysis of RCTs.

This article is organized as follows. Section 2 reviews background material, including design-based RCT analysis and covariate adjustment. Section 3 discusses incorporating remnant data and presents our main methodological contribution. In Section 4, we apply the method to estimate treatment effects in 33 TestBed experiments. Section 5 concludes this article.

2 Methodological background

2.1 Causal inference from experiments

Consider a randomized experiment to estimate the average effect of a binary treatment T on an outcome Y. There are N subjects, indexed by $i = 1, \ldots, N$. Let $T_i = 1$ if subject i is assigned to treatment, and $T_i = 0$ if control. Let $\mathcal{T} = \{i|T_i = 1\}$ and $C = \{i|T_i = 0\}$, and let $n_t = |\mathcal{T}|$ and $n_c = |C|$.

Following refs [42,43], let potential outcomes y_i^t and y_i^c represent the outcome value Y_i that i would have exhibited if he or she had (perhaps counterfactually) been assigned to treatment or control, respectively.

We model the potential outcomes as fixed (not random). Observed outcomes are a function of treatment assignment and potential outcomes:

$$Y_i = T_i y_i^t + (1 - T_i) y_i^c.$$

Define the treatment effect for i as $\tau_i = y_i^t - y_i^c$. Our goal will be to estimate the average treatment effect (ATE), $\bar{\tau} \equiv \sum_i \tau_i / N = \bar{y}^t - \bar{y}^c$, where $\bar{y}^t = \sum_{i=1}^N y_i^t / N$ is the mean of y^t over all N units in the experiment and $\bar{y}^c$ is defined similarly.

If both y_i^c and y_i^t were known for each subject i, statistical modeling would be unnecessary – researchers could calculate $\bar{\tau}$ exactly, without error, by simply averaging observed τ. In practice, we never observe both y_i^c and y_i^t. Instead, we rely on the experimental setup to estimate and infer causation. Since the treatment and control groups are each random samples of the N participants, survey sampling literature provides design-based unbiased estimators of $\bar{y}^t$ and $\bar{y}^c$ based on observed Y and the known distribution of T. These estimators, and their associated inference, depend only on the experimental design and not on modeling assumptions. The survey sample structure of randomized experiments allows us to infer counterfactual potential outcomes (at least on average) and estimate $\bar{\tau}$ as if τ_i were available for each i, albeit with sampling error.

We will use this framework to analyze the 33 TestBed experiments. These experiments are examples of "Bernoulli experiments," in which each T_i is an independent Bernoulli trial: $\mathbb{P}(T_i = 1) = p$, with $0 < p < 1$, and $T_i \perp\!\!\!\perp T_j$ if $i \neq j$. In the TestBed experiments, $p = 1/2$. Estimation and inference about $\bar{\tau}$ is based on the observed values of Y and T, and the known value of p.

We will now introduce some statistical elements that we will use as the ingredients for our approach. Let $M_i = T_i y_i^c + (1 - T_i) y_i^t$ denote i's unobserved counterfactual outcome – when i is treated, $M_i = y_i^c$ and when i is in the control condition $M_i = y_i^t$. Then i's treatment effect may be expressed as $\tau_i = (-1)^{T_i}(M_i - Y_i)$, i.e., $\tau_i = M_i - Y_i$ if i is in the control group, or $\tau_i = Y_i - M_i$ if i is in the treatment group. Although M_i is, by definition, unobserved, it plays a central role in causal inference; its expectation

$$m_i \equiv \mathbb{E} M_i = p y_i^c + (1 - p) y_i^t$$

will also play a prominent role. Note that m_i is a weighted average of subject i's potential outcomes.

Let

$$U_i = \begin{cases} \dfrac{1}{p} & T_i = 1 \\ -\dfrac{1}{1-p} & T_i = 0 \end{cases}$$

be subject i's signed inverse probability weights; U_i is merely a rescaled treatment indicator. Note that $\mathbb{E} U_i = 0$, and $\mathbb{E} U_i Y_i = \tau_i$. To see the latter, note that when $T_i = 1$, with probability p, $Y_i = y_i^t$ and $U_i Y_i = y_i^t / p$; when $T_i = 0$, with probability $1 - p$, $U_i Y_i = -y_i^c / (1 - p)$. Thus, $U_i Y_i$ may be thought of as an unbiased estimate of τ_i, and $\hat{\tau}^{\text{IPW}} \equiv \sum_i U_i Y_i / N$ is an unbiased estimate of $\bar{\tau}$. Note that $\hat{\tau}^{\text{IPW}}$ is identical to the "Horvitz-Thompson" estimator of ref. [16]

$$\hat{\tau}^{\text{IPW}} = \frac{1}{N} \sum_{i \in \mathcal{T}} \frac{Y_i}{p} - \frac{1}{N} \sum_{i \in C} \frac{Y_i}{1 - p} \tag{1}$$

since it is the difference between the Horvitz-Thomson estimates of $\bar{y}^t$ and $\bar{y}^c$ [44].

The sampling variance of $\hat{\tau}^{\text{IPW}}$ proceeds from the same principals. The variance of $U_i Y_i$ is

$$\mathbb{V}(U_i Y_i) = \left(y_i^t \sqrt{\frac{1-p}{p}} + y_i^c \sqrt{\frac{p}{1-p}} \right)^2 = \frac{m_i^2}{p(1-p)} \tag{2}$$

and $\mathbb{V}(\hat{\tau}^{\text{IPW}}) = \sum_i m_i^2 / [N^2 p(1-p)]$ because treatment assignments are independent. Note that because y_i^t and y_i^c are never simultaneously observed, $\mathbb{V}(\hat{\tau}^{\text{IPW}})$ is not identified. However, $\hat{\mathbb{V}}(\hat{\tau}^{\text{IPW}}) = \sum_i U_i^2 Y_i^2 / N^2$ is an

upper bound, i.e., $\mathbb{E}\hat{\mathbb{V}}(\hat{\tau}^{\text{IPW}}) \geq \mathbb{V}(\hat{\tau}^{\text{IPW}})$. (See ref. [16] for equivalent expressions for more general experimental designs.)

Strangely, $\hat{\tau}^{\text{IPW}}$ and $\mathbb{V}(\hat{\tau}^{\text{IPW}})$ are not translation independent, i.e., adding a constant to each Y changes both the value of $\hat{\tau}^{\text{IPW}}$ and $\mathbb{V}(\hat{\tau}^{\text{IPW}})$ without changing the estimand $\bar{\tau}$. The more popular simple "difference-in-means" estimator [42]

$$\hat{\tau}^{\text{DM}} = \frac{1}{n_t}\sum_{i\in\mathcal{T}} Y_i - \frac{1}{n_c}\sum_{i\in\mathcal{C}} Y_i = \bar{Y}_{\mathcal{T}} - \bar{Y}_{\mathcal{C}} \tag{3}$$

and its associated variance estimator

$$\hat{\mathbb{V}}(\hat{\tau}^{\text{DM}}) = \frac{S^2(Y_{\mathcal{C}})}{n_c} + \frac{S^2(Y_{\mathcal{T}})}{n_t}, \tag{4}$$

where $S^2(Y_{\mathcal{C}}) = \sum_{i\in\mathcal{C}}(Y_i - \bar{Y}_{\mathcal{C}})^2/(n_c - 1)$ is the sample variance of the control group and $S^2(Y_{\mathcal{T}})$ is defined similarly, do not have this undesirable property. Our presentation here focuses on $\hat{\tau}^{\text{IPW}}$ as a jumping-off point for subsequent methodological development, but $\hat{\tau}^{\text{DM}}$ will also play a prominent role.

2.2 Design-based covariate adjustment

The reason for error when estimating τ is our inability to observe counterfactual potential outcomes M. As we have seen, randomized trials, coupled with design-based estimators like $\hat{\tau}^{\text{IPW}}$, use comparison groups and survey sampling theory to implicitly fill in this missing information. Baseline covariates – a vector $\boldsymbol{x}_i$ of data for subject i gathered prior to treatment randomization – may potentially help us improve upon this strategy. Suppose a researcher has constructed algorithms $\hat{y}^c(\cdot)$ and $\hat{y}^t(\cdot)$ designed to impute y^c and y^t, respectively, from $\boldsymbol{x}$. Then $\hat{M}_i = T_i\hat{y}^c(\boldsymbol{x}_i) + (1 - T_i)\hat{y}^t(\boldsymbol{x}_i)$ is an imputation of i's missing counterfactual outcome, and the researcher may estimate τ_i as $(-1)^{T_i}(\hat{M}_i - Y_i)$. In general, the bias of algorithms such as $\hat{y}^c(\cdot)$ and $\hat{y}^t(\cdot)$ will be unknown without further assumptions, so these effect estimates may be inadvisable. On the other hand, imperfect or potentially biased imputations of potential outcomes can, *when combined with randomization*, yield substantial benefits.

The approach we will take to combining covariate adjustment with randomization has antecedents in [2,16–18, 21,45–53], among others. We will focus on exactly unbiased estimators, despite the fact that a small amount of bias in finite sample is often acceptable, especially in the presence of other considerations. In fact, the covariate adjustment techniques we will develop have advantageous properties beyond unbiasedness (see, e.g., Section 4.3.3). That said, our main methodological contributions (in Section 3) are compatible with alternative techniques, including those that may be biased in finite samples. We will frame our arguments around bias since we find it to be the easiest way to formalize confounding, which we see as the most pressing threat to estimators that include observational data.

In a Bernoulli experiment, note that

$$\begin{aligned} U_i(Y_i - m_i) &= \begin{cases} \dfrac{1}{p}(y_i^t - py_i^c - (1-p)y_i^t) & T_i = 1 \\ -\dfrac{1}{1-p}(y_i^c - py_i^c - (1-p)y_i^t) & T_i = 0 \end{cases} \\ &= \begin{cases} \dfrac{p(y_i^t - y_i^c)}{p} & T_i = 1 \\ \dfrac{(1-p)(y_i^t - y_i^c)}{1-p} & T_i = 0 \end{cases} \\ &= \tau_i, \end{aligned}$$

and this therefore suggests using imputations $\hat{y}^c(\boldsymbol{x}_i)$ and $\hat{y}^t(\boldsymbol{x}_i)$ to estimate m_i as $\hat{m}_i = p\hat{y}^c(\boldsymbol{x}_i) + (1-p)\hat{y}^t(\boldsymbol{x}_i)$, and then estimating τ_i as

$$\hat{\tau}_i \equiv U_i(Y_i - \hat{m}_i).$$

For $\hat{\tau}_i$ to be unbiased, it is sufficient that algorithms $\hat{y}^c(\cdot)$ and $\hat{y}^t(\cdot)$ are constructed in such a way that

$$\{\hat{y}^c(\boldsymbol{x}_i), \hat{y}^t(\boldsymbol{x}_i)\} \perp\!\!\!\perp T_i. \tag{5}$$

Since by design the distribution of T_i does not depend on $\boldsymbol{x}_i$, (5) is tantamount to requiring that T_i, and variables such as Y_i that depend on T_i, play no role in constructing algorithms $\hat{y}^c(\cdot)$ and $\hat{y}^t(\cdot)$. Then, under (5),

$$\mathbb{E}(\hat{\tau}_i) = \mathbb{E}(U_iY_i) - \mathbb{E}(U_i\hat{m}_i) = \mathbb{E}(U_iY_i) - \mathbb{E}(U_i)\mathbb{E}(\hat{m}_i) = \mathbb{E}(U_iY_i) = \tau_i,$$

where we use the facts that $\mathbb{E}(U_i) = 0$ and $\mathbb{E}(U_iY_i) = \tau_i$. Finally, define the ATE estimate:

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_i = \frac{1}{N}\sum_{i\in\mathcal{T}}\frac{Y_i - \hat{m}_i}{p} - \frac{1}{N}\sum_{i\in C}\frac{Y_i - \hat{m}_i}{1-p}. \tag{6}$$

The unbiasedness of $\hat{\tau}$ for $\bar{\tau}$ follows from the unbiasedness of each of its summands, $\hat{\tau}_i$ for τ_i.

Crucially, this unbiasedness holds even if $\hat{y}^c(\boldsymbol{x}_i)$ and $\hat{y}^t(\boldsymbol{x}_i)$ are biased; algorithms $\hat{y}^c(\cdot)$ and $\hat{y}^t(\cdot)$ need not be unbiased, consistent, or correct in any sense. As long as $\hat{y}^c(\boldsymbol{x}_i)$ and $\hat{y}^t(\boldsymbol{x}_i)$ are constructed to be independent of T_i, then $\hat{\tau}_i$ will be unbiased. The same cannot be said for regression-based covariate adjustment, the common technique of regressing Y on T and $\boldsymbol{x}$ [54].

The estimate $\hat{\tau}$ given in (6) is identical to the "augmented IPW" (AIPW) estimate familiar from the double-robustness literature in observational studies [48], but with known propensity scores p [55,56]. Though AIPW estimators are typically derived in a model-based framework, the previous results show that in the context of an RCT, provided (5) holds, the AIPW estimator (6) is unbiased under a design-based framework as well.

