Using Auxiliary Data to Boost Precision in the Analysis of A/B Tests on an Online Educational Platform: New Data and New Results*

Adam C. Sales
Worcester Polytechnic Institute
asales@wpi.edu

Ethan B. Prihar
Worcester Polytechnic Institute
ebprihar@gmail.com

Johann A. Gagnon-Bartsch
University of Michigan
johanngb@umich.edu

Neil T. Heffernan
Worcester Polytechnic Institute
nth@wpi.edu

Randomized A/B tests within online learning platforms represent an exciting direction in learning sciences. With minimal assumptions, they allow causal effect estimation without confounding bias and exact statistical inference even in small samples. However, often experimental samples and/or treatment effects are small, A/B tests are underpowered, and effect estimates are overly imprecise. Recent methodological advances have shown that power and statistical precision can be substantially boosted by coupling design-based causal estimation to machine-learning models of rich log data from historical users who were not in the experiment. Estimates using these techniques remain unbiased and inference remains exact without any additional assumptions. This paper reviews those methods and applies them to a new dataset including over 250 randomized A/B comparisons conducted within ASSISTments, an online learning platform. We compare results across experiments using four novel deep-learning models of auxiliary data and show that incorporating auxiliary data into causal estimates is roughly equivalent to increasing the sample size by 20% on average, or as much as 50-80% in some cases, relative to t-tests, and by about 10% on average, or as much as 30-50%, compared to cutting-edge machine learning unbiased estimates that use only data from the experiments. We show that the gains can be even larger for estimating subgroup effects, hold even when the remnant is unrepresentative of the A/B test sample, and extend to post-stratification population effects estimators.

Keywords: A/B tests, deep learning, evaluation

1. INTRODUCTION

In randomized A/B tests on an online learning platform, students are randomized between different educational conditions or strategies, and their subsequent educational outcomes of interest are compared between different conditions. For instance, Harrison et al. (2020) studied

*Data and code used in this work can be found at https://osf.io/k8ph9/.

data from 2,152 middle- and high-school students whose teachers assigned a specific module—a "skill builder"—on the ASSISTments online tutoring platform (Heffernan and Heffernan, 2014). Prior to the students' work, the authors designed four different educational conditions, which differed in how the numbers and symbols in arithmetic expressions were spaced. As students logged on to the platform, during their usual schoolwork, they were each individually randomized to one of the four conditions, and completed their work under that condition. Subsequently, the authors of the study compared the average number of problems students in each condition had to work on before achieving mastery, defined as answering three problems correctly in a row. They found that students who were assigned the "congruent" condition—in which the spacing between numbers corresponded to the order of operations—needed to work on roughly one fewer problem, on average, than students in the "incongruent" condition. This finding, and others reported in the paper, validated their previous scientific hypotheses regarding embodied cognition, the relationship between abstract learning, and the arrangement of objects in physical (or virtual) space.

In general, A/B tests have two significant advantages over observational study designs, which do not include randomization, and additional advantages over studies conducted in a lab. First, they are (famously) free of confounding bias—since students are randomly allocated between conditions, differences in outcomes must be due to either a causal effect of the randomized conditions or to random error, but not to systematic baseline differences between students, observed or unobserved. Perhaps less famously, randomization forms a "reasoned basis for inference" (Fisher, 1935): the (known) probabilities of allocation of students between experimental conditions provide nearly all of the necessary justification for the unbiased estimation of causal effects, as well as standard errors, confidence intervals, and p-values. No other distributional assumptions or modeling assumptions are necessary. These properties allowed Harrison et al. (2020) to estimate causal effects of spacing conditions, as well as to statistically rule out other alternative explanations.[1] Causal effect and standard error estimators that rely only on the experimental design are referred to as "design-based" (Schochet, 2015).

On the other hand, A/B tests can be hobbled by statistical imprecision. For instance, Harrison et al. (2020) was unable to confirm or disconfirm one of their initial hypotheses, regarding differences in causal effects between subgroups of students, because the standard errors of the relevant estimates were too high. Unlike observational studies using data from online tutors, the sample size in A/B tests is necessarily limited to those students who worked on the relevant modules while the study was taking place. In contrast, a typical observational study would use data from all students who have ever worked on the relevant modules, including the (often large) number of students who worked on them before the onset of the study, and might sometimes use data from students who worked on similar modules as well. Analysis of A/B tests must discard data from these students, who were not randomized between treatment conditions and are subject to confounding. Unlike studies conducted in carefully controlled laboratory environments, A/B tests are subject to the haphazard unpredictability of real life, which only increases statistical imprecision—even a sample as large as the 2,152 of Harrison et al. (2020) may not be enough to answer some causal questions.

However, recent methodological innovations (Gagnon-Bartsch et al., Forthcoming; Sales et al., 2018) have argued that data from the "remnant" from an experiment—students who were

[1] Actually the authors of that paper did make modeling assumptions in their analysis, but they could have conducted a non-parametric analysis.

not randomized between conditions, but for whom covariate and outcome data are available—need not be discarded, but can play a valuable role in causal estimation. In fact, researchers can use data from the remnant to decrease experimental standard errors without sacrificing the unbiased estimation and design-based inference that recommend A/B testing. The basic idea is to first use the remnant data to train a machine learning model predicting outcomes as a function of covariates; then, use that fitted model to generate predicted outcomes for participants in the experiment. Finally, use those predictions as a covariate in a design-based covariate-adjusted causal estimator (Aronow and Middleton, 2013; Wager et al., 2016; Wu and Gagnon-Bartsch, 2018; Chernozhukov et al., 2018, for eg.). Variants of the method use the predictions from the remnant alongside other covariates to estimate causal effects.

These methods can help alleviate another weakness, shared by A/B tests and observational studies—the dependence of conclusions on statistical modeling choices. By observing outcome data prior to selecting and fitting statistical models, researchers (often inadvertently) choose models most favorable to their desired conclusions and undermine statistical objectivity and the logic of inference. Two proposed solutions to this issue are (1) to split the sample prior to data analysis and use one part to choose a model and the second part to estimate effects (Heller et al., 2009) or (2) to rely on flexible non-parametric models that can be specified prior to data collection (Van der Laan and Rose, 2011). Design-based estimators incorporating remnant data rely on both these techniques: model-fitting in the remnant can be interactive and based on human judgment, without adversely affecting the objectivity or validity of statistical inference using the experimental sample. Design-based covariate adjustment often uses robust or non-parametric models.

This paper reviews design-based effect estimation from A/B tests, along with a set of design-based causal estimators that use remnant data (Section 2). Next (Section 3) we describe a new dataset that we used to test these methods: a collection of 68 multi-armed A/B tests run on the ASSISTments TestBed (Ostrow et al., 2016, now called "E-Trials"), which together include 227 different two-way comparisons and 38,035 students. Alongside this experimental data, we collected log data for an additional 193,218 students who worked on similar skill builders in ASSISTments but did not participate in any of the 68 experiments—the remnant. The following section (Section 4) describes the Deep Learning model that we trained in the remnant to predict student outcomes as a function of prior log data.

The next four sections use that data and those models to address four research questions regarding the use of remnant data to assist in the analysis of A/B tests. The first research question (Section 5) regards the overall efficacy of our approach: to what extent might remnant data improve the precision of effect estimates from A/B tests? Does it ever harm precision, in practice? As part of this research question, we also investigated the roles that various types of remnant data may play in the process. The second research question (Section 6) regards subgroup effects—treatment effects may be present for some groups of students but not others, or may differ between groups of students. However, breaking A/B test data into subsets further exacerbates sample size issues—is this something remnant data may help with? The third research question (Section 7) regards differences between the remnant and A/B testing data—in particular, what if the remnant is known to be drawn from a different population than the participants in A/B tests? Can it still be useful? To answer this question, we purposely constructed a new remnant that we believe is composed mostly of white and Asian males and used it to analyze A/B testing data from primarily other demographic groups. The last research question (Section 8) asks if remnant data may be helpful in generalizing effects estimated from an A/B

Table 1: Descriptions of sets of subjects described in the text, and associated causal estimands.