Compare $\hat{\tau}$ to the estimate $\hat{\tau}^{\text{IPW}}$ given in (1). The only difference is that Y_i in (6) has been replaced by $Y_i - \hat{m}_i$ in (1). The goal of this covariate adjustment is to improve precision – we are residualizing our outcomes, in effect, to reduce variation. Its success in this regard depends on the predictive accuracy of $\hat{y}^c(\boldsymbol{x}_i)$ and $\hat{y}^t(\boldsymbol{x}_i)$. The variance of $\hat{\tau}_i$ depends on $\hat{m}_i$ and is given by

$$\mathbb{V}(\hat{\tau}_i|\hat{m}_i) = \frac{(\hat{m}_i - m_i)^2}{p(1-p)}. \tag{7}$$

Compared with (2), (7) replaces m_i with $\hat{m}_i - m_i$ – that is, replaces potential outcomes with their residuals. Accurate imputations of y_i^c and y_i^t, and hence of m_i, yield precise estimation of τ_i. On the other hand, inaccurate imputations, i.e., when $(\hat{m}_i - m_i)^2$ is greater than m_i^2, will decrease precision – though, again, without causing bias. The sampling variance of the full estimator $\hat{\tau}$ depends on how the parameters of $\hat{y}^c(\cdot)$ and $\hat{y}^t(\cdot)$ are estimated, which may induce dependence between $\hat{\tau}_i$ and $\hat{\tau}_j$ for $i \neq j$. The most important case, for our purposes, is discussed in the next section.

2.3 Sample splitting

Successful covariate adjustment requires imputations $\hat{y}^c(\boldsymbol{x}_i)$ and $\hat{y}^t(\boldsymbol{x}_i)$ that are accurate and independent of T_i. To satisfy the independence condition, i's observed outcome Y_i, which is a function of T_i, cannot play a role in the construction of the algorithms $\hat{y}^c(\cdot)$ and $\hat{y}^t(\cdot)$; they must be trained using other data.

This may be achieved by sample splitting, also referred to in this context as cross-estimation or cross-fitting. In a Bernoulli experiment, rather than fitting global imputation algorithms $\hat{y}^t(\cdot)$ and $\hat{y}^c(\cdot)$ (which would violate 5), fit a separate set of imputation models $\hat{y}^t_{-i}(\cdot)$ and $\hat{y}^c_{-i}(\cdot)$ for each experimental participant i, using data from the other participants. In other words, for each i, one first drops observation i, and then use the remaining $N-1$ observations to construct imputation models for the control and treatment potential

outcomes, denoted by $\hat{y}^c_{-i}(\cdot)$ and $\hat{y}^t_{-i}(\cdot)$, respectively. These models may be fit by any method, for example linear regression or random forests [57] (which, conveniently, automatically provides out-of-bag predictions for each subject). In particular, methods that allow for regularization to prevent overfitting may be used. (For a discussion of sample-splitting for AIPW estimation, see, e.g., refs [18,58,59].)

In this leave-one-out context,

$$\hat{m}_i = p\hat{y}^c_{-i}(\boldsymbol{x}_i) + (1-p)\hat{y}^t_{-i}(\boldsymbol{x}_i),$$

and the estimated average treatment effect is then again given by $\hat{\tau}^{\mathrm{SS}} = \sum_i \hat{\tau}_i / N$ as in (6), and where the superscript denotes "sample splitting." Note that in a Bernoulli experiment $\hat{m}_i \perp\!\!\!\perp T_i$ due to the fact that $\hat{m}_i$ is computed using $\boldsymbol{x}_i$ and a model fit without using observation i. It follows that $\hat{\tau}^{\mathrm{SS}}$ is unbiased. Other randomization designs would call for modifications to the algorithm, e.g. ref. [60].

When we wish to explicitly specify the covariates and imputation method that are used within $\hat{\tau}^{\mathrm{SS}}$, we will write $\hat{\tau}^{\mathrm{SS}}[\,\text{covariates};\text{imputation method}\,]$. For example, if we wished to use random forests and all available covariates, we would write $\hat{\tau}^{\mathrm{SS}}[\boldsymbol{x};\mathrm{RF}\,]$, or if we wished to use only the fourth covariate and ordinary least squares regression, we would write $\hat{\tau}^{\mathrm{SS}}[x_4;\mathrm{LS}\,]$. If we wished to ignore the covariates and always set $\hat{m}_i = 0$, we would write $\hat{\tau}^{\mathrm{SS}}[\varnothing;0]$. Note in particular that $\hat{\tau}^{\mathrm{SS}}[\varnothing;0] = \hat{\tau}^{\mathrm{IPW}}$.

Building upon (7), and following [53], the variance of $\hat{\tau}^{\mathrm{SS}}$ may be estimated as follows. Let

$$\hat{E}_c^{\,2} = \frac{1}{n_c}\sum_{i\in C}[\hat{y}^c_{-i}(\boldsymbol{x}_i) - y_i^c]^2 \tag{8}$$

be the mean-squared-error of control imputations $\hat{y}^c(\boldsymbol{x}_i)$ with respect to potential outcomes y^c, and define $\hat{E}_t^{\,2}$ similarly. Note $\hat{E}_c^{\,2}$ and $\hat{E}_t^{\,2}$ are leave-one-out cross validation mean squared errors. The estimated variance is then given by

$$\hat{\mathbb{V}}(\hat{\tau}^{\mathrm{SS}}) = \frac{1}{N}\left[\frac{p}{1-p}\hat{E}_c^{\,2} + \frac{1-p}{p}\hat{E}_t^{\,2} + 2\sqrt{\hat{E}_c^{\,2}\hat{E}_t^{\,2}}\right]. \tag{9}$$

This variance estimate will typically be somewhat conservative. This is due to the fact that $\mathbb{V}(\hat{\tau}^{\mathrm{SS}})$ is unidentifiable, because the correlation of the potential outcomes is not estimable, and instead an upper bound is used [53]. This difficulty is not unique to $\hat{\tau}^{\mathrm{SS}}$; as noted in Section 2.1, similar comments apply to $\hat{\tau}^{\mathrm{IPW}}$, and the same is true of $\hat{\tau}^{\mathrm{DM}}$ as well [42,61].

Note that by (9),

$$\hat{\mathbb{V}}(\hat{\tau}^{\mathrm{SS}}) \le \frac{\hat{E}_c^{\,2}}{N(1-p)} + \frac{\hat{E}_t^{\,2}}{Np} \approx \frac{\hat{E}_c^{\,2}}{n_c} + \frac{\hat{E}_t^{\,2}}{n_t}, \tag{10}$$

which is similar in form to the variance estimate typically used in a two-sample t-test, namely, $\frac{S^2(Y_C)}{n_c} + \frac{S^2(Y_T)}{n_t}$. In (10), $S^2(Y_C)$ and $S^2(Y_T)$ are replaced by $\hat{E}_c^{\,2}$ and $\hat{E}_t^{\,2}$, respectively. In other words, the sample variances are replaced by the estimated mean squared errors of the imputations.

A special case occurs when the potential outcomes are imputed by simply taking the mean of the observed outcomes (after dropping observation i). That is, we set

$$\hat{y}^c_{-i}(\boldsymbol{x}_i) = \frac{1}{|C\setminus i|}\sum_{j\in C\setminus i} y_j^c \tag{11}$$

and similarly for $\hat{y}^t_{-i}(\boldsymbol{x}_i)$. Note that the covariates are simply ignored, and we denote this special case by $\hat{\tau}^{\mathrm{SS}}[\varnothing;\mathrm{mean}\,]$. It can be shown that $\hat{\tau}^{\mathrm{SS}}[\varnothing;\mathrm{mean}\,] = \hat{\tau}^{\mathrm{DM}}$, i.e., the sample splitting estimator using leave-one-out mean imputation is exactly equal to the simple difference-in-means estimator. Moreover, in this special case, $\hat{E}_c^{\,2} = \frac{n_c}{n_c-1}S^2(Y_C)$ and $\hat{E}_t^{\,2} = \frac{n_t}{n_t-1}S^2(Y_T)$, and thus, the variance estimate given by (10) is nearly identical to the ordinary t-test variance estimate [53].

In short, when using mean imputation for the potential outcomes, the leave-one-out sample splitting procedure essentially simplifies to a standard t-test. The effect estimate is identical, and the variance

estimate is nearly identical.[1] This is highly reassuring. Any imputation strategy that improves upon mean imputation in terms of mean squared error will reduce the variance of $\hat{\tau}^{\mathrm{SS}}$ relative to $\hat{\tau}^{\mathrm{DM}}$. Most modern machine learning methods employ some form of regularization to guard against overfitting, and thus typically perform no worse, or at least not substantially worse, than mean-imputation. Thus, in practice, there is relatively little risk of hurting precision.[2]

3 Incorporating observational data

Modern field trials are often conducted within a very data-rich context, in which rich high-dimensional covariate data are automatically, or already, collected for all experiment participants. For instance, in the TestBed experiments, system administrators have access to log data for every problem and skill builder each participating student worked before the onset of the experiment. In other contexts, such as healthcare or education, rich administrative data are often available. In fact, these covariates are available for a much wider population than just the experimental participants – in the TestBed case, there are log data for all ASSISTments users. In other education or healthcare examples, administrative data are often available for every student or patient in the system, not just for those who were randomized to a treatment or control condition. Often, as in the TestBed case, the outcome variable Y is also drawn from administrative or log data. We refer to subjects within the same data system in which the experiment took place – i.e., for whom covariate and outcome data are available – but who were not part of the experiment, as the "remnant" from the experiment. The remnant from a TestBed experiment consists of all ASSISTments users for whom log data are available but who did not participate in the experiment, of whom there are several hundred thousand.

Simply pooling data from the remnant with data from the experiment undermines the randomization, since students in the remnant were not randomized between conditions. This section will describe an alternative approach – a set of unbiased effect estimators that use the remnant to improve precision. The estimators all begin by using the remnant to fit or train a model predicting potential outcomes as a function of covariates and using that model to impute potential outcomes for units in the experiment. They differ in how they use those imputations and build on each other. The following subsection discusses a simple residualizing estimator, Section 3.2 discusses sample splitting to improve that estimator, and Section 3.3 discusses incorporating an additional set of covariate-adjustment models fit to data from the experimental subjects themselves.

---

1 These statements are conditional on $n_c \geq 2$ and $n_t \geq 2$. When $n_c < 2$, then $S^2(Y_C)$ and the expression in (11) are not defined. When $n_c = 0$, $\bar{Y}_C$ and $\hat{\tau}^{\mathrm{DM}}$ are also undefined. More generally, several of our estimators are undefined when $n_c = 0$, namely, $\hat{\tau}^{\mathrm{DM}}$ defined in (3), $\hat{E}_c^2$ defined in (8), as well as $\hat{\tau}^{\mathrm{RE}}$ in (12) and $\hat{\tau}^{\mathrm{GR}}(b)$ in (14) defined in the next section. Thus, it is worth noting that when we assert $\hat{\tau}^{\mathrm{DM}}$ is unbiased, we implicitly condition on $n_c, n_t > 0$. (It is well known that $\hat{\tau}^{\mathrm{DM}}$ is unbiased conditional on any n_c, n_t, so long as $n_c, n_t > 0$. Without conditioning on n_c and n_t, the moments of $\hat{\tau}^{\mathrm{DM}}$ are undefined in a Bernoulli trial. See, e.g., ref. [62].) The same applies to $\hat{\tau}^{\mathrm{RE}}$ and $\hat{\tau}^{\mathrm{GR}}(b)$ in the next section. For $\hat{\tau}^{\mathrm{SS}}$, we do not implicitly condition $n_c, n_t > 0$ but rather assume that $\hat{m}_i$ is defined for all possible randomizations, including those in which $n_c < 2$ or $n_t < 2$. This may be accomplished, for example, by setting $\hat{m}_i = 0$ in cases where $\hat{y}_{-i}^c(\boldsymbol{x}_i)$ or $\hat{y}_{-i}^t(\boldsymbol{x}_i)$ are otherwise undefined, in which case $\hat{\tau}_i$ reverts to the Horvitz-Thompson estimator. As for $\hat{E}_c^2$ defined in (8), we note that we could alternatively replace the n_c in the denominator with $N(1-p)$, in which case $\hat{E}_c^2$ would be an unbiased estimate of $\frac{1}{N}\sum_{i=1}^{N}\mathrm{MSE}[\hat{y}_{-i}^c(\boldsymbol{x}_i)]$. In practice, we prefer to divide by n_c, although, unlike $\hat{\tau}^{\mathrm{DM}}$, we cannot claim that $\hat{E}_c^2$ is unbiased conditional $n_c, n_t > 0$. See ref. [53].

2 Beyond the question of *hurting* precision, one might reasonably ask – as an anonymous reviewer did – whether, or in what sense, $\hat{\tau}^{\mathrm{SS}}$ is optimal. Since $\hat{\tau}^{\mathrm{SS}}$ is a version of the AIPW estimator, we may refer to the extensive literature on its optimality. For example, ref. [49] gives a set of conditions under which AIPW is efficient or locally efficient, ref. [56] discusses the case of a known propensity score, and refs [18,58] discuss the sample-splitting AIPW estimator. In general, the theoretical literature surrounding AIPW tends take potential outcomes as random, whereas in our development, they are fixed; we defer an examination of the consequences of that distinction for future research.

We will focus on the case in which the treatment condition in the remnant is constant, irrelevant, or just unobserved. For instance, in the TestBed dataset, the RCTs typically test an experimental intervention against "business as usual," and subjects in the remnant were all exposed to the control condition. Extension to cases in which T is observed in the remnant is straightforward and will be discussed briefly in Section 5.

3.1 Covariate adjustment using the remnant

Design based covariate adjustment requires imputation models $\hat{y}^c(\cdot)$ and $\hat{y}^t(\cdot)$; ref. [16] suggests training those models using "auxiliary data" such as the remnant. In the TestBed, there is no basis for separate imputation of y^c and y^t; instead, we use data from the remnant to train an algorithm $\hat{y}^r(\cdot)$ to predict (generic) outcomes as a function of covariates. In some cases, $\hat{y}^r(\cdot)$ may be interpreted as predicting control outcomes, but in other cases, the interpretation may be more opaque.

Regardless of the interpretation, the logic of Section 2.2 would suggest using $\hat{y}^r(\cdot)$ to construct the estimator $\hat{\tau}$ (6), by setting $\hat{m}_i = \hat{y}^r(\boldsymbol{x}_i)$, where $\hat{y}^r(\boldsymbol{x}_i)$, $i = 1, \ldots N$ denotes predictions obtained by applying $\hat{y}^r(\cdot)$ to members of the RCT.[3] This estimator is equivalent to the IPW estimator $\hat{\tau}^{\text{IPW}}$ (1), but with observed outcomes Y replaced by residuals $R_i \equiv Y_i - \hat{y}^r(\boldsymbol{x}_i)$, that is, $\sum_i U_i R_i / N$. Along similar lines, ref. [63] proposes conditioning on n_c and n_t and using a difference in means estimator (also see ref. [25] for a similar suggestion):

$$\hat{\tau}^{\text{RE}} = \frac{1}{n_t} \sum_{i \in \mathcal{T}} R_i - \frac{1}{n_c} \sum_{i \in C} R_i = \bar{R}_{\mathcal{T}} - \bar{R}_C. \tag{12}$$

In what follows, we will refer specifically to (12) as "the remnant estimator."

The remnant estimator $\hat{\tau}^{\text{RE}}$ and its IPW variant work because R_i is itself an outcome variable, with its own potential outcomes $r_i^c = y_i^c - \hat{y}^r(\boldsymbol{x}_i)$ and $r_i^t = y_i^t - \hat{y}^r(\boldsymbol{x}_i)$, and because $\hat{y}^r(\boldsymbol{x}_i)$ is invariant to treatment assignment. Thus, treatment effects on the original outcomes are equal to treatment effects on the residualized outcomes, i.e.,

$$r_i^t - r_i^c = [y_i^t - \hat{y}^r(\boldsymbol{x}_i)] - [y^c - \hat{y}^r(\boldsymbol{x}_i)] = y_i^t - y_i^c = \tau_i,$$

and $\hat{\tau}^{\text{RE}}$ – a difference-in-means estimate of this effect – is therefore an unbiased estimate of $\bar{\tau}$. This property holds regardless of whether $\hat{y}^r(\cdot)$ itself is unbiased, consistent, or "correct" in any sense; indeed, as suggested earlier, it may not even be clear precisely what $\hat{y}^r(\cdot)$ is estimating.

The goal of residualization is to improve precision. Since $\hat{\tau}^{\text{RE}}$ is a difference-in-means estimator, its sampling variance can be conservatively estimated in a similar way as $\hat{\tau}^{\text{DM}}$ (4), but, again, with R replacing Y:

$$\hat{\mathbb{V}}(\hat{\tau}^{\text{RE}}) = \frac{S^2(R_C)}{n_c} + \frac{S^2(R_{\mathcal{T}})}{n_t}. \tag{13}$$

Comparing this expression to $\hat{\mathbb{V}}(\hat{\tau}^{\text{DM}})$ given in (4), we see that the residualized estimator will have a lower variance than $\hat{\tau}^{\text{DM}}$ if $S^2(R_C) < S^2(Y_C)$ and $S^2(R_{\mathcal{T}}) < S^2(Y_{\mathcal{T}})$. In other words, we wish for $\hat{y}^r(\boldsymbol{x})$ to capture at least some of the variation in Y, so that R is less variable than Y. This will be achieved in practice when $\hat{y}^r(\cdot)$ does indeed successfully predict outcomes in the RCT – or, more generally, when the sample covariances between $\hat{y}^r$ and Y for subjects with $T = 0$ and $T = 1$, respectively, are sufficiently large.

Importantly for practitioners, as long as only remnant data are used, $\hat{y}^r(\cdot)$ may be trained and assessed in any way. This process can be iterative, so that an analyst may train a candidate model, assess its performance (perhaps with k-fold cross-validation), modify the algorithm, and repeat until achieving

3 This estimator was also suggested by an anonymous reviewer.

suitable performance. Any modeling approach may be taken, so long as no data from the RCT are used. Postselection inference, which would be a serious concern if model selection were done using the RCT data (especially when the dimension of $\boldsymbol{x}$ is large and the sample size is small), does not apply here.

Unfortunately, in some cases (see, e.g., Section 4), the remnant estimator may have greater sampling variance than the $\hat{\tau}^{\mathrm{DM}}$. This will be the case if $\hat{y}^r(\cdot)$, trained in the remnant, extrapolates poorly to the experimental sample – for instance, if the distribution of Y conditional on $\boldsymbol{x}$ differs substantially between the remnant and the RCT. To make matters worse, the performance of $\hat{y}^r(\cdot)$ in the experimental sample – where it counts – may not be checked directly to select a best model, since when fitting $\hat{y}^r(\cdot)$ outcomes from the RCT cannot be touched.[4]

Thus, residualizing with $\hat{y}^c_{-i}(\boldsymbol{x}_i)$ – i.e., replacing Y with R in an unbiased estimator of $\bar{\tau}$ – will result in an unbiased, design-based estimator that may be substantially more precise than $\hat{\tau}^{\mathrm{DM}}$, but may also be less precise. In other words, covariate adjustment using the remnant in this way is potentially fruitful, but risky.

3.2 Flexibly incorporating Remnant-based imputations

Consider a "generalized remnant estimator"

$$\hat{\tau}^{\mathrm{GR}}(b) \equiv \frac{1}{n_t}\sum_{i\in\mathcal{T}}[Y_i - b\hat{y}^r(\boldsymbol{x}_i)] - \frac{1}{n_c}\sum_{i\in C}[Y_i - b\hat{y}^r(\boldsymbol{x}_i)], \tag{14}$$

where b is some prespecified constant. Note that in the special case $b = 1$, this is the remnant estimator $\hat{\tau}^{\mathrm{RE}}$, and in the special case $b = 0$, it is the simple difference-in-means $\hat{\tau}^{\mathrm{DM}}$. Thus, following the discussion earlier, when $\hat{y}^r(\cdot)$ extrapolates well to the RCT, we wish to set $b = 1$, and when $\hat{y}^r(\cdot)$ extrapolates poorly to the RCT, we wish to set $b = 0$. More typically, an intermediate value for b may be optimal.

The challenge is that we do not know *a priori* how well $\hat{y}^r(\cdot)$ extrapolates to the RCT, and therefore do not know the optimal choice for b. We will use sample splitting to overcome that challenge. First define $x^r \equiv \hat{y}^r(\boldsymbol{x})$. That is, we compute the remnant-based predictions of RCT outcomes as earlier (i.e., $\hat{y}^r(\boldsymbol{x})$), but now regard these predictions simply as a covariate to be used within the sample splitting estimator (i.e., x^r). Then we construct a sample splitting estimator $\hat{\tau}^{\mathrm{SS}}$ using the following imputation method:

$$\begin{aligned}\hat{y}^c_{-i}(x^r_i) &= a^c_{-i} + b^c_{-i}x^r_i\\ \hat{y}^t_{-i}(x^r_i) &= a^t_{-i} + b^t_{-i}x^r_i,\end{aligned} \tag{15}$$

where we obtain a^c_{-i}, b^c_{-i}, a^t_{-i}, and b^t_{-i} by ordinary least squares, i.e., let

$$(a^c_{-i}, b^c_{-i}) = \underset{(a,b)}{\arg\min}\sum_{j\in C\setminus i}[Y_j - (a + bx^r_j)]^2 \tag{16}$$

and similarly for (a^t_{-i}, b^t_{-i}). We denote the resulting estimator $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$.

The estimator $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$ will typically be preferable to the remnant estimator $\hat{\tau}^{\mathrm{RE}}$ because, for each observation i, the remaining $N - 1$ observations of the RCT help determine the best use of x^r_i in constructing $\hat{m}_i$. For example, suppose that the x^r are highly accurate imputations of the y^c in the RCT. In this case, we might expect $a^c_{-i} \approx 0$ and $b^c_{-i} \approx 1$ so that $\hat{y}^c_{-i}(x^r_i) \approx x^r_i$, or in other words, the remnant-based predictions would "pass through" the linear regression largely unmodified, so that $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}] \approx \hat{\tau}^{\mathrm{RE}}$. However, in contrast to the remnant estimator, poor imputations x^r will not necessarily harm precision in $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$. Consider the extreme case in which the x^r are pure noise. We would then expect $a^c_{-i} \approx \bar{Y}_{C\setminus i}$ and $b^c_{-i} \approx 0$ so that $\hat{y}^c_{-i}(x^r_i) \approx \bar{Y}_{C\setminus i}$. That is, we would revert approximately to mean-imputation, so that $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}] \approx \hat{\tau}^{\mathrm{DM}}$. In

4 However, one may use covariate data from the RCT to anticipate $\hat{y}^r(\cdot)$'s performance; Appendix D describes our (unfortunately unsuccessful) attempt to do so. Future research may result in improved methods.

other words, the role of x^r may be tempered according to the prediction accuracy of $\hat{y}^r(\cdot)$ in the RCT. We might therefore expect $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$ to nearly always outperform, or at least perform no worse than, $\hat{\tau}^{\text{RE}}$ and $\hat{\tau}^{\text{DM}}$. This intuition is formalized in the following proposition:

Proposition 1. *Let* $(y_1^c, y_1^t, x_1^r), \ldots, (y_N^c, y_N^t, x_N^r)$ *be IID samples from a population in which* y^c, y^t, *and* x^r *have finite fourth moments, and where* $-1 < \text{corr}(y^c, x^r) < 1$ *and* $-1 < \text{corr}(y^t, x^r) < 1$. *Let* b *be a fixed constant. Let* $\hat{\mathbb{V}}[\hat{\tau}^{\text{GR}}(b)]$ *denote the estimated variance of* $\hat{\tau}^{\text{GR}}(b)$, *defined analogously to* (4) *and* (13). *Let* $\hat{\mathbb{V}}\{\hat{\tau}^{\text{SS}}[x^r, \text{LS}]\}$ *denote the estimated variance of* $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$, *defined as in* (9). *Then as* $N \to \infty$,

$$\frac{\hat{\mathbb{V}}\{\hat{\tau}^{\text{SS}}[x^r, \text{LS}]\}}{\hat{\mathbb{V}}[\hat{\tau}^{\text{GR}}(b)]} \xrightarrow{p} \phi(b) \leq 1$$

where $\phi(b)$ *is some constant that depends on* b.

Proof. See Appendix B. □

Notably, although this proposition is asymptotic in nature, we expect it to be relevant even in relatively small samples, given that $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$ effectively only requires estimating two more parameters than $\hat{\tau}^{\text{GR}}(b)$ (i.e., the slope coefficients b_{-i}^c and b_{-i}^t). The ASSISTments experiments we analyze in Section 4 appear to generally support this intuition; $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$ nearly always outperforms $\hat{\tau}^{\text{DM}}$. Indeed, we see the greatest performance gain in the RCT with the smallest sample size.

Importantly, because the x^r are used only as a covariate, they do not necessarily need to accurately impute the potential outcomes in the RCT; rather, it suffices that they are merely predictive. If the RCT is systematically different from the remnant, e.g., the potential outcomes in the RCT differ in scale from those in the remnant, the x^r will still be useful as long as they are correlated with the experimental potential outcomes. Indeed, counterintuitively, it is even possible for $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$ to achieve precision gains if the x^r are *anticorrelated* with outcomes in the RCT.

In any event, regardless of the properties of $\hat{y}^r(\cdot)$ or quality of the data in the remnant, $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$ remains unbiased, and its associated variance estimator remains conservative, because it relies on $\hat{\tau}^{\text{SS}}$, which has both of those properties, and because x^r is a covariate, and invariant to treatment assignment.

3.3 Combining Remnant-based and within-RCT covariate adjustment

The estimator $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$ effectively solves the remnant estimator's main deficiencies. However, $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$ largely neglects the RCT covariates, except to the extent that x^r depends on $\boldsymbol{x}$ through $\hat{y}^r(\cdot)$. Neglecting the RCT covariate data may be suboptimal, especially when $\hat{y}^r(\cdot)$ is poorly predictive of outcomes in the RCT, perhaps due to systematic differences between the RCT and the remnant. Our goal in this section is to augment the strategy of the previous section, so that the RCT covariate data may be more fully exploited.

Define

$$\tilde{\boldsymbol{x}}_i \equiv (x_{i1}, x_{i2}, \ldots, x_{ip}, x_i^r) \tag{17}$$

or in other words, $\tilde{\boldsymbol{x}}_i$ is $\boldsymbol{x}_i$ augmented with x_i^r. We may now compute $\hat{\tau}^{\text{SS}}$ using the augmented set of covariates $\tilde{\boldsymbol{x}}$ instead of $\boldsymbol{x}$. The hope is that by including x^r we can exploit information in the remnant in much the same way that $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$ does, while simultaneously performing a more standard within-RCT covariate adjustment. For example, we might use random forests and compute $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}, \text{RF}]$.