Name	Abbreviation	Explanation	Avg. Effect
RCT Set	RCT	Participants in the RCT, randomized between $Z = 0$ and $Z = 1$ conditions	$\overline{\tau}_{RCT}$
Population of interest	POP	The total population for which researchers wish to estimate effects	$\mathbb{E}_{POP}[\tau]$
Subgroup k	$G = k$	One of K disjoint subsets of RCT or POP	$\overline{\tau}_{G=k}$ or $\mathbb{E}_{POP}[\tau \mid G = k]$
Remnant	REM	Subjects with covariate ($\boldsymbol{x}$) and outcome (Y) data available, but who were randomized between conditions in the RCT	n/a

test to a wider population, even when subjects in the A/B test were not randomly drawn from that population.

Across the board, we find that estimates using the remnant are often substantially more precise than estimates that do not, and very rarely are much less precise. This holds for overall estimates, estimated subgroup effects, population average effects, and even when the remnant is unrepresentative of the A/B test by construction. Our results give a much clearer picture of the potential impacts of using remnant data in design-based causal inference than was previously available.

2. Background

2.1. Framework: Different (Groups of) Users, Different (Average) Treatment Effects

For the method we are describing, it will be useful to define several different sets of subjects or users, summarized in Table 1 and Figure 1 (also see (Imbens, 2004)).

Consider an A/B test in which subjects $i = 1, \ldots, n$ are randomized between two conditions, which we denote as $Z_i = 0$ or $Z_i = 1$, with the goal of estimating effects of Z_i on an outcome Y_i. Call the set of randomized subjects i the "RCT set," or RCT. Typically, researchers running A/B tests are interested in the effect of Z on a broader population than RCT, such as all users of the system, or all users of a particular type; denote this target population as POP. For instance, students in a set of participating classrooms (RCT), working on a mastery-based homework assignment, may be randomized to receive tutoring in the form of either multi-step hints ($Z = 1$) or complete explanations of problem solutions ($Z = 0$), with the ultimate goal of estimating the effects of hints versus explanations on assignment completion (Y) for all users of the educational software (POP). (We focus on binary treatments for the sake of simplicity, though the methods and concepts we discuss extend easily to experiments with more than two conditions.)

Following (Neyman, 1923; Rubin, 1978) let $y_i(z)$, $z = 0, 1$ represent the outcome that

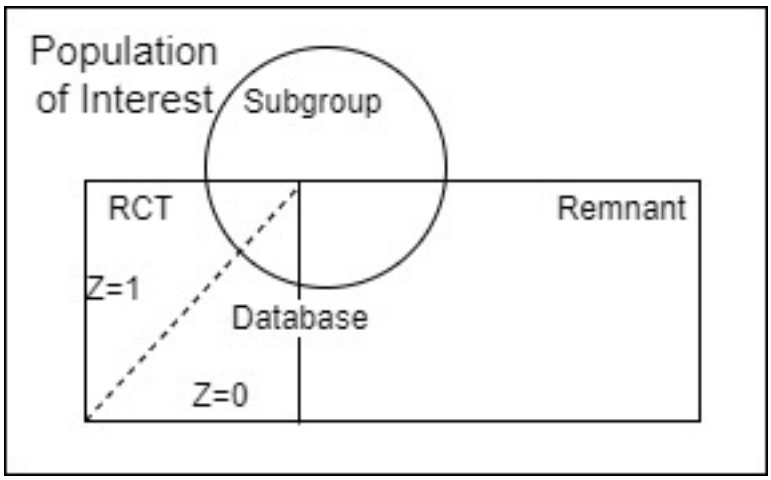

Figure 1: A Venn Diagram for the sets of subjects described in the text and Table 1.

subject i would experience if randomized to z—that is, if $Z_i = 0$, the observed outcome $Y_i = y_i(0)$, and if $Z_i = 1$ then $Y_i = y_i(1)$. Then, define the treatment effect for subject i as $\tau_i = y_i(1) - y_i(0)$, the difference between the outcome i would experience under condition 1 versus what they would experience under condition 0.

The challenge of causal inference is that for each i, only one of $y_i(0)$ or $y_i(1)$ is observed. Hence, individual treatment effects τ_i cannot be estimated directly (at least, not precisely), but under some circumstances, average treatment effects can be estimated.

2.1.1. The Sample Average Treatment Effect

First, consider the sample average treatment effect,

$$\overline{\tau}_{RCT} = \sum_{i=1}^{n} \tau_i / n = \overline{y(1)} - \overline{y(0)},$$

where $\overline{y(1)}$ is the sample average of $y(1)$ over every subject in the RCT (whether $Z = 1$ or $Z = 0$), and $\overline{y(0)}$ is the sample average of $y(0)$. Hence, $\overline{\tau}_{RCT}$ is never observed, but can often be estimated. Claims about $\overline{\tau}_{RCT}$ pertain only to the participants in RCT, not (necessarily) to treatment effects among other subjects.

2.1.2. The Population Average Treatment Effect

When researchers' interest goes beyond the average effect in RCT, and actually pertains to the larger population POP, then the estimand of interest is the population average effect, denoted $\mathbb{E}_{POP}[\tau]$[2]. If RCT is a random sample of POP, then there is little difference between estimating $\overline{\tau}_{RCT}$ and estimating $\mathbb{E}_{POP}[\tau]$. However, it is often the case that experimental participants are not representative of POP.

[2]We use the notation of expected value $\mathbb{E}[\cdot]$ instead of the sample average $\bar{\cdot}$ since it will often be mathematically convenient to think of POP as an infinite "super-population" from which subjects are drawn randomly (see, e.g. (Ding et al., 2017)).

2.1.3. Subgroup Effects

If prior to randomization the population is partitioned into K subgroups—for instance, students with high or low prior academic performance, students in different school districts, or students in different demographic categories—then let $G_i \in 1, \ldots, K$ denote subject i's group membership, so that if $G_i = k$, then i is in the k^{th} subgroup. Then $\overline{\tau}_{G=k}$ and $\mathbb{E}_{POP}[\tau \mid G = k]$ are average treatment effects for members of the subgroup k in RCT or the population POP, respectively. In general, $\overline{\tau}_{RCT} = \sum_{k=1}^{K} p_k \overline{\tau}_{G=k}$ and $\mathbb{E}_{POP}[\tau] = \sum_{k=1}^{K} \pi_k \mathbb{E}_{POP}[\tau \mid G = k]$, where p_k and π_k are the proportions of RCT and POP, respectively, that belonging to group k.

2.2. Estimation and Types of Causal Bias

Bias in estimating average effects depends on the causal estimand of interest and can be due to bias in estimating $\overline{\tau}_{RCT}$, which we will call "internal" bias, bias in estimating $\mathbb{E}_{POP}[\tau]$ due to differences between subjects in the experiment and the population, which we will call "external bias," or a combination of the two. Our terminology mirrors the distinction between internal and external validity (McDermott, 2011, for eg.).

2.2.1. Aside: Why do We Care about Statistical Bias?

While a good amount of early work in theoretical statistics focused on unbiased estimators, recent decades have seen increasing acknowledgment that unbiased estimators are often sub-optimal according to alternative estimation criteria and that a small amount of statistical bias may be a reasonable price to pay for improved statistical precision. That being the case, what accounts for our focus on unbiased estimation in this paper?

Although exact unbiasedness may not be an important goal for estimation in general, the concept of bias remains a useful formalization of some very important problems in estimation. For instance, the widely-known problems of estimating population quantities from unrepresentative or non-random samples or estimating causal effects from observational studies with unobserved confounding variables are both—in our opinion—most easily and clearly expressed in terms of bias. Extrapolation from unrepresentative samples and confounding can cause estimators to be inconsistent or inadmissible, and for confidence intervals and hypothesis tests to under-cover or over-reject, respectively. Our focus is on bias since we take it to be the simplest and most straightforward way to formalize confounding and unrepresentative sampling.

2.2.2. Estimating $\overline{\tau}_{RCT}$ and Internal Bias

In a completely randomized experiment, the set of subjects with $Z = 1$ are a random sample of all the experimental participants, so $\overline{Y}_{Z=1} = (\sum_{i=1}^{n} Y_i Z_i)/(\sum_{i=1}^{n} Z_i)$, the average observed outcome for treated subjects, is an unbiased estimate of $\overline{y(1)}$, and likewise $\overline{Y}_{Z=0}$ is an unbiased estimator of $\overline{y(0)}$. (In general, let $\overline{X}_G$ be the sample mean of X for subjects for whom G is true $(\sum_{i=1}^{n} X_i \mathbf{1}\{G_i\})/(\sum_{i=1}^{n} \mathbf{1}\{G_i\})$, where $\mathbf{1}\{G_i\} = 1$ if G is true for i and 0 otherwise.) Then

$$\hat{\tau}^{DM} = \overline{Y}_{Z=1} - \overline{Y}_{Z=0},$$

the "difference-in-means" or "T-Test" estimator, is (internally) unbiased for $\overline{\tau}_{RCT}$. The same reasoning extends to estimates of subgroup effects $\overline{\tau}_{G=k}$—the difference in mean outcomes for RCT subjects with $G = k$ between $Z = 1$ and $Z = 0$ is internally unbiased, i.e. unbiased for $\overline{\tau}_{G=k}$.

However, if treatment Z is not randomized—or if randomization is "broken" due to attrition or some other irregularity—then $\hat{\tau}^{DM}$ will be biased due to confounding. Similarly, if treatment was randomized, but with different probabilities of treatment assignment for different subjects, $\hat{\tau}^{DM}$ may be a biased estimate of $\overline{\tau}_{RCT}$.

Even in a completely randomized experiment without other complications, some common effect estimators are biased for $\overline{\tau}_{RCT}$. For instance, say a vector of covariates $\boldsymbol{x}_i$ is observed for each subject. The ANCOVA estimator for $\overline{\tau}_{RCT}$, the estimated coefficient on Z from an ordinary least squares (OLS) regression of Y on Z and $\boldsymbol{x}$, is biased for $\overline{\tau}_{RCT}$ unless the linear model is correct. In general, non-linear relationships between $\boldsymbol{x}$ and $y(0)$ or $y(1)$, or un-modeled interactions between the treatment indicator and $\boldsymbol{x}$, will lead to bias in the ANCOVA estimator. That said, when $\boldsymbol{x}$ has low dimension relative to n, the bias of the ANCOVA estimator is negligible (under suitable regularity conditions it decreases roughly with $1/n$; Freedman (2008)). However, if $\boldsymbol{x}$ has high dimension relative to n, or if a prediction algorithm other than OLS is used (improperly), the bias might be substantial.

2.2.3. Estimating $\mathbb{E}_{POP}[\tau]$ and External Bias

An unbiased estimator of $\overline{\tau}_{RCT}$ may still be biased for $\mathbb{E}_{POP}[\tau]$, depending on the population of interest POP. For instance, consider a stylized example in which G encoded income level: poor $G = 1$ versus rich $G = 2$, and that the effect of an intervention differs by income level—say $\mathbb{E}_{POP}[\tau \mid G = 1] < \mathbb{E}_{POP}[\tau \mid G = 2]$—and that sample proportions $p_1 < p_2$ while population proportions $\pi_1 > \pi_2$, so the experiment was conducted among subjects who were wealthier, on average, than the population of interest. Finally, say that within income groups G, the experimental subjects are representative of the corresponding subgroups in the population so that $\mathbb{E}[\overline{\tau}_{G=k}] = \mathbb{E}_{POP}[\tau \mid G = k]$. Let $\hat{\tau}$ be an unbiased estimator of $\overline{\tau}_{RCT}$. As an estimate of the population average effect $\mathbb{E}_{POP}[\tau]$, $\hat{\tau}$ will be biased:

$$\begin{aligned}
&\mathbb{E}[\hat{\tau}] - \mathbb{E}_{POP}[\tau] = \mathbb{E}[\overline{\tau}_{RCT}] - \mathbb{E}_{POP}[\tau] \\
=&p_1 \mathbb{E}\left[\overline{\tau}_{G=1}\right] + (1-p_1)\mathbb{E}\left[\overline{\tau}_{G=2}\right] - \pi_1 \mathbb{E}_{POP}[\tau \mid G = 1] - (1-\pi_1)\mathbb{E}_{POP}[\tau \mid G = 2] \\
=&(p_1 - \pi_1)\mathbb{E}_{POP}[\tau \mid G = 1] + (\pi_1 - p_1)\mathbb{E}_{POP}[\tau \mid G = 2] \\
=&(p_1 - \pi_1)(\mathbb{E}_{POP}[\tau \mid G = 1] - \mathbb{E}_{POP}[\tau \mid G = 2]) > 0
\end{aligned} \tag{1}$$

since $p_2 = 1 - p_1$ and $\pi_2 = 1 - \pi_1$. It is clear from (1) that if either $p_1 = \pi_1$, so that the subjects in the experiment are representative of POP, or if $\mathbb{E}_{POP}[\tau \mid G = 1] = \mathbb{E}_{POP}[\tau \mid G = 2]$, so that the average effect of the treatment doesn't vary with G, that $\hat{\tau}$ will be unbiased. In general, for an estimate to be externally biased, there must be at least one (observed or unobserved) characteristic in which the subjects in the experiment do not represent the population, *and* which predicts variation in the treatment effect. If the ways in which the experimental sample is unrepresentative are unrelated to treatment effect variation, then there will be no external bias.

Since, in the example above, $\hat{\tau}$ was unbiased for $\overline{\tau}_{RCT}$, the bias of (1) is purely external bias. However, if internal bias is also present, then the two biases add, so that

$$\mathbb{E}\left[\hat{\tau}\right] - \mathbb{E}_{POP}[\tau] = \text{internal bias} + \text{external bias}. \tag{2}$$

Note, however, that if internal and external bias have opposite signs, they may (partially) cancel each other out—that said, it is hard to know when this fortunate situation may or may not hold.

Sometimes subgroup effect estimates can be combined to mitigate external bias from an unrepresentative RCT sample, via post-stratification (Miratrix et al., 2013). Say $\mathbb{E}[\overline{\tau}_{G=k}] \approx \mathbb{E}_{POP}[\tau \mid G = k]$ as in the example above, and that population proportions π_k are known. Let $\hat{\tau}_k$ be unbiased estimates of $\overline{\tau}_{G=k}$; then,

$$\mathbb{E}\left[\sum_k \pi_k \hat{\tau}_k\right] = \sum_k \pi_k \mathbb{E}[\hat{\tau}_k] \approx \sum_k .\pi_k \mathbb{E}_{POP}[\tau \mid G = k] = \mathbb{E}_{POP}[\tau] \tag{3}$$

That is, if estimated subgroup effects $\hat{\tau}_k$ are unbiased, then the post-stratification estimator $\sum_i \pi_k \hat{\tau}_k$ will also be externally unbiased. Hence, accurate estimation of subgroup effects can reduce external bias of overall population effects.