In general, the precision of the estimator will depend on the performance of the imputation strategy, and in particular, its ability to integrate information from the remnant, via x^r, with information from other covariates $\boldsymbol{x}$. On the one hand, x^r is a function of the other covariates and thus, in at least some sense, does not contain any additional information. However, the function $\hat{y}^r(\cdot)$ is fitted on the remnant, which may be much larger than the experimental sample, and thus $\hat{y}^r(\cdot)$ may be a more accurate imputation function than

what we would be able to obtain using the RCT data alone. In this sense, x^r does contain additional information, which can be exploited by the imputation method by heavily weighting x^r over the other covariates.

On the other hand, if the x^r are highly accurate, using them as a covariate within a nonlinear model like a random forest may be statistically inefficient compared to a linear model, as in $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$. Therefore, it may not always be clear whether a highly flexible method such as $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}, \text{RF}]$ will outperform $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$; it depends on the quality of the imputations x^r as well as the predictive power of the covariates in the experimental sample.

This suggests imputing potential outcomes using a specialized ensemble learner [64]: a weighted average of linear regression using just x_i^r, as in $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$, and random forests using $\tilde{\boldsymbol{x}}$, as in $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}, \text{RF}]$. More specifically, let $\hat{y}_{-i}^{c,\text{LS}}(\tilde{\boldsymbol{x}}_i)$ be the least squares imputation defined in (15) and (16), i.e., the imputation used within $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$; note in particular that $\hat{y}_{-i}^{c,\text{LS}}(\tilde{\boldsymbol{x}}_i)$ ignores all of the entries of $\tilde{\boldsymbol{x}}_i$ except x_i^r. Let $\hat{y}_{-i}^{c,\text{RF}}(\tilde{\boldsymbol{x}}_i)$ denote the imputation from a random forest regression of $Y_{C\setminus i}$ on $\tilde{\boldsymbol{x}}_{C\setminus i}$. We then define an ensemble imputation

$$\hat{y}_{-i}^{c,\text{EN}}(\tilde{\boldsymbol{x}}_i) = \gamma_i^c \hat{y}_{-i}^{c,\text{LS}}(\tilde{\boldsymbol{x}}_i) + (1 - \gamma_i^c)\hat{y}_{-i}^{c,\text{RF}}(\tilde{\boldsymbol{x}}_i), \tag{18}$$

which is an interpolation between $\hat{y}_{-i}^{c,\text{LS}}(\tilde{\boldsymbol{x}}_i)$ and $\hat{y}_{-i}^{c,\text{RF}}(\tilde{\boldsymbol{x}}_i)$, where the interpolation parameter γ_i^c is such that $0 \le \gamma_i^c \le 1$ and is given by

$$\gamma_i^c = \underset{\gamma\in[0,1]}{\arg\min} \sum_{j\in C\setminus i} \{Y_j - [\gamma \hat{y}_{-i,j}^{c,\text{LS}}(\tilde{\boldsymbol{x}}_j) + (1-\gamma)\hat{y}_{-i,j}^{c,\text{RF}}(\tilde{\boldsymbol{x}}_j)]\}^2,$$

where $\hat{y}_{-i,j}^{c,\text{LS}}(\tilde{\boldsymbol{x}}_j)$ is defined analogously to $\hat{y}_{-i}^{c,\text{LS}}(\tilde{\boldsymbol{x}}_i)$, but with both observations i and j removed, and similarly for $\hat{y}_{-i,j}^{c,\text{RF}}(\tilde{\boldsymbol{x}}_j)$. That is, the interpolation parameter γ_i^c is obtained empirically to minimize mean squared error and is obtained from a leave-one-out procedure, which ensures that $\gamma_i^c \perp\!\!\!\perp T_i$, and thus $\hat{y}_{-i}^{c,\text{EN}}(\tilde{\boldsymbol{x}}_i) \perp\!\!\!\perp T_i$. We denote the resulting ensemble-based estimator $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}, \text{EN}]$. The imputation strategy (18) allows $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}, \text{EN}]$ to triangulate between $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$ and $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}, \text{RF}]$ and therefore combines the advantages of both, at the cost of estimating only one additional parameter (i.e., γ_i^c).

4 Estimating effects in 33 online experiments

4.1 Data from the ASSISTments TestBed

We apply and evaluate the methods described in this work to a set of 33 randomized controlled experiments run within the ASSISTments TestBed, described in Section 1. These A/B tests contrast a variety of pedagogical conditions in modules teaching 6th, 7th, and 8th grade mathematics content. For our purposes, the outcome of interest was completion of the module, a binary variable.

In general, once a TestBed proposal is approved, based on Institutional Review Board and content quality criteria, its experimental conditions are embedded into an ASSISTments assignment. This is then assigned to students, either by a group of teachers recruited by the researcher or, more commonly, by the existing population of teachers using ASSISTments in their classrooms. As an example, consider an experiment comparing text-based hints to video hints. The proposing researcher would create the alternative hints and embed them into particular assignable content, a "problem set." Then, any time a teacher assigns that problem set to his or her students, those students are randomized to one of the conditions, and, when they request hints, receive them as either text or video.

There are several types of problem sets that researchers can utilize when developing their experiments. In the case of the 33 experiments observed in this work, the problem sets are mastery-based assignments called skill builders. As opposed to more traditional assignments requiring students to complete all problems assigned, skill builders require students to demonstrate a sufficient level of understanding in order to

complete the assignment. By default, students must simply answer three consecutive problems correctly without the use of computer-provided aid such as hints or scaffolding (a type of aid that breaks the problem into smaller steps). In this way, completion acts as a measure of knowledge and understanding as well as persistence and learning, as students will be continuously given more problems until they are able to reach the completion threshold. ASSISTments also includes a daily limit of ten problems to encourage students to seek help if they are struggling to reach the threshold.

After the completion of a TestBed experiment, the proposing researcher may download a dataset which includes students' treatment assignments and their performance within the skill builder, including an indicator for completion. In addition, the dataset includes aggregated features that describe student performance within the learning platform prior to random assignment for each respective experiment. Summary statistics for the nine covariates we used in our analyses, pooled across experiments, are displayed in Table 1. These include the numbers of problems worked, and assignments and homework assigned, percent of problems correct on first try, assignments completed, and homework completed at the student and class level, and students' genders, as guessed by an internal ASSISTments algorithm based on first names. We imputed missing covariate values separately within each experiment. When possible, we used the mean of observed values from students in the same classroom; otherwise we used the grand mean. We combined this data with disaggregated log data from students' individual prior assignments.

4.2 Imputations from the Remnant

We also gathered analogous data from a large remnant of students who did not participate in any of the 33 experiments we analyzed. Ideally, the remnant would consist of previous ASSISTments students who had worked on the skill builders on which the 33 experiments had been run. If that were the case, we would have considered 33 outcomes of interest, say Y_s, denoting completion of skill builder s. Unfortunately, due to labeling conventions in the ASSISTments database, this was only feasible for 11 of the 33 experiments. Instead, for all 33 experiments, we used prior ASSISTments data to impute one outcome, completion of a generic skill builder.

Rather than use the entire set of past ASSISTments users to build a remnant, we selected students who resembled those who participated in the 33 experiments. For the 11 experiments that we were able to match to other prior work, the remnant consisted of previous students who had worked on at least one of the skill builders in the experiments. For the remaining 22 experiments, we first observed the collection of problem sets given to students in the experiments before being assigned. The remnant consisted of all other ASSISTments users who had been assigned to at least one of those assignments. In other words, the

Table 1: Summary statistics for aggregate prior ASSISTments performance used as within-sample covariates: number of problems worked, and assignments and homework assigned, percent of problems correct on first try, assignments completed, and homework completed at the student and class level, and students' genders, as guessed by ASSISTments based on first names

	Mean	SD	% missing
Problem count	601.13	784.45	2
Percent correct	0.68	0.13	2
Assignments assigned	104.25	413.94	13
Percent completion	0.89	0.21	13
Class percent completion	0.90	0.13	22
Homework assigned	25.97	29.90	50
Homework percent completion	0.93	0.16	59
Class homework percent completion	0.93	0.09	56
Guessed gender	Male: 36%	Female: 36%	Unknown: 28%

remnant consisted of students who did not participate in any of the 33 experiments, but had worked on some of the same content as those who did. In all, the remnant consisted of 141,039 distinct students. Sample sizes and skill builder completion rates in the 33 experiments are given in Table A1.

We gathered records of up to 10 assigned skill builders for each student in the remnant, and for each skill builder recorded the number of problems the student started, completed, requested help on, and answered correctly, the total amount of time spent, and assignment completion (i.e., skill mastery). Then, we fit a type of recurrent neural network [65] called long short-term memory (LSTM) [66] to the resulting panel data. The model attempts to detect within-student trends in assignment completion and speed (i.e., the number of problems needed for skill mastery); please see Appendix C for further details. By using 10-fold cross validation within the remnant, we estimated the area under the ROC curve as 0.82 and a root mean squared error of 0.34 for the dependent measure of next assignment completion.

After fitting and validating the model in the remnant, we used it to predict skill builder completion for each subject in each of the 33 experiments. To do so, we gathered log data for each student from up to ten previous assigned skill builders. (Students in the experiments with no prior data were dropped from all analyses.) By using the model fit in the remnant, we predicted whether each student would complete his or her next assigned skill builder. The resulting predictive probabilities were used as x^r in the following analyses.

4.3 Results

In each of the 33 experiments, we calculated five different unbiased ATE estimates: (1) the simple difference-in-means estimator $\hat{\tau}^{\mathrm{DM}}$ (equation (3)); (2) the remnant estimator $\hat{\tau}^{\mathrm{RE}}$ (equation (12)); (3) $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$ (Section 3.2); (4) $\hat{\tau}^{\mathrm{SS}}[\boldsymbol{x};\mathrm{RF}]$ (Section 2.3) where $\boldsymbol{x}$ denotes only those covariates supplied within the TestBed, as listed in Table 1; and (5) $\hat{\tau}^{\mathrm{SS}}[\tilde{\boldsymbol{x}}, \mathrm{EN}]$ (Section 3.3), using both x^r and the provided TestBed covariates $\boldsymbol{x}$. These five methods are all design-based and unbiased, but they differ in their adjustment for covariates – both in the data they use for the adjustment and in how the adjustment is effected. Notably, in this application, the remnant-based predictions x^r are not functions only of $\boldsymbol{x}$. The covariates in $\boldsymbol{x}$ are limited to aggregated data that summarize a student's previous performance (Table 1), whereas the predictions x^r are based on a more fine-grained longitudinal analysis of each student's log data.

Since each of these estimates is unbiased, we will focus on their estimated sampling variances. To aid interpretability, we will express contrasts between the sampling variances of two methods in terms of sample size. The estimated sampling variance of each estimator we consider is inversely proportional to sample size (see, e.g., equation (9)). Therefore, reducing the sampling variance of an estimator by, say, 1/2 is equivalent to doubling its sample size. Under that reasoning, the following discussion will refer to the ratio of estimated sampling variances as a "sample size multiplier."

4.3.1 Remnant-based adjustment: comparing $\hat{\tau}^{\mathrm{RE}}$ and $\hat{\tau}^{\mathrm{SS}}[\boldsymbol{x}^r, \mathrm{LS}]$

Figure 1 compares $\hat{\tau}^{\mathrm{DM}}$, $\hat{\tau}^{\mathrm{RE}}$, and $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$ on the 33 ASSISTments TestBed experiments. Each dot in the figure corresponds to a sample size multiplier comparing two estimated sampling variances in a particular experiment. The vertical line at 1.0 indicates experiments in which the two methods gave approximately equal sampling variances. Dots to the right of the line correspond to experiments in which the variance in the denominator of the fraction was lower, and dots to the left of the line correspond to experiments in which the variance in the numerator was lower.

The leftmost plot contrasts $\hat{\tau}^{\mathrm{RE}}$ with $\hat{\tau}^{\mathrm{DM}}$. In four experiments, the variances of $\hat{\tau}^{\mathrm{RE}}$ and $\hat{\tau}^{\mathrm{DM}}$ were approximately equal, and in 27 experiments, $\hat{\tau}^{\mathrm{RE}}$ outperformed $\hat{\tau}^{\mathrm{DM}}$. Notably, in one case (experiment 33), the adjustment provided by $\hat{\tau}^{\mathrm{RE}}$ was equivalent to a roughly 85% increase in the sample size, and in another (experiment #27), the adjustment was equivalent to a roughly 50% increase. On the whole, $\hat{\tau}^{\mathrm{RE}}$ offers substantial gains in precision relative to $\hat{\tau}^{\mathrm{DM}}$. On the other hand, in two experiments, the sampling

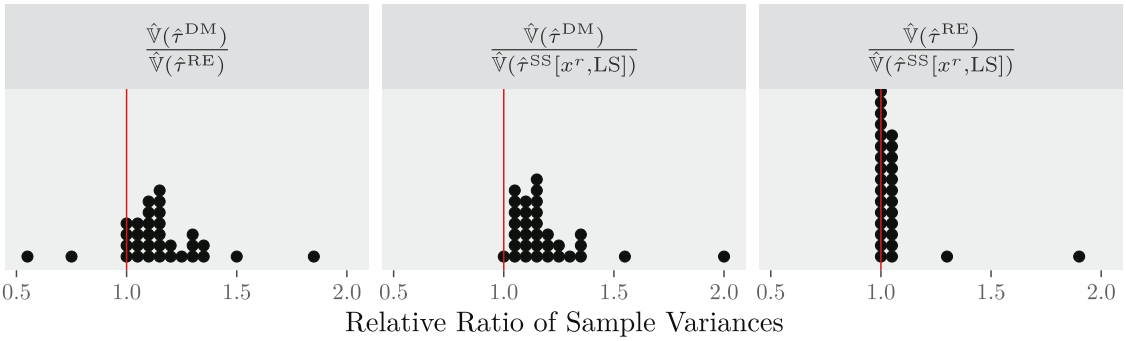

Figure 1: A dotplot showing sample size multipliers (i.e., sampling variance ratios) comparing $\hat{\tau}^{\mathrm{DM}}$, $\hat{\tau}^{\mathrm{RE}}$, and $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$ on the 33 ASSISTments TestBed experiments.

variance of $\hat{\tau}^{\mathrm{RE}}$ was higher than that of $\hat{\tau}^{\mathrm{DM}}$. Most notably, in one experiment (#2), the adjustment given by $\hat{\tau}^{\mathrm{RE}}$ was equivalent to a roughly 45% *decrease* in sample size. In this case, apparently, the imputations from the model fit to the remnant were particularly inaccurate in the experimental sample. Because experimental outcomes played no role in determining the adjustment provided by $\hat{\tau}^{\mathrm{RE}}$, the adjustment was blind to this inaccuracy, and was unable to anticipate the resulting increase in variance in those cases.