2.3. Internally Unbiased Estimators using Auxiliary Data

2.3.1. The Remnant

While the difference-in-means estimator $\hat{\tau}^{DM}$ is unbiased for $\overline{\tau}_{RCT}$ in a completely randomized experiment, it may be imprecise, especially when the sample size is small. This problem may be exacerbated if a researcher is interested in estimating subgroup effects, either because of scientific interest in subgroups or for the sake of post-stratification. The reason is that $\overline{\tau}_{RCT}$ depends on unobserved counterfactual potential outcomes, $y_i(0)$ if $Z_i = 1$ and $y_i(1)$ if $Z_i = 0$, which must be imputed. $\hat{\tau}^{DM}$ relies on very rudimentary imputation strategy: the imputed $\hat{y}_i(0) = \overline{Y}_{Z=0}$ for all i such that $Z_i = 1$, and $\hat{y}_i(1) = \overline{Y}_{Z=1}$ for all i such that $Z_i = 0$. This strategy ignores all observed differences between subjects in the experiment, instead imputing one of the same two values for every subject.

In many cases, covariate and outcome data from an experiment are drawn from a larger database. For instance, educational field trials may use state longitudinal data systems to collect covariate data on student demographics and prior achievement as well as on outcomes of interest such as standardized test scores, and medical trials may gather baseline and outcome data from databases of medical records. Most relevant for our purposes, analysis of A/B tests within online applications can access rich baseline data from users' logs prior to the onset of the experiment and often draw outcome data from that same source. In these cases, researchers have the option of gathering additional auxiliary data—covariate and outcome data from users who were not part of the experiment. This includes historical data from before the onset of the experiment, as well as data from concurrent users who were not part of the experiment for some other reason. We refer to this set of users as the "remnant" from the experiment (Sales et al., 2018) (rounding out the list of sets described in Table 1 and Figure 1).

2.3.2. A Naive Estimator using the Remnant

Say, for the sake of argument, that every subject in the remnant was in the $Z = 0$ condition; this will be the case if, for instance, $Z = 0$ represents a "business as usual" condition. Then, say researchers used the remnant to train an algorithm $\hat{y}^{REM}(\boldsymbol{x}; \boldsymbol{\beta})$ predicting outcomes from covariates $\boldsymbol{x}$, with parameters $\boldsymbol{\beta}$, estimated with remnant data as $\hat{\boldsymbol{\beta}}$. Define this algorithm's prediction for each experimental subject i as $\hat{y}_i^r \equiv \hat{y}^{REM}\left(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}\right)$ (where "$\equiv$" denotes definition). Researchers could use these to impute control potential outcomes $y(0)$ for participants in the experiment as $\hat{y}_i(0) = \hat{y}_i^r$. That is, for each experimental participant with $Z_i = 1$, estimate an individual treatment effect of $\hat{\tau}_i = Y_i - \hat{y}_i^r$ and estimate $\bar{\tau}$ or $\mathbb{E}_{POP}[\tau]$ as $\hat{\tau}_{naive} = \overline{\hat{\tau}}_{Z=1}$.

The estimator $\hat{\tau}_{naive}$ has the potential to be much more precise than $\hat{\tau}^{DM}$ since it can account for observed baseline differences between experimental subjects, and use those differences to tailor its imputations to each individual subject. On the other hand, it has two serious disadvantages. First, the participants in the experiment are not necessarily drawn from the same population as the remnant, so there is no guarantee that the conditional distribution of $y(0)$ given $\boldsymbol{x}$ is the same in both groups. If the remnant is not representative of the experiment so that $p(y(0)|\boldsymbol{x})$ differs between the two sets, $\hat{\tau}_{naive}$ may be biased for both $\bar{\tau}$ and $\mathbb{E}_{POP}[\tau]$. Second, even if the remnant is representative of the sample, there is typically no guarantee that the predictions $\hat{y}^{REM}(\boldsymbol{x};\boldsymbol{\beta})$ are unbiased—in this case, the often erratic behavior of supervised learning algorithms in finite samples can also lead to bias.

2.3.3. Better Estimation using the Remnant

Both of these disadvantages can be corrected by relying on *both* randomization and supervised learning from the remnant. Specifically, the problems that cause internal bias in $\hat{\tau}_{naive}$ will also be present when comparing Y_i to $\hat{y}_i^r$ for subjects in the control group, leading to the "remnant-based residualization" or "rebar" estimator (Sales et al., 2018),

$$\hat{\tau}_{rebar} \equiv \hat{\tau}_{naive} - \overline{Y - \hat{\boldsymbol{y}}^{\boldsymbol{r}}}_{Z=0} = \overline{Y - \hat{\boldsymbol{y}}^{\boldsymbol{r}}}_{Z=1} - \overline{Y - \hat{\boldsymbol{y}}^{\boldsymbol{r}}}_{Z=0} = \hat{\tau}^{DM} - \overline{\hat{\boldsymbol{y}}^{\boldsymbol{r}}}_{Z=1} - \overline{\hat{\boldsymbol{y}}^{\boldsymbol{r}}}_{Z=0}, \quad (4)$$

where $\hat{\boldsymbol{y}}^{\boldsymbol{r}}$ is the vector of imputations $\{\hat{y}_i^r\}_{i=1}^n$. As (4) suggests, there are (at least) two ways to conceptualize the rebar estimator: first, it corrects the bias of $\hat{\tau}_{naive}$ by subtracting the analogous contrast in the $Z = 0$ group, $\overline{Y - \hat{\boldsymbol{y}}^{\boldsymbol{r}}}_{Z=0}$, and second, it corrects for imprecision in $\hat{\tau}^{DM}$ by subtracting the finite-sample difference in $\hat{\boldsymbol{y}}^{\boldsymbol{r}}$ between students in the two treatment conditions. $\hat{\tau}_{rebar}$ is precise if $\hat{\boldsymbol{y}}^{\boldsymbol{r}}$ is close to $y(0)$, on average, and is always unbiased for $\bar{\tau}$, due to the randomization of treatment assignment. Importantly, because the parameters β from the algorithm $\hat{y}^{REM}(\boldsymbol{x};\boldsymbol{\beta})$ are estimated using a separate sample, and $\boldsymbol{x}$ is fixed at baseline, $\hat{\tau}_{rebar}$ will be unbiased for $\bar{\tau}$ regardless of whether imputations $\hat{\boldsymbol{y}}^{\boldsymbol{r}}$ are themselves accurate or biased. This property is guaranteed by the randomization of treatment assignment. In fact, it applies regardless of whether subjects in the remnant were in the $Z = 0$ or $Z = 1$ condition, or some other condition altogether.

The problem with $\hat{\tau}_{rebar}$ is that if the algorithm $\hat{y}^{REM}(\boldsymbol{x};\boldsymbol{\beta})$ performs poorly for subjects in RCT, then $\hat{\tau}_{rebar}$ will have high variance—sometimes even higher than $\hat{\tau}^{DM}$. A better solution is based on the fact that, in essence, $\hat{y}_i^r$ is itself a covariate, since it is a function of covariates $\boldsymbol{x}_i$ and parameters $\boldsymbol{\beta}$ estimated using a separate sample. That being the case, it can be used as a covariate, perhaps along with others, in an existing covariate-adjusted estimator of $\overline{\tau}_{RCT}$.

For instance, consider the ANCOVA estimator based on the following OLS model:

$$Y_i = \alpha_0 + \alpha_1 \hat{y}^r + \tau_{OLS} Z_i + \epsilon_i, \quad (5)$$

where ϵ_i is a mean-0 error term. The estimated coefficient on Z from this model, $\hat{\tau}_{OLS}$, triangulates between $\hat{\tau}_{rebar}$ and $\hat{\tau}^{DM}$, essentially picking whichever estimator is better. If $\hat{y}^r$ is highly correlated with Y, then we might expect its estimated coefficient in (5) $\hat{\alpha}_1 \approx 1$, in which case $\hat{\tau}_{OLS} \approx \hat{\tau}_{rebar}$. If, on the other hand, $\hat{y}^r$ is a poor prediction of Y, then $\hat{\alpha}_1 \approx 0$ and $\hat{\tau}_{OLS} \approx \hat{\tau}^{DM}$. However, as discussed earlier, while $\hat{\tau}_{OLS}$ is a consistent estimator of $\overline{\tau}_{RCT}$, it is slightly biased, and its associated standard error estimates require either large samples or additional modeling assumptions. Researchers may consider including additional covariates as predictors alongside

$\hat{y}^r$ in an ANCOVA model like (5); however, as the number of covariates, interactions and/or non-linear terms increases, so may bias or other inferential issues.