In contrast, the $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$ estimator incorporates information on imputation accuracy into its covariate adjustment. The middle panel of Figure 1 shows that across the board, $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$ variances were smaller or roughly equal to those of $\hat{\tau}^{\mathrm{DM}}$. That is, $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$ successfully avoided the risk that poor imputations pose to $\hat{\tau}^{\mathrm{RE}}$, and never increased variance relative to $\hat{\tau}^{\mathrm{DM}}$. Moreover, in those cases in which $\hat{\tau}^{\mathrm{RE}}$ performed well, $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$ tended to perform even better. For instance, in experiment 33, the adjustment provided by $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$ was equivalent to a roughly 100% increase in sample size (relative to $\hat{\tau}^{\mathrm{DM}}$).

The rightmost panel of Figure 1 compares $\hat{\tau}^{\mathrm{RE}}$ to $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$ explicitly: $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$ sample variances dominated those of $\hat{\tau}^{\mathrm{RE}}$. In roughly half of the experiments, $\hat{\tau}^{\mathrm{RE}}$ and $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$ performed similarly, and in the remaining half, $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$ improved upon $\hat{\tau}^{\mathrm{RE}}$. Proposition 1, guarantees that $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$ will dominate both $\hat{\tau}^{\mathrm{DM}}$ and $\hat{\tau}^{\mathrm{RE}}$ in the limit as $N \to \infty$; Figure 1 gives examples of this property in finite samples.

4.3.2 Incorporating standard covariates

Figure 2 compares $\hat{\tau}^{\mathrm{SS}}[\tilde{\boldsymbol{x}}, \mathrm{EN}]$ to $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$, $\hat{\tau}^{\mathrm{SS}}[\boldsymbol{x};\mathrm{RF}]$, and $\hat{\tau}^{\mathrm{DM}}$, on the same 33 ASSISTments TestBed experiments. The left panel, comparing $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$ to $\hat{\tau}^{\mathrm{SS}}[\tilde{\boldsymbol{x}}, \mathrm{EN}]$, shows the impact of including standard covariates, incorporating them as described in the ensemble approach of Section 3.3. In all but one case, the sampling variance of $\hat{\tau}^{\mathrm{SS}}[\tilde{\boldsymbol{x}}, \mathrm{EN}]$ was less than or roughly equal to that of $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$ – that is, including

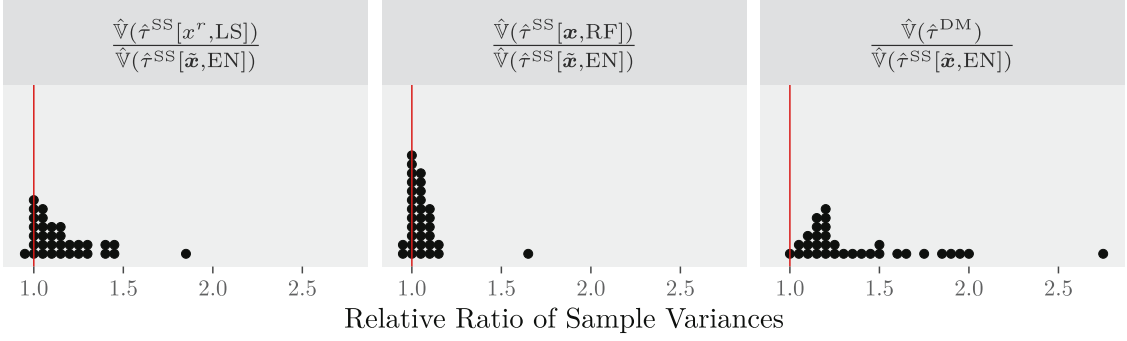

Figure 2: A dotplot showing sample size multipliers (i.e., sampling variance ratios) comparing $\hat{\tau}^{\mathrm{SS}}[\tilde{\boldsymbol{x}}, \mathrm{EN}]$ to $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$, $\hat{\tau}^{\mathrm{SS}}[\boldsymbol{x};\mathrm{RF}]$, and $\hat{\tau}^{\mathrm{DM}}$, respectively, on the 33 ASSISTments TestBed experiments.

standard covariates improved precision. In 16 cases, this improvement was equivalent to increasing the sample size by more than 10%; in eight of those cases, the improvement was more than 25% and in one experiment, it was more than 80%.

The middle panel compares the sampling variances of $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}, \text{EN}]$ and $\hat{\tau}^{\text{SS}}[\boldsymbol{x};\text{RF}]$, showing the extent to which including x^r improved precision relative to using only standard covariates. In all but two experiments, the sampling variance of $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}, \text{EN}]$ was less than or roughly equal to the sampling variance of $\hat{\tau}^{\text{SS}}[\boldsymbol{x};\text{RF}]$. In six experiments, the improvement was equivalent to an increase in sample size of more than 10%, and in one of those cases, experiment 33, the improvement was equivalent to an over 65% increase in the sample size.

The rightmost panel compares the sampling variances of $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}, \text{EN}]$ and the simple difference-in-means estimator, showing the total impact of covariate adjustment on statistical precision. Across every one of the 33 experiments, the estimated sampling variances for $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}, \text{EN}]$ were lower or roughly equal to those of $\hat{\tau}^{\text{DM}}$. In 28 experiments, the improvement was equivalent to increasing the sample size by more than 10%; in 15 of those, the improvement was equivalent to a more than 25% increase in the sample size, and in the case of experiment 33, the improvement was equivalent to a 175% increase in the sample size.

4.3.3 Covariate adjustment with ANCOVA

The methodological development in Section 3 focused on the covariate-adjusted estimator $\hat{\tau}^{\text{SS}}$, which can incorporate nearly any imputation method – including least squares regression, random forests, and ensemble methods such as (18) – while maintaining the advantages of design-based estimation, namely unbiased effect estimation and conservative standard error estimation. However, the strategy of covariate adjustment using x^r or $\tilde{\boldsymbol{x}}$ is compatible with any covariate-adjusted estimator. For instance, an anonymous reviewer suggested estimating $\bar{\tau}$ via ANCOVA – that is, fitting the model

$$Y_i = \mu + \beta T_i + \boldsymbol{\gamma}^T \boldsymbol{X}_i + \varepsilon_i \tag{19}$$

with ordinary least squares, where $\boldsymbol{X}_i = x_i^r$ or $\tilde{\boldsymbol{x}}_i$ and estimating $\bar{\tau}$ with the estimated coefficient $\hat{\beta}$, which we will denote as $\hat{\beta}[x^r]$ or $\hat{\beta}[\tilde{\boldsymbol{x}}]$, respectively (also see [67]). ANCOVA estimators $\hat{\beta}[\cdot]$ are typically slightly biased, but consistent, with bias decreasing with $1/N$ [54].

Figure 3 compares the estimated sampling variances of $\hat{\tau}^{\text{DM}}$, $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$, $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}, \text{EN}]$, $\hat{\beta}[x^r]$, and $\hat{\beta}[\tilde{\boldsymbol{x}}]$ when applied to the 33 TestBed experiments. (The ANCOVA standard errors were estimated using the HC2 sandwich formula [68], the default for the `lm_robust()` routine of the `estimatr` package in `R` [69,70].) For the sake of comparison, the top panels of Figure 3 reproduce results from Figures 1 and 2, comparing $\hat{\tau}^{\text{DM}}$ to $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$ and $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}, \text{EN}]$. The middle two panels contrast the sampling variances of $\hat{\beta}[x^r]$ and $\hat{\beta}[\tilde{\boldsymbol{x}}]$ to $\hat{\tau}^{\text{DM}}$. Like $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$ and $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}, \text{EN}]$, the ANCOVA estimates are, in many cases, much more precise than $\hat{\tau}^{\text{DM}}$. On the other hand, in some cases, $\hat{\beta}[\tilde{\boldsymbol{x}}]$ had a noticeably higher sampling variance than $\hat{\tau}^{\text{DM}}$ – in one case, the effect of ANCOVA adjustment was roughly equivalent to reducing the sample size by about 15%.

Across the board, the precision gains afforded by $\hat{\beta}[\tilde{\boldsymbol{x}}]$ were typically slightly less than those afforded by $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}, \text{EN}]$. This is displayed in the bottom row of Figure 3, which compares the ANCOVA estimators directly to $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$ and $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}, \text{EN}]$. While $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$ and $\hat{\beta}[x^r]$ tend to have very similar sampling variances, $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}, \text{EN}]$ is often (but not always) much more precise than $\hat{\beta}[\tilde{\boldsymbol{x}}]$. Presumably, this advantage is due to the flexibility of the ensemble learner in $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}, \text{EN}]$, which is in contrast to the linear additive adjustment of ANCOVA.

5 Discussion

Randomized experiments and observational studies have complementary strengths. Randomized experiments allow for unbiased estimates with minimal statistical assumptions, but often suffer from small

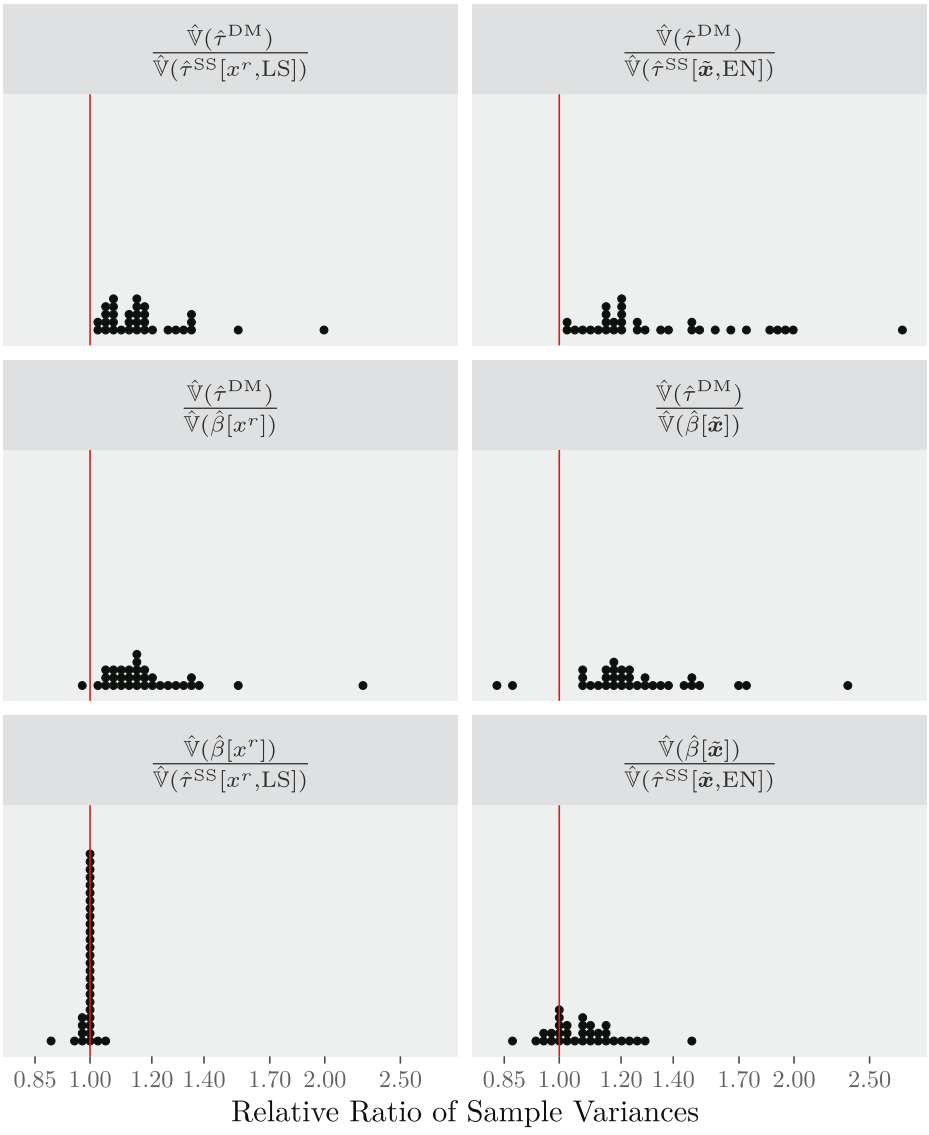

Figure 3: A dotplot showing sample size multipliers (i.e., sampling variance ratios), from contrasts between the difference-in means estimator $\hat{\tau}^{\mathrm{DM}}$, sample-splitting estimators $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$ and $\hat{\tau}^{\mathrm{SS}}[\tilde{\boldsymbol{x}}, \mathrm{EN}]$, and ANCOVA estimators $\hat{\boldsymbol{\beta}}[x^r]$ and $\hat{\boldsymbol{\beta}}[\tilde{\boldsymbol{x}}]$ with HC2 standard errors, on the 33 ASSISTments TestBed experiments.

sample sizes. Observational studies, by contrast, may offer huge sample sizes, but typically suffer from confounding biases, which must be adjusted for, often through statistical modeling with questionable assumptions. In this article, we have attempted to combine the strengths of both. More specifically, we have sought to improve the precision of randomized experiments by exploiting the rich information available in a large observational dataset.