Alternatively, Gagnon-Bartsch et al. (Forthcoming) suggests incorporating $\hat{\boldsymbol{y}}^{\boldsymbol{r}}$, perhaps alongside other covariates, into a flexible, internally-unbiased effect estimator that adjusts for baseline covariates (Wager et al., 2016; Aronow and Middleton, 2013, for eg.). We will focus here on the "LOOP" estimator (Wu and Gagnon-Bartsch, 2018). As above, for each subject i, let $\boldsymbol{x}_i$ be a vector of covariates. In general, LOOP is an alternative to ANCOVA for A/B tests with Bernoulli randomization, in which each subject is independently randomized with $Pr(Z_i = 1) = p$ for all i. Like ANCOVA, LOOP estimates $\overline{\tau}_{RCT}$ after adjusting for baseline differences in $\boldsymbol{x}$ between subjects assigned to $Z = 0$ and $Z = 1$. Unlike ANCOVA, LOOP estimates are exactly unbiased for $\overline{\tau}_{RCT}$ and are not limited linear models of the covariates—in principle, they can accommodate any model relating outcomes to covariates, including models that can incorporate high-dimensional covariate matrices. Write the standard LOOP estimator, adjusting for covariates $\boldsymbol{x}$, as $\hat{\tau}_{LOOP}(\boldsymbol{x})$—this estimator adjusts for $\boldsymbol{x}$ but does not use the model trained in the remnant. We recommend two alternatives: $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}})$, which adjusts the estimate for $\hat{\boldsymbol{y}}^{\boldsymbol{r}}$ instead of $\boldsymbol{x}$—this estimator is quite similar to $\hat{\tau}_{OLS}$ in (5) but unbiased—and $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$, which adjusts for *both* $\hat{y}^{REM}(\boldsymbol{x})$ and $\boldsymbol{x}$, incorporating all of the best properties of $\hat{\tau}^{DM}$, $\hat{\tau}_{LOOP}(\boldsymbol{x})$, and $\hat{\tau}_{rebar}$. These estimators are design-based, exactly unbiased for the $\overline{\tau}_{RCT}$, give conservative standard error estimates, and make no modeling assumptions, beyond the design of the experiment itself.

The following sub-section gives a more technical description of the LOOP estimator and $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$ for interested readers.

2.3.4. The LOOP Estimator with Remnant-Based Predictions

In a Bernoulli-randomized A/B test, specify an algorithm $\widehat{y(z)}^{RCT}(\boldsymbol{x}, \hat{y}^r; \boldsymbol{\alpha})$ to impute potential outcomes $y(0)$ and $y(1)$ from remnant-based imputations $\hat{\boldsymbol{y}}^{\boldsymbol{r}}$, and (optionally) covariates $\boldsymbol{x}$, with parameters $\boldsymbol{\alpha}$. (Note that there are two separate algorithms predicting Y from $\boldsymbol{x}$: $\hat{y}^{REM}(\boldsymbol{x}; \boldsymbol{\beta})$ is fit using data from the remnant and produces imputations $\hat{y}^r_i$, while $\widehat{y(z)}^{RCT}(\boldsymbol{x}, \hat{y}^r; \boldsymbol{\alpha})$ is fit using RCT data.) For instance, (Gagnon-Bartsch et al., Forthcoming) considers models

$$\widehat{y(z)}^{RCT}(\hat{y}^r; \boldsymbol{\alpha})_{OLS} = \alpha_0^z + \alpha_1^z \hat{y}^r, \tag{6}$$

where $\boldsymbol{\alpha} = [\alpha_0, \alpha_1]$, an OLS intercept and slope estimated separately in each treatment arm, as well as a random forest (RF) predictor, $\widehat{y(z)}^{RCT}(\boldsymbol{x}, \hat{y}^r; \boldsymbol{\alpha})_{RF}$ incorporating covariates $\boldsymbol{x}$ alongside $\hat{\boldsymbol{y}}^{\boldsymbol{r}}$ as predictors, but ultimately recommends an ensemble of the two.

To estimate $\overline{\tau}_{RCT}$ without bias, it is essential that the predictions from $\widehat{y(z)}^{RCT}(\boldsymbol{x}, \hat{y}^r; \boldsymbol{\alpha})$ be statistically independent from the treatment assignment Z. The recommended estimators in (Gagnon-Bartsch et al., Forthcoming) ensure that this is the case by using leave-one-out sample-splitting. For each subject in the experiment $i = 1, \ldots, n$, estimate $\boldsymbol{\alpha}$ using data from the other $n - 1$ subjects. Denote this estimate of $\boldsymbol{\alpha}$, using data from all RCT subjects except i, as $\hat{\boldsymbol{\alpha}}_{(i)}$. Impute missing potential outcomes using predictions $\widehat{y_i(0)}^{RCT}(\hat{y}^r, \boldsymbol{x}) = \widehat{y(0)}^{RCT}\left(\hat{y}^r_i, \boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}_{(i)}\right)$ and $\widehat{y_i(1)}^{RCT}(\hat{y}^r, \boldsymbol{x}) = \widehat{y(1)}^{RCT}\left(\hat{y}^r_i, \boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}_{(i)}\right)$.

Finally, estimate $\overline{\tau}_{RCT}$: first, let $\hat{m}_i(\hat{y}^r, \boldsymbol{x}) = p\widehat{y_i(0)}^{RCT}(\hat{y}^r, \boldsymbol{x}) + (1-p)\widehat{y_i(1)}^{RCT}(\hat{y}^r, \boldsymbol{x})$, an

imputation of i's expected counterfactual potential outcome. Then estimate $\overline{\tau}$ as:

$$\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x}) = \sum_{i:Z_i=1} \frac{Y_i - \hat{m}_i(\hat{y}^r, \boldsymbol{x})}{np} - \sum_{i:Z_i=0} \frac{Y_i - \hat{m}_i(\hat{y}^r, \boldsymbol{x})}{n(1-p)}, \tag{7}$$

where p, as above, is the probability of an individual participant being assigned to the $Z = 1$ condition. The $(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$ in the notation $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$ refer to the data included in the imputation algorithms that give rise to $\hat{m}$. In the following section, we contrast $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$ with variants $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}})$, in which $\hat{m}_i = \hat{m}(\hat{y}_i^r)$ is a function of $\hat{y}_i^r$ only, and $\hat{\tau}_{LOOP}(\boldsymbol{x})$, in which $\hat{m}_i = \hat{m}(\boldsymbol{x}_i)$ is a function of $\boldsymbol{x}_i$ but not $\hat{y}_i^r$.

$\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$ and its variants are inverse-probability-weighted estimates (also called Horvitz Thompson)—they are similar in form to $\hat{\tau}^{DM}$, except with the treatment and control sample sizes replaced with their expected values, np and $n(1-p)$. Aside from that difference, $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$ with $\hat{m}_i = 0$ would correspond to $\hat{\tau}^{DM}$, and $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}})$ with $\hat{m}_i = \hat{y}_i^r$ would be equivalent to $\hat{\tau}_{rebar}$. In general, $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$ is much more flexible than either $\hat{\tau}^{DM}$ or $\hat{\tau}_{rebar}$, since it allows $\hat{y}^r$'s role to vary depending on its prognostic value, and because it allows flexible incorporation of other baseline covariates $\boldsymbol{x}$.