Our approach may be summarized as "first, do no harm." A randomized experiment may be analyzed by taking a simple difference in means, which on its own provides a valid design-based unbiased estimate. The rationale for a more complicated analysis would be to improve precision. Our goal has therefore been to ensure that, in attempting to improve precision by incorporating observational data, we have not actually made matters worse. In particular, we have sought to ensure that (1) no biases in the observational data may "leak" into the analysis, (2) we can reasonably expect to improve precision, not harm it, and (3) inference may be justified by the experimental randomization, without the need for additional statistical modeling assumptions.

In this article, we focused on covariate adjustment using $\hat{\tau}^{\text{SS}}$, which is exactly unbiased; if a different covariate adjustment method were used instead of $\hat{\tau}^{\text{SS}}$, such as those proposed by [71] or [72], then the resulting estimator would inherit its properties, instead. We focus on the sample splitting estimator for two reasons. First, because we believe that a guarantee of exact unbiasedness will remove barriers to the method's adoption. Incorporating observational data into the analysis of RCTs may appear to be inherently risky, or to undermine the rationale for randomization. A general guarantee that effect estimates are unbiased, even in finite samples, may alleviate those concerns. Second, $\hat{\tau}^{\text{SS}}$ is compatible with nearly any imputation algorithm, and this flexibility may be especially valuable when incorporating x^r. The analysis in Section 4.3.3 provides a nice illustration of this: while there is little difference between the standard errors of $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$ and analogous ANCOVA estimates, $\hat{\tau}^{\text{SS}}[\tilde{\boldsymbol{x}}, \text{EN}]$ – which uses an ensemble imputation algorithm including random forests – tended to perform substantially better than an ANCOVA estimator using the same data.

The results from the 33 A/B tests we analyzed suggest that incorporating information gleaned from the remnant of an experiment can indeed improve causal inference – but it does not always do so. The extent to which the remnant can help improve precision depends on the quality of the remnant-based predictions, and this in turn depends on both the quality of the remnant data and the algorithm $\hat{y}^r(\cdot)$. It is therefore important to include observational data judiciously – our methods dynamically adapt, taking advantage of observational data when it is useful and minimizing its role when it isn't.

The focus of this article was to show that these methods can improve statistical precision without incurring a statistical cost – i.e., without potentially increasing bias or standard errors. However, gathering remnant data and using it to train an algorithm may require substantial human and/or computational resources. Therefore, it is crucial for applied researchers to be able to anticipate in advance the extent to which our methods will outperform estimators that use only RCT data. These cost–benefit calculations can take place at two different points in the research process: before collecting any remnant data, and after collecting data from the remnant but before using it to train a predictive algorithm. Before collecting data from the remnant, researchers may be able to use observed properties of RCT data, along with anticipated, but yet unobserved, properties of the remnant to decide whether to proceed. For instance, some initial empirical results, currently under review, suggest that our methods have the potential to improve statistical precision across a wide range of RCT sample sizes, but that the most dramatic improvements tend to occur when the RCT sample size is small or moderate. Intuition suggests that the greatest contribution of auxiliary data will occur when a large number of covariates are available, but there is little prior information on which covariates are the most important. If remnant data are available, analysts may decide whether to use them to train a predictive algorithm based on explicit comparisons between covariate distributions in the remnant and in the RCT (Appendix D). Intuition suggests that our methods hold the greatest promise when covariates in the remnant and RCT are most similar.

These, and other questions will be best answered by applying our methods in a wide variety of contexts. While we have focused on the ASSISTments platform in this article, the future work will explore what other sources of auxiliary data, and corresponding prediction algorithms, may be particularly well suited to improving the precision of RCTs typically encountered in education research. Indeed, one of the advantages of developing models on observational data in this manner is that a wide variety of models may be explored, tested, and iteratively improved upon before they are applied to an RCT.

In particular, it will be interesting to consider cases in which the experimental condition varies – and is recorded – in the remnant. For instance, the remnant from an RCT contrasting two common medical procedures may include medical records from previous patients who underwent one or the other procedure. In that case, analysts may train remnant models to impute both potential outcomes as, say, $\hat{y}^{rc}(\boldsymbol{x})$ and $\hat{y}^{rt}(\boldsymbol{x})$. Then (following Section 3.1) they may set $\hat{m}_i = p\hat{y}^{rc}(\boldsymbol{x}_i) + (1-p)\hat{y}^{rt}(\boldsymbol{x}_i)$ and estimate average treatment effects using $\hat{\tau}$, or (following Sections 3.2 and 3.3), include $\hat{y}^{rc}(\boldsymbol{x})$ and $\hat{y}^{rt}(\boldsymbol{x})$ within a sample-splitting estimator $\hat{\tau}^{\text{SS}}$, perhaps alongside other covariates. We expect that the inclusion of both types of exposures in the remnant may enhance remnant-based estimators even further, and hope to explore these possibilities in future research.

Acknowledgements: We would like to thank Ben Hansen and Charlotte Mann for helpful discussions and Ethan Prihar for help with computation. We would also like to thank the two anonymous reviewers for their comments.

Funding information: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D210031. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. E. Wu was supported by NSF RTG grant DMS-1646108. N. Heffernan oversaw the creation of the 33 experiments and provided the data from ASSISTments; we want to acknowledge the funding that created/related to ASSISTments from (1) NSF (e.g., 2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, and 1535428), (2) IES (e.g., R305N210049, R305D210031, R305A170137, R305A170243, R305A180401, R305D210036, R305A120125, and R305R220012), (3) GAANN (e.g., P200A180088 and P200A150306), (4) EIR (U411B190024 and S411B210024), (5) ONR (N00014-18-1-2768), and (6) Schmidt Futures. None of the opinions expressed here are those of the funders.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: Code and data are available at https://osf.io/d9ujq/.

References

[1] Schochet PZ. Statistical theory for the RCT-YES software: Design-based causal inference for RCTs. NCEE 2015-4011. Washington, D.C.: National Center for Education Evaluation and Regional Assistance; 2015.
[2] Rosenbaum PR. Covariance adjustment in randomized experiments and observational studies. Stat Sci. 2002;17(3):286–327.
[3] Sales AC, Hansen BB, Rowan B. Rebar: Reinforcing a matching estimator with predictions from high-dimensional covariates. J Educ Behav Stat. 2018;43(1):3–31.
[4] Heffernan NT, Heffernan CL. The ASSISTments ecosystem: building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. Int J Artif Intell Educ. 2014;24(4):470–97.
[5] Ostrow KS, Selent D, Wang Y, Van Inwegen EG, Heffernan NT, Williams JJ. The assessment of learning infrastructure (ALI): the theory, practice, and scalability of automated assessment. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. ACM; 2016. p. 279–88.
[6] Fyfe ER. Providing feedback on computer-based algebra homework in middle-school classrooms. Comput Human Behav. 2016;63:568–74.
[7] Walkington C, Clinton V, Sparks A. The effect of language modification of mathematics story problems on problem-solving in online homework. Instruct Sci. 2019;47:1–31.
[8] Prihar E, Syed M, Ostrow K, Shaw S, Sales A, Heffernan N. Exploring common trends in online educational experiments. In: Proceedings of the 15th International Conference on Educational Data Mining; 2022. p. 27.
[9] Vanacore K, Gurung A, Mcreynolds A, Liu A, Shaw S, Heffernan N. Impact of non-cognitive interventions on student learning behaviors and outcomes: an analysis of seven large-scale experimental inventions. In: LAK23: 13th International Learning Analytics and Knowledge Conference. LAK2023. New York, NY, USA: Association for Computing Machinery; 2023. p. 165–74. doi: 10.1145/3576050.3576073.
[10] Gurung A, Baral S, Vanacore KP, Mcreynolds AA, Kreisberg H, Botelho AF, et al. Identification, exploration, and remediation: can teachers predict common wrong answers? In: LAK23: 13th International Learning Analytics and Knowledge Conference. LAK2023. New York, NY, USA: Association for Computing Machinery; 2023. p. 399–410. doi: 10.1145/3576050.3576109.
[11] Gurung A, Vanacore KP, McReynolds AA, Ostrow KS, Sales AC, Heffernan N. How common are common wrong answers? exploring remediation at scale. In: Proceedings of the Tenth ACM Conference on Learning@ Scale (L@S'23). New York, NY, USA: ACM; 2023.
[12] Selent D, Patikorn T, Heffernan N. Assistments dataset from multiple randomized controlled experiments. In: Proceedings of the Third (2016) ACM Conference on Learning@ Scale. ACM; 2016. p. 181–4.
[13] Diamond A, Sekhon JS. Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. Rev Econom Stat. 2013;95(3):932–45.
[14] Künzel SR, Stadie BC, Vemuri N, Ramakrishnan V, Sekhon JS, Abbeel P. Transfer learning for estimating causal effects using neural networks. INFORMS. 2019.

[15] Rzepakowski P, Jaroszewicz S. Decision trees for uplift modeling with single and multiple treatments. Knowledge Inform Syst. 2012;32(2):303–27.
[16] Aronow PM, Middleton JA. A class of unbiased estimators of the average treatment effect in randomized experiments. J Causal Inference. 2013;1(1):135–54.
[17] Wager S, Du W, Taylor J, Tibshirani RJ. High-dimensional regression adjustments in randomized experiments. Proc Natl Academy Sci. 2016;113(45):12673–8.
[18] Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double/debiased machine learning for treatment and structural parameters. Econometrics J. 2018;21(1):C1–68.
[19] Bloniarz A, Liu H, Zhang CH, Sekhon JS, Yu B. Lasso adjustments of treatment effect estimates in randomized experiments. Proc Natl Acad Sci. 2016;113(27):7383–90.
[20] Rosenblum M, Van Der Laan MJ. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. Int J Biostat. 2010;6(1). https://doi.org/10.2202/1557-4679.1138.
[21] Van der Laan MJ, Rose S. Targeted learning: causal inference for observational and experimental data. New York: Springer Science & Business Media; 2011.
[22] Pocock SJ. The combination of randomized and historical controls in clinical trials. J Chronic Diseases. 1976;29(3):175–88.
[23] Viele K, Berry S, Neuenschwander B, Amzal B, Chen F, Enas N, et al. Use of historical control data for assessing treatment effects in clinical trials. Pharmaceut Stat. 2014;13(1):41–54.
[24] Yuan J, Liu J, Zhu R, Lu Y, Palm U. Design of randomized controlled confirmatory trials using historical control data to augment sample size for concurrent controls. J Biopharmaceut Stat. 2019;29(3):558–73.
[25] Deng A, Xu Y, Kohavi R, Walker T. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining; 2013. p. 123–32.
[26] Gui G. Combining observational and experimental data using first-stage covariates. 2020. arXiv: http://arXiv.org/abs/arXiv:201005117.
[27] Opper IM. Improving average treatment effect estimates in small-scale randomized controlled trials. EdWorkingPapers. 2021. https://edworkingpapers.org/sites/default/files/ai21-344.pdf.
[28] Bareinboim E, Pearl J. Causal inference and the data-fusion problem. Proce Natl Acad Sci. 2016;113(27):7345–52.
[29] Hartman E, Grieve R, Ramsahai R, Sekhon JS. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. J R Stat Soc Ser A. 2015;10:1111.
[30] Athey S, Chetty R, Imbens G. Combining experimental and observational data to estimate treatment effects on long term outcomes. 2020. http://arXiv.org/abs/arXiv:200609676.
[31] Rosenman ET, Owen AB. Designing experiments informed by observational studies. J Causal Inference. 2021;9(1):147–71.
[32] Rosenman ET, Basse G, Owen AB, Baiocchi M. Combining observational and experimental datasets using shrinkage estimators. Biometrics. 2020;1–13. https://doi.org/10.1111/biom.13827.
[33] Rosenman ET, Owen AB, Baiocchi M, Banack HR. Propensity score methods for merging observational and experimental datasets. Stat Med. 2022;41(1):65–86.
[34] Chen S, Zhang B, Ye T.Minimax rates and adaptivity in combining experimental and observational data. 2021. http://arXiv.org/abs/arXiv:210910522.
[35] Kallus N, Puli AM, Shalit U. Removing hidden confounding by experimental grounding. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in neural information processing systems. Vol. 31. Curran Associates, Inc.; 2018. p. 10888–97.
[36] Degtiar I, Rose S. A review of generalizability and transportability. Annual Rev Stat Appl. 2023;10:501–24.
[37] Colnet B, Mayer I, Chen G, Dieng A, Li R, Varoquaux G, et al. Causal inference methods for combining randomized trials and observational studies: a review. 2020. arXiv: http://arXiv.org/abs/arXiv:201108047.
[38] Breidt FJ, Opsomer JD. Model-assisted survey estimation with modern prediction techniques. Stat Sci. 2017;32(2):190–205.
[39] Erciulescu AL, Cruze NB, Nandram B. Statistical challenges in combining survey and auxiliary data to produce official statistics. J Official Stat (JOS). 2020;36(1):63–88.
[40] Dagdoug M, Goga C, Haziza D. Model-assisted estimation through random forests in finite population sampling. J Amer Stat Assoc. 2021;118:1234–51.
[41] McConville KS, Moisen GG, Frescino TS. A tutorial on model-assisted estimation with application to forest inventory. Forests. 2020;11(2):244.
[42] Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Stat Sci. 1923;5:463–80. 1990; transl. by D.M. Dabrowska and T.P. Speed.
[43] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688.
[44] Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Amer Stat Assoc. 1952;47(260):663–85.
[45] Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Amer Stat Assoc. 1994;89(427):846–66.
[46] Scharfstein DO, Rotnitzky A, Robins JM. Rejoinder. J Amer Stat Assoc. 1999;94(448):1135–46.