Because parameters $\boldsymbol{\alpha}$ are estimated independently of i's outcome data, and $\boldsymbol{x}_i$ is fixed prior to treatment assignment, the sample splitting estimator is unbiased for the sample average treatment effect $\overline{\tau}$.

In Gagnon-Bartsch et al. (Forthcoming), incorporating $\hat{y}_i^r$ into the LOOP estimator led, in many cases, to substantial gains in precision compared to either $\hat{\tau}^{DM}$ or to the LOOP estimator with other covariates but not $\hat{y}_i^r$.

None of the methods considered here assumes that either imputation model, $\hat{y}^{REM}(0)(\cdot; \boldsymbol{\beta})$ or $\widehat{y(z)}^{RCT}(\boldsymbol{x}, \hat{y}^r; \boldsymbol{\alpha})$ is correct, unbiased, or consistent in any sense. Regardless of the quality of the imputation methods, randomization of treatment assignment ensures that effect estimates are unbiased.

2.3.5. Specific Estimators and Associated Terminology

Our two recommended estimators, which we term RELOOP and RELOOP+, combine ideas from $\hat{\tau}_{rebar}$ (4) and the leave-one-out covariate adjustment strategy LOOP (Wu and Gagnon-Bartsch, 2018)—hence the name "RELOOP." We will compare RELOOP and RELOOP+ to the T-Test estimator $\hat{\tau}^{DM}$, and a LOOP estimator that does not use remnant data. All told, we consider four different estimators:

- "T-Test": the difference-in-means estimator $\hat{\tau}^{DM}$, with no covariate adjustment
- "LOOP": $\hat{\tau}_{LOOP}(\boldsymbol{x})$ adjusts for covariates using a random forest imputation model fit to RCT data. It does not use any remnant data.
- "RELOOP": $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}})$ adjusts only for $\hat{y}_i^r$, imputations from the model trained in the remnant, using LOOP with the OLS RCT imputation model (6). It adjusts for no other covariates.
- "RELOOP+": $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$ uses an ensemble of OLS and random forests trained in RCT to adjust for both $\hat{y}_i^r$ and other covariates.

When an imputation model $\widehat{y(z)}^{RCT}(\boldsymbol{x}, \hat{y}^r; \boldsymbol{\alpha})$ is trained using RCT data, we refer to the associated covariate adjustment as "within-sample" or "within-RCT" adjustment. When an imputation model is trained in the remnant (i.e. $\hat{y}^{REM}(\boldsymbol{x}; \boldsymbol{\beta})$), we refer to the associated covariate adjustment as "remnant-based." Comparing the two types of adjustment, within-sample adjustment has the advantage of hewing more closely to the actual RCT data on which it's trained, while remnant-based adjustment can rely on models fit using the remnant, which may boast a much larger sample size than the RCT. ReLOOP and ReLOOP+ make use of both types of adjustment.

2.3.6. Estimating Sampling Variance, p-values, and Confidence Intervals

The true sampling variances of $\hat{\tau}^{DM}$, $\hat{\tau}_{rebar}$, and $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$, as estimates of $\overline{\tau}_{RCT}$, depend on the correlation between $y(0)$ and $y(1)$, which is not identified without making further assumptions, since $y(0)$ and $y(1)$ are never observed simultaneously. However, it is possible to *conservatively* estimate the sampling variances of all three estimators. Specifically, for $z = 0, 1$, let

$$\hat{E}_z^2 = \frac{1}{n_z} \sum_{i:Z_i=z} \left[\widehat{y(z)}^{RCT} \left(\hat{y}_i^r, \boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}^z{}_{(i)} \right) - Y \right]^2 .$$

Then estimate the sampling variance of $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$ as:

$$\widehat{\mathbb{V}}\left(\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})\right) \frac{1}{n} \left[\frac{p}{1-p} \hat{E}_0^2 + \frac{1-p}{p} \hat{E}_1^2 + 2\hat{E}_0 \hat{E}_1 \right].$$

As Wu and Gagnon-Bartsch (2018) show, $\mathbb{E}\left[\widehat{\mathbb{V}}\left(\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})\right)\right] \geq \mathbb{V}\left(\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})\right)$—that is, $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$'s estimated sampling variance is conservative in expectation.

Let the estimated standard error of $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$ $\widehat{SE} = \widehat{\mathbb{V}}\left(\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})\right)^{1/2}$. The usual $1-\alpha$ confidence interval has asymptotic coverage of at least $1-\alpha$—i.e.

$$Pr(\bar{\tau} \in \hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x}) \pm \mathfrak{z}_{1-\alpha/2} SE) \to 1 - \tilde{\alpha} \geq 1 - \alpha$$

as $n \to \infty$, where $\mathfrak{z}_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. Similarly, a hypothesis test that rejects the null hypothesis of $\bar{\tau} = 0$ when $|\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})/SE| \geq \mathfrak{z}_{1-\alpha/2}$ will have a type-I error rate of at most α in large samples.

The possible upward bias in these variance estimates will, if anything, cause confidence intervals to include the true parameter too often, or cause type-I error rates to be too low. While an unbiased sampling variance estimator would be preferable, conservative estimators are (arguably) the next best thing.

3. Data from 68 Educational A/B Tests

The remainder of the paper will discuss a set of illustrations and case-studies in using the ReLOOP and ReLOOP+ to estimate causal treatment effects from A/B tests run on an educational technology platform. This section describes the dataset—first the A/B tests themselves, and then the remnant—and the following section describes $\hat{y}^{REM}(\boldsymbol{x}; \boldsymbol{\beta})$, the deep-learning model trained using remnant data. Subsequent sections will use data from the A/B tests and imputations from the model trained in the remnant to answer our research questions.

E-Trials is a platform that allows researchers to design educational experiments that will then be run within the ASSISTments online tutor. Education researchers can specify experimental conditions, including variation on how subject matter is portrayed, available hints, and feedback to students. Researchers also choose learning modules on which their experiments run. When teachers subsequently assign these modules to their students, the students are randomized between the conditions. After the period of the experiment has ended, the researcher is provided with a dataset, including classroom and student identifiers, log data from during the experiment, and outcome data such as which students completed the assignment and how many problems they worked. Students are randomized between conditions independently, one at a time; when there are only two conditions, this is Bernoulli randomization.

We gathered data from a set of 84 A/B tests run on E-Trials. Since our interest here is primarily methodological, with the goal of reducing standard errors, we focus on estimated standard errors as opposed to treatment effects. Our analyses focus on assignment completion as a binary outcome.

We also gathered a set of nine student-level aggregated predictors, to be used for within-RCT covariate adjustment. These were the numbers of skill builders (mastery-based modules in ASSISTments) and problem sets each student began and completed, as well as each student's prior median first response time when working ASSISTments problems, median time on task, overall correctness, completed problem count, and average attempt count.

Several experiments included multiple conditions, rather than only treatment and control. We assume that the primary interest in these experiments lies in head-to-head comparisons between conditions, and, as such, we analyze all unique pairs of conditions within randomized experiments separately. All in all, this included 383 pairs. However, not every pair was amenable to analysis. Six pairwise contrasts were dropped because the outcome variance in one or both of the conditions was zero. Further exclusions were motivated by two factors: first, the LOOP estimator (which also underlies the RELOOP and RELOOP+ estimators) presumes that $p_i = Pr(Z_i = 1)$ is known. When the experiments were run, the E-Trials platform was only equipped to run Bernoulli-randomized experiments in which students were independently assigned to available conditions with equal probability. Hence, in theory, $p = 1/2$ should hold in all pairwise comparisons. However, there were strong indications that a handful of experiments used a different randomization scheme—we suspect that in some cases two conditions were combined, leading to $p = 2/3$ or 1/3. To exclude cases in which $p \neq 1/2$, we estimated p-values testing the null hypothesis that $p = 1/2$ for each comparison we considered; we dropped contrasts in which the p-value testing $p = 1/2$ was < 0.1. Secondly, there were some contrasts that included extremely small samples, with the smallest being $n = 16$. The LOOP estimators rely on OLS regression or more complex models, and cannot be expected to perform well when sample sizes are so small. In the main analyses, we dropped experiments in which the sample size in either condition was less than $5(k + 2) + 1$, where $k = 9$ is the number of predictors, which would allow for at least 5 observations per predictor in any model. In the subgroup analyses of Section 6, we analyzed subgroups with smaller sample sizes.