[47] Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. In: Proceedings of the American Statistical Association. vol. 1999. Indianapolis, IN; 2000. p. 6–10.
[48] Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005;61(4):962–73.
[49] van der Laan MJ, Rubin D. Targeted maximum likelihood learning. Int J Biostat. 2006;2(1). https://doi.org/10.2202/1557-4679.1043.
[50] Tsiatis AA, Davidian M, Zhang M, Lu X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. Stat Med. 2008;27(23):4658–77.
[51] Moore KL, van der Laan MJ. Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. Stat Med. 2009;28(1):39–64.
[52] Belloni A, Chernozhukov V, Hansen C. Inference on treatment effects after selection among high-dimensional controls. Rev Econom Stud. 2014;81(2):608–50.
[53] Wu E, Gagnon-Bartsch JA. The LOOP estimator: adjusting for covariates in randomized experiments. Evaluat. Rev. 2018;42(4):458–88.
[54] Freedman DA. On regression adjustments to experimental data. Adv Appl Math. 2008;40(2):180–93.
[55] Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica. 1998;66:315–31.
[56] Rothe C. The value of knowing the propensity score for estimating average treatment effects. IZA Discussion Papers. 2016. (9989).
[57] Breiman L. Random forests. Machine Learn. 2001;45(1):5–32.
[58] Jiang K, Mukherjee R, Sen S, Sur P. A new central limit theorem for the augmented IPW estimator: variance inflation, cross-fit covariance and beyond. 2022. arXiv: http://arXiv.org/abs/arXiv:220510198.
[59] Smucler E, Rotnitzky A, Robins JM. *A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts*. 2019. arXiv: http://arXiv.org/abs/arXiv:19040373.
[60] Wu E, Gagnon-Bartsch JA. Design-based covariate adjustments in paired experiments. J Educ Behav Stat. 2021;46(1):109–32.
[61] Aronow PM, Green DP, Lee DKK. Sharp bounds on the variance in randomized experiments. Ann Statist. 2014;42(3):850–71.
[62] Freedman D, Pisani R, Purves R, Adhikari A. Statistics. New York: WW Norton & Company; 2007.
[63] Sales AC, Botelho A, Patikorn TM, Heffernan NT. Using big data to sharpen design-based inference in A/B tests. In: Proceedings of the 11th International Conference on Educational Data Mining. International Educational Data Mining Society; 2018. p. 479–86.
[64] Opitz D, Maclin R. Popular ensemble methods: an empirical study. J Artif Intell Res. 1999;11:169–98.
[65] Williams RJ, Zipser D. A learning algorithm for continually running fully recurrent neural networks. Neural Comput. 1989;1(2):270–80.
[66] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
[67] Walsh D, Miller D, Hall D, Walsh J, Fisher C, Schuler A. Prognostic covariate adjustment: a novel method to reduce trial sample sizes while controlling type I error; 2022. Talk presented at the Joint Statistical Meetings. https://ww2.amstat.org/meetings/jsm/2022/onlineprogram/AbstractDetails.cfm?abstractid=320608.
[68] MacKinnon JG, White H. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. J Econom. 1985;29(3):305–25.
[69] Blair G, Cooper J, Coppock A, Humphreys M, Sonnet L. Estimatr: fast estimators for design-based inference; 2021. R package version 0.30.2. https://CRAN.R-project.org/package=estimatr.
[70] R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria; 2011. ISBN 3-900051-07-0. http://www.R-project.org/.
[71] Lin W. Agnostic notes on regression adjustments to experimental data: reexamining freedmanas critique. Ann Appl Stat. 2013;7(1):295–318.
[72] Guo K, Basse G. The generalized oaxaca-blinder estimator. J Amer Stat Assoc. 2021;118:1–13.
[73] Seber GA, Lee AJ. Linear regression analysis. Vol. 329. Hoboken, NJ: John Wiley & Sons; 2012.
[74] Piech C, Bassen J, Huang J, Ganguli S, Sahami M, Guibas LJ, et al. Deep knowledge tracing. In: Advances in neural information processing systems. Red Hook, NY: Curran Associates, Inc.; 2015. p. 505–13.
[75] Botelho AF, Baker RS, Heffernan NT. Improving sensor-free affect detection using deep learning. In: International Conference on Artificial Intelligence in Education. Cham, Switzerland: Springer; 2017. p. 40–51.
[76] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Netw. 1989;2(5):359–66.
[77] Schäfer AM, Zimmermann HG. Recurrent neural networks are universal approximators. In: International Conference on Artificial Neural Networks. Berlin: Springer; 2006. p. 632–40.
[78] Werbos PJ. Backpropagation through time: what it does and how to do it. Proc IEEE. 1990;78(10):1550–60.
[79] Kingma DP, Ba J. Adam: a method for stochastic optimization. 2014. arXiv: http://arXiv.org/abs/arXiv:14126980.
[80] Caruana R. Multitask learning. Machine Learn. 1997;28(1):41–75.

Appendix

A Summary of A/B test data

Table A1 gives sample sizes and skill builder completion rates in the 33 experiments discussed in the paper.

B Proof of Proposition

Proposition 1. *Let* $(y_1^c, y_1^t, x_1^r), \ldots, (y_N^c, y_N^t, x_N^r)$ *be IID samples from a population in which* y^c, y^t, *and* x^r *have finite fourth moments, and where* $-1 < \text{corr}(y^c, x^r) < 1$ *and* $-1 < \text{corr}(y^t, x^r) < 1$. *Let* b *be a fixed constant. Let* $\hat{\mathbb{V}}[\hat{\tau}^{\text{GR}}(b)]$ *denote the estimated variance of* $\hat{\tau}^{\text{GR}}(b)$, *defined analogously to* (4) *and* (13). *Let* $\hat{\mathbb{V}}\{\hat{\tau}^{\text{SS}}[x^r, \text{LS}]\}$ *denote the estimated variance of* $\hat{\tau}^{\text{SS}}[x^r, \text{LS}]$, *defined as in* (9). *Then as* $N \to \infty$,

$$\frac{\hat{\mathbb{V}}\{\hat{\tau}^{\text{SS}}[x^r, \text{LS}]\}}{\hat{\mathbb{V}}[\hat{\tau}^{\text{GR}}(b)]} \xrightarrow{p} \phi(b) \leq 1$$

where $\phi(b)$ *is some constant that depends on* b.

Proof. We first explicitly define $\hat{\mathbb{V}}[\hat{\tau}^{\text{GR}}(b)]$. Let $R_i^{\text{GR}} = Y_i - bx_i^r$ and define

$$\hat{\mathbb{V}}[\hat{\tau}^{\text{GR}}(b)] = \frac{S^2(R_C^{\text{GR}})}{n_c} + \frac{S^2(R_{\mathcal{T}}^{\text{GR}})}{n_t}. \tag{A1}$$

Comparing (A1) to (10), we see that in order to prove the desired result, it is sufficient to show that $\hat{E}_c^2 / S^2(R_C^{\text{GR}}) \xrightarrow{p} \phi_c(b) \leq 1$ and $\hat{E}_t^2 / S^2(R_{\mathcal{T}}^{\text{GR}}) \xrightarrow{p} \phi_t(b) \leq 1$, where $\phi_c(b)$ and $\phi_t(b)$ are constants that depend on b. We will show $\hat{E}_c^2 / S^2(R_C^{\text{GR}}) \xrightarrow{p} \phi_c(b) \leq 1$; the argument for $\hat{E}_t^2 / S^2(R_{\mathcal{T}}^{\text{GR}}) \xrightarrow{p} \phi_t(b) \leq 1$ is analogous.

Let $\tilde{E}_c^2$ be defined similarly to $\hat{E}_c^2$, except that $\tilde{E}_c^2$ does not use leave-one-out predictions, and instead uses predictions based on all of the data. That is,

$$\tilde{E}_c^2 = \frac{1}{n_c} \sum_{i \in C} [\tilde{y}^c(x_i^r) - y_i^c]^2, \tag{A2}$$

Table A1: Sample sizes and % homework completion – the outcome of interest – by treatment group in each of the 33 A/B tests

	n		% complete			*n*		% complete	
Experiment	**Trt**	**Ctl**	**Trt**	**Ctl**	**Experiment**	**Trt**	**Ctl**	**Trt**	**Ctl**
1	956	961	94	93	18	165	170	92	89
2	329	363	98	96	19	259	246	82	85
3	649	610	86	88	20	199	213	85	88
4	201	228	97	95	21	258	276	82	80
5	910	887	73	72	22	188	193	89	85
6	931	900	61	64	23	242	266	81	76
7	360	344	88	88	24	279	235	72	69
8	492	463	79	81	25	269	288	65	59
9	215	211	93	92	26	225	232	73	74
10	231	197	92	91	27	267	256	63	62
11	607	578	68	63	28	228	244	68	64
12	370	384	83	82	29	239	258	54	48
13	338	289	88	84	30	74	92	91	84
14	478	476	76	73	31	69	67	91	87
15	193	209	89	93	32	76	81	62	70
16	404	451	73	69	33	15	11	73	55
17	264	274	84	85					

where $\tilde{y}^c(x_i^r) = \tilde{a}^c + \tilde{b}^c x_i^r$ and where $\tilde{a}^c$ and $\tilde{b}^c$ are the intercept and slope coefficients, respectively, from a univariate regression of Y_C on x_C^r (not dropping observation i). Now note the following: (a) both $S^2(R_C^{\mathrm{GR}})$ and $\tilde{E}_c^2$ converge to finite constants; (b) the constant to which $S^2(R_C^{\mathrm{GR}})$ converges is not 0; and (c) $\tilde{E}_c^2 \leq S^2(R_C^{\mathrm{GR}})$ for all $n_c \geq 2$. (a) is ensured by the moment conditions. (b) is ensured by the condition $-1 < \mathrm{corr}(y^c, x^r) < 1$. (c) follows from the fact that $\tilde{E}_c^2 = \frac{1}{n_c}\min_{(a,b)}\sum_{j\in C}[Y_j - (a + bx_j^r)]^2$, whereas $S^2(R_C^{\mathrm{GR}}) = \frac{1}{n_c-1}\sum_{j\in C}[Y_j - bx_j^r - \overline{Y_j - bx_j^r}]^2 = \frac{1}{n_c-1}\min_a\sum_{j\in C}[Y_j - (a + bx_j^r)]^2$ for a fixed value of b, and thus, the minimization problem of the former is less constrained than the latter. As a result of (a), (b), and (c), it follows that $\tilde{E}_c^2 / S^2(R_C^{\mathrm{GR}}) \xrightarrow{p} \tilde{\phi}_c(b) \leq 1$.

To complete the proof, it suffices to show that $\hat{E}_c^2 \xrightarrow{p} \tilde{E}_c^2$. After some algebra,

$$\hat{E}_c^2 = \frac{1}{n_c}\sum_{i\in C}[\tilde{y}^c(x_i^r) - y_i^c]^2 / (1 - h_i)^2, \tag{A3}$$

where

$$h_i = \frac{1}{(n_c - 1)S^2(x_C^r)}[\overline{(x_C^r)^2} - 2\overline{x_C^r}x_i^r + (x_i^r)^2]. \tag{A4}$$

Here, the h_i are the diagonal entries of the hat matrix from the regression of Y_C on x_C^r and we use the well-known shortcut formula for calculating leave-one-out residuals [73]. Note that $0 < h_i \leq 1$. Thus,

$$|\hat{E}_c^2 - \tilde{E}_c^2| \leq \left[\frac{1}{(1 - h^*)^2} - 1\right]\tilde{E}_c^2 \tag{A5}$$

where $h^* = \max_C h_i$. However, because of the moment conditions on x^r, it is straightforward to show that $h^* \xrightarrow{p} 0$, and therefore $\hat{E}_c^2 \xrightarrow{p} \tilde{E}_c^2$. □

C Deep learning in the Remnant to impute completion

We used the remnant to train a variant of a recurrent neural network [65] called a long short-term memory (LSTM) network [66] to predict students' assignment completion. Deep learning models, and particularly LSTM networks, have been previously applied successfully to model similar temporal relationships in various areas of educational research [74,75].

Neural networks, including recurrent networks such as those explored here, are universal function approximators [76,77]. These models are commonly represented as "layers" of neurons; these feed from a set of inputs, through one or more "hidden" layers, to an output layer, where, in the basic case, the output of each layer is determined by equation (A6). In that equation, W is a set of learned weights, comparable to the coefficients learned in a regression model. The activation function $a(.)$ is commonly a nonlinearity that is applied to each layer in the network.

$$h_\ell = a(W * h_{\ell-1} + b), \tag{A6}$$

where h_0 is the input vector X.

Recurrent networks build upon this formulation to add layers that utilize not only the outputs of preceding layers, but also incorporate values from previous time steps within a supplied series; in time series data, the model estimates for a particular time step may be better informed by information from previous time steps, and a recurrent network structure is designed to take advantage of this likelihood. The LSTM networks explored here incorporate a set of "gates" that regulate the flow of data from both preceding layers and a "cell memory" that is calculated through previous time steps. The output of this LSTM layer is given by equations A7–A12.

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f), \tag{A7}$$

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i), \tag{A8}$$

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o), \tag{A9}$$

$$\tilde{C}_t = \tanh(W_C * [h_{t-1}, x_t] + b_C), \tag{A10}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t, \tag{A11}$$

$$h_t = o_t * \tanh(C_t), \tag{A12}$$

where t is given as recurrent layer ℓ on the given timestep.

In the aforementioned equations, gates f_t and i_t inform the cell memory C_t how much of the previously-computed memory should be forgotten and updated with the output of the previous time step and preceding layer, respectively.

As a recurrent network, the model is trained by iteratively updating the weight matrices (W in the above equations) through a procedure known as backpropagation through time [78] combined with a stochastic gradient descent method called Adam [79]. These methods are informed by a cost function (sometimes called a loss function) that is calculated through the comparison of model predictions with supplied ground truth labels. In this work, we adopted a network structure that incorporates multi-task learning [80] as a means of regularization. In other words, our model ultimately produces two sets of predictions corresponding with two outcomes of interest: student completion and inverse mastery speed, each on the subsequent assignment. By optimizing model weights in regard to these two outcomes, the process helps prevent the model from overfitting to either outcome; as student completion of their next assignment is the outcome explored in this work, the second outcome of inverse mastery speed is used only for this regularization purpose and is not utilized in subsequent analyses. Given that student completion is binary and inverse mastery speed is a continuous measure, the formula of which is described in Table A2, the cost function for our model training was calculated as a linear combination of two separate cost functions. Binary cross-entropy is used in the case of next assignment completion, as shown in equation (A13), while RMSE (equation (A14)) is used in the case of inverse mastery speed on the next assignment. The final cost function is then given as equation (A15), which is calculated over smaller smaller "batches" of samples over multiple training cycles known as epochs.

$$\text{BCE} = -(y * \log(\hat{y}) + (1 - y) * \log(1 - \hat{y})). \tag{A13}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum(y - \hat{y})^2}. \tag{A14}$$

Table A2: Assignment-level features in LSTM model

Input feature	Description
Problems started	The number of problems started by the student. (Untransformed & Sq.Root)
Problems completed	The number of problems completed by the student. (Untransformed & Sq.Root)
Inverse mastery speed	The inverse of the number of problems needed to complete the mastery assignment, or 0 where the student did not complete. (Untransformed & Sq.Root)
Percent correct	The percentage of problems answered correctly on the first attempt without the use of hints. (Untransformed & Sq.Root)
Assignment completion	Whether the current assignment was completed by the student.
Attempts per problem	The number of attempts taken to correctly answer each problem. (Avg. & Sq.Root)
First response time	The time taken per problem before making the first action. (Avg.)
Problem duration	The time, in seconds, needed to solve each problem. (Avg.)
Days with activity	The number of distinct days on which the student worked on each problem in the assignment. (Avg.)
Attempted problem first	Whether, on each problem, the first action was an attempt to answer (as opposed to a help request). (Avg.)
Requested answer hint	Whether, on each problem, the student needed to be given the answer to progress. (Avg.)

$$\text{Cost}_{\text{batch}} = \frac{\text{BCE}_{\text{batch}} + \text{RMSE}_{\text{batch}}}{2}. \tag{A15}$$

The training of the model continues by calculating the cost and iteratively updating model weights over multiple epochs until a stopping criterion is met. In this regard, we hold out 30% of the training data as a validation set. Model performance is calculated on this validation set after each epoch of training. Training ceases once the model performance on this validation set stops improving (i.e., the difference of model performance from one epoch to the next falls below a designed threshold). To avoid stopping the training process too early due to small fluctuations in model performance on the validation set early in the training procedure, a 5-epoch moving average of validation cost is used as the stopping criterion.

The specific model structure used in this work observed an LSTM network comprised of 3 layers. We used 16 covariates to describe each single time step, which then feeds into a hidden LSTM layer of 100 nodes, which is used to inform an output layer of two units corresponding with the previously described two outcomes of interest. The input features used in this model, described in Table A2, represent transformed and nontransformed versions of several metrics that describe different aspects of student performance within a single assignment. We considered sequences of at most ten worked skill builder assignments (c.f. Section 4.1), to predict student completion on a subsequent skill builder assignment.

We specified the LSTM model's hyperparameters (e.g., number of LSTM nodes, delta of stopping criterion, weight update step size) based on previously successful model structures and training procedures within the context of education. We evaluated the model using a 10-fold cross validation within the remnant to gain a measure of model fit (leading to an ROC area under the curve of 0.82 and root mean squared error of 0.34 for the dependent measure of next assignment completion). After this evaluation, the model is then re-trained using the full set of remnant data. This trained model is then used within the analyses described in Section 4.

D Comparing covariates in the remnant to the RCT

The requirement (5) that imputations $\hat{y}^c(\boldsymbol{x}_i)$ and $\hat{y}^t(\boldsymbol{x}_i)$ are independent of treatment assignment T_i precludes any use of RCT outcomes in training the imputation algorithm $\hat{y}^r(\cdot)$. This is due to the fact that, if there is indeed a treatment effect for any RCT subject, RCT outcomes are a function of T. This restriction includes the use of Y to select between competing $\hat{y}^r(\cdot)$ algorithms, or to decide whether to use remnant-based predictions $\hat{y}^r$ for covariate adjustment at all. That is, so long as analysts use only remnant outcomes, they may assess and modify $\hat{y}^r(\cdot)$ without restriction without violating (5), and they may not use outcome data from the RCT.

This restriction, however, does not extend to covariate data $\boldsymbol{x}$ from the RCT. An anonymous reviewer suggested developing a method comparing covariate distributions between the RCT and the remnant that may indicate the gain in precision an analyst may expect from including $\hat{y}^r$ in a covariate adjustment estimator.

Here, we discuss a technique we attempted, although we do not believe that it achieved its aim.

The intuition behind our approach is based roughly on "K-Nearest Neighbors" classification – if a subject in the RCT closely resembles other subjects in the remnant, an algorithm trained on the remnant may be able to predict that subject's outcome accurately, whereas if an RCT subject is unlike many other remnant subjects, the prediction is not likely to be accurate. Formally, let $K > 0$ be an integer, and $D(\cdot,\cdot)$ be a distance measure. Then, for subject i in the RCT and subject j in the remnant, let $d_{ij} = D(\boldsymbol{x}_i, \boldsymbol{x}_j)$, then, for each i, sort these distances so that $d_{i(1)} \leq d_{i(2)} \leq \ldots$. Finally, compute $\bar{d}_i^K = \sum_{k=1}^{K} d_{i(k)} / K$, the average distance between $\boldsymbol{x}_i$ and it's K nearest neighbors. The thought is that outcomes for subjects with low $\bar{d}_i^K$ should typically be easier to predict than subjects with larger $\bar{d}_i^K$. Distances within the remnant may form a reasonable basis of comparison – that is, one may compare $\bar{d}_i^k$ to the distribution of average distances between remnant subjects and their K nearest neighbors.

To calculate this measure for TestBed A/B tests, we first flattened each subject's covariate data by averaging their assignment-level statistics and also including a covariate equal to the number of included assignments. Then, we chose $K = 5$ and $D(\cdot,\cdot)$ to be the Mahalanobis distance, with the covariance matrix estimated using the remnant.

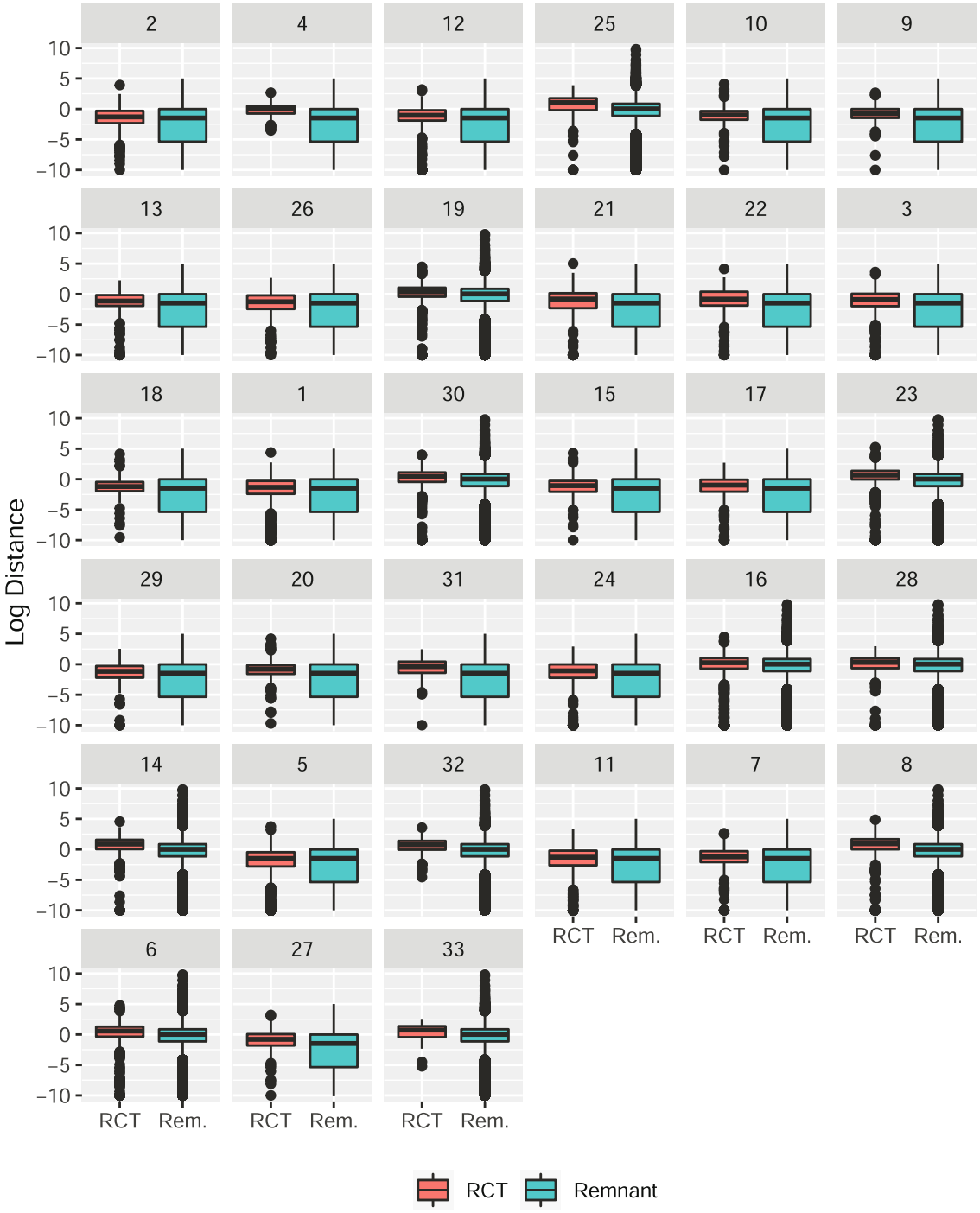

Figure A1: Boxplots comparing the distribution of $\bar{d}_i^5$ for each ASSISTments TestBed A/B test against the analogous distribution for the corresponding remnant. Panels are ordered from lowest to highest according to $\hat{\mathbb{V}}(\hat{\tau}^{\mathrm{DM}})/\hat{\mathbb{V}}(\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}])$.

Figure A1 shows the results. Each panel corresponds to a different A/B test and displays a boxplot of $\bar{d}_i^K$ for subjects in the RCT next to an analogous boxplot for the remnant. Note that across the 33 experiments, there were two distinct remnants, corresponding to two separate data draws. The panels are sorted lowest to highest according to the ratio $\hat{\mathbb{V}}(\hat{\tau}^{\mathrm{DM}})/\hat{\mathbb{V}}(\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}])$.

Unfortunately, no pattern is apparent, suggesting that d_i^5 is not a useful indicator of the variance reduction potential of algorithms trained in the remnant. Future research may lead to modifications of d_i^K or another measure entirely that may better anticipate $\hat{y}^r(\cdot)$'s out-of-sample performance. Fortunately, estimators $\hat{\tau}^{\mathrm{SS}}[x^r, \mathrm{LS}]$ and $\hat{\tau}^{\mathrm{SS}}[\tilde{\boldsymbol{x}}, \mathrm{EN}]$ often perform well, and (in our examples) never harm precision, even when $\hat{y}^r(\cdot)$ performs poorly in the RCT.