These exclusions left a total of 227 randomized contrasts—pairs of treatment conditions between which students were randomly assigned—drawn from 68 separate A/B tests.

3.1. Data Collection

The data was collected from ASSISTments in two sets: remnant data and experiment data. Remnant data was used to train the imputation models, and experiment data was used to impute the outcomes in each experiment using the imputation models. The skill builders started by the students in the remnant data were not the same skill builders as the experimental skill builders in the experiment data, nor is there any overlap in students between the two datasets. **No information from the students or skill builders in the experiment data was in the remnant data used to train the imputation models**.

For both the remnant and experiment data, the same information was collected. For each instance of a student starting a skill builder for the first time, we collected data on whether they completed the skill builder, and if so, how many problems they had to complete before mastering the material. The imputation models, discussed more in section 4, were trained to predict these two dependent measures. The data used to predict these dependent measures was aggregated from all of the previous work done by the student. Three different sets of data were collected for each sample in the datasets: prior student statistics, prior assignment statistics, and prior daily actions. Prior student statistics included the past performance of each student, for example, their prior percent correct, prior time on task, and prior assignment completion percentage. Prior assignment statistics were aggregated for each assignment the student started prior to the skill builder. Prior assignment statistics included things like the skill builders' unique identifiers (or in the remnant data, the ID of the experimental version of a skill builder, if it existed), how many problems had to be completed in the assignment, students' percent correct on the assignment, and how many separate sessions students used to complete the assignment. Prior daily actions contained the total number of times students performed each possible action in the ASSISTments tutor for each day prior to the day they started the skill builder. The possible actions included things like starting a problem, completing an assignment, answering a problem, and requesting support. Complete lists of features included in prior student, assignment, and daily action datasets are included in Tables 5, 6, and 7 in the appendix. 193,218 sets of prior statistics on students, 837,409 sets of statistics on prior assignments, and 695,869 days of students' actions were aggregated for the remnant data, and 113,963 sets of prior statistics on students, 2,663,421 sets of statistics on prior assignments, and 926,486 days of students' actions were aggregated for the experiment data.

4. Remnant-Trained Imputation Models

4.1. Model Design

Each of the three types of data in the remnant dataset was used to predict both skill builder completion and the number of problems completed prior to demonstrating mastery. For each type of data: prior student statistics, prior assignment statistics, and prior daily actions, a separate neural network was trained. Additionally, a fourth neural network was trained using a combination of the previous three models. The prior student statistics model, shown in Figure 2 in red was a simple feed-forward network with a single hidden layer of nodes using sigmoid activation and dropout. Both the prior assignment statistics model and the prior daily actions model, shown in Figure 2 in blue and yellow respectively, were recurrent neural networks with a single hidden layer of LSTM nodes (Gers et al., 2000) with both layer-to-layer and recurrent dropout. The prior assignment statistics model used the last 20 started assignments as input, and

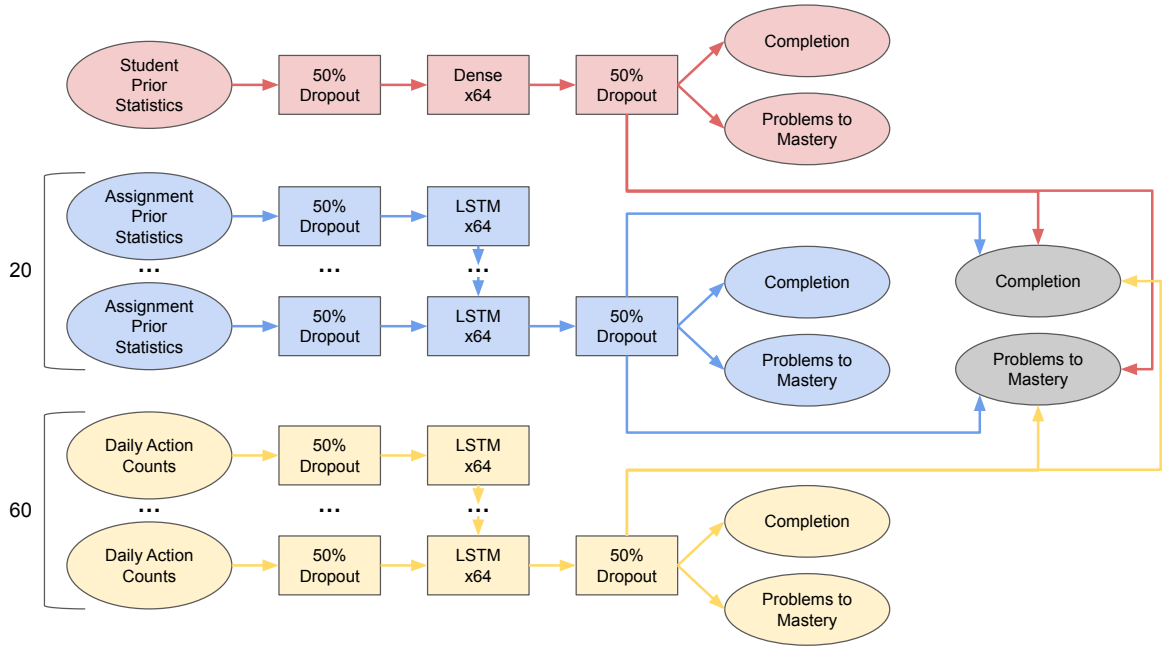

Figure 2: All four of the imputation models in one. The red model predicts performance using only prior statistics of the student, the blue model uses statistics on the last 20 assignments completed by the student to predict performance, and the yellow model uses the last 60 days of actions the student took in the tutor. The combined model, shown in grey, uses all three models to predict performance.

the prior daily actions model used the last 60 days of actions as input. The last 20 assignments were chosen because of success in prior work with similar numbers of prior assignments (Sales et al., 2018), and the last 60 days of actions were chosen based on usage data that indicated that after two months, students are unlikely to be working on content that is relevant for predicting their assignment completion. The combined model in Figure 2 takes the three models above and couples their predictions such that the prediction is a function of all three models' weights and the loss is backpropagated through each model during training. The hyperparameters of the model, including the dropout frequency, layer depth, and the number of nodes in each layer, were determined via grid search prior to using the model in the ReLOOP process. This model was chosen because this set of hyperparameters led to the lowest loss when predicting a held-out subset of the data.

Dropout was used to regularize model training but was not used in model validation, testing, or prediction.

4.2. Model Training

To select the best model hyperparameters and to measure the quality of each imputation model, 5-fold cross-validation was used to train and calculate various metrics for each model. For all training, the ADAM method (Kingma and Ba, 2015) was used during backpropagation, binary cross-entropy loss was used for predicting completion, and mean squared error loss was used for problems to mastery. The total loss for each model was the sum of the two individual losses.

Table 2: Metrics Calculated from 5-Fold Cross Validation for each Model.

Metric	Prior Student Statistics	Prior Assignment Statistics	Prior Daily Action Counts	Combined
Completion AUC	0.743	0.755	0.658	**0.770**
Completion Accuracy	0.761	0.767	0.743	**0.774**
Completion R^2	0.143	0.161	0.045	**0.184**
# of Problems MSE	8.489	8.505	8.719	**8.363**
# of Problems R^2	0.033	0.032	0.007	**0.048**

Because mean squared error and binary cross-entropy have different scales, a gain of 16 was applied to the binary cross-entropy loss, which brought the loss into the same range as the mean squared error loss for this particular dataset. The gain of 16 was determined via grid search based on which gain led to the most accurate completion predictions during cross-validation because assignment completion was the outcome of interest for experiment analysis. Table 2 shows various metrics of the models' quality. Interestingly, even though all the models are bad at predicting problems to mastery, removing problems to mastery from the loss function reduced the models' ability to predict completion.

Based on Table 2, statistics on prior assignments were the most predictive of students' assignment performance, followed by the students' overall prior performance statistics, and then their daily action history, which was the least predictive of their performance on their next assignment. Combining these datasets together led to predictions of a higher quality than any individual dataset could achieve.

The effort we put into optimizing the model likely contributed to our methods' successes. However, our methods do not assume that the imputation model is optimal, accurate, or correct in any sense. A well-fitting model will lead to precise effect estimates, but estimates using a poorly-fitting model will still be unbiased, and their associated statistical inference will still be valid.

5. Research Question 1: Can Imputations from Remnant-Trained Models Improve Standard Errors for Average Effects?

To gauge the potential of remnant-based imputations to improve the precision of impact estimates, we compared estimated sampling variances from the four different treatment effect estimators listed in Section 2.3.5: T-Test ($\hat{\tau}^{DM}$), which includes no covariate adjustment; LOOP, which uses random forests for within-sample covariate adjustment using only the nine student-aggregated covariates in Section 3 but not the remnant; ReLOOP, which uses remnant-based imputations $\hat{y}_i^r$ in a within-sample OLS adjustment model; and ReLOOP+, which uses an ensemble algorithm to adjust for both $\hat{y}_i^r$ and the nine student-aggregate covariates in LOOP. In this analysis, we used the "combined" model, including all available remnant data, to generate remnant-based imputations $\hat{y}_i^r$. We used these four estimators to estimate $\overline{\tau}_{RCT}$ in each of the 227 randomized contrasts described above.

Figure 3 shows the ratios of estimated sampling variances from the four estimators. Since sampling variance scales as $1/n$, ratios of sampling variances can be thought of as "sample size multipliers"—that is, decreasing the variance by a factor of q is analogous to increasing the

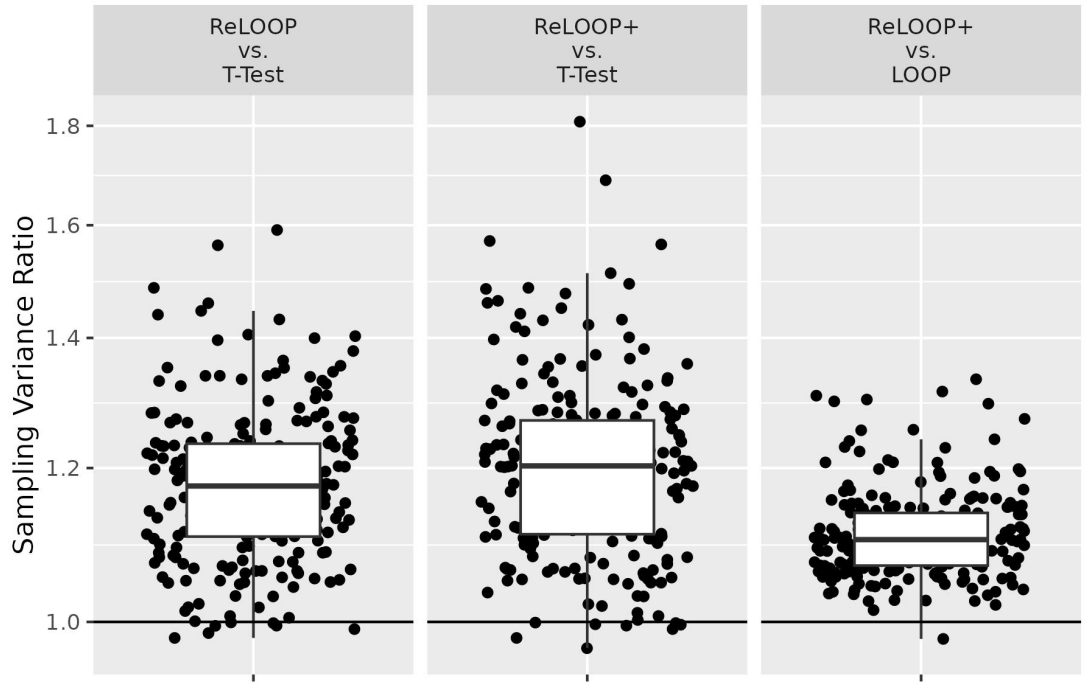

Figure 3: Boxplots and jittered scatter plots of the ratios of estimated sampling variances of $\hat{\tau}^{DM}$ (i.e. "T-Test," which includes no covariate adjustment), $\hat{\tau}_{LOOP}(\boldsymbol{x})$ ("LOOP", which adjusts for covariates within sample, but does not use the remnant), $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}})$ ("RELOOP," which adjusts for remnant-based imputations but not within-sample covariates), and $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$ ("RELOOP+," which adjusts for both within-sample covariates and remnant-based imputations) in 227 randomized contrasts. The Y-axis is on a logarithmic scale, so that, say, doubling the sample size appears as the same magnitude of an effect as halving the sample size.

sample size by the same factor. The results in Figure 3 were previously reported in a conference poster (Sales et al., 2022).

The panel on the left of 3 compares $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}})$ to $\hat{\tau}^{DM}$, the T-Test estimator. In nearly every case the estimator using remnant data substantially outperformed the t-test estimator. In the majority of cases, including remnant-based predictions was roughly equivalent to increasing the sample by between 15 and 60%. The middle panel of Figure 3 compares $\hat{\tau}^{DM}$ to $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$. Here the results are slightly more impressive than those of the left panel—the median improvement is equivalent to increasing the sample size by about 20%, and in the best case the improvement is equivalent to an 80% increase in sample size.

The rightmost panel of Figure 3 compares $\hat{\tau}_{LOOP}(\boldsymbol{x})$, which uses leave-one-out sample splitting and a random forest to adjust for covariates—but does not use the remnant—to $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$ which does. In this case, we see more modest relative gains, which is to be expected since $\hat{\tau}_{LOOP}(\boldsymbol{x})$ can accomplish a good deal of covariate adjustment using only experimental data. Nevertheless, the contribution of the remnant is still significant—in roughly half of cases, including data from the remnant was equivalent to increasing the sample size by about 10–20%, and in a handful of cases the improvement was closer to 30%.

In summary, covariate adjustment can lead to substantial gains in precision, with the greatest improvement resulting from adjustment using both within-sample aggregated covariates and

remnant-based imputations. In particular, estimators including remnant-based imputations consistently outperformed those that used only within-sample covariate adjustment.

5.1. Did the Remnant Help Us Discover any Effects?

Researchers may naturally want to know if our claim to increase the power of A/B tests to detect effects actually lead, in practice, to more effects detected. In other words, did covariate adjustment lead to any p-values dipping below the $\alpha = 0.05$ threshold? Counting significant p-values is a problematic approach to gauging the success of our method since it depends on the size of the true effects. In particular, if the true $\overline{\tau}_{RCT}$ is equal to 0, then a p-value less than 0.05 would be a type-I error, but if the $\overline{\tau}_{RCT}$ is not equal to 0, a p-value less than 0.05 would be a true discovery. Since the ground truth is unknown, we cannot know which one is the case.

Table 3: The number of p-values less than $\alpha = 0.05$ using each of the four estimators. The table counts significant p-values unadjusted for multiple comparisons, and adjusted with the Benjamini-Hochberg and Benjamini-Yekutieli procedures.

	T-Test	LOOP	ReLOOP	ReLOOP+
Unadjusted	38	39	41	41
Benjamini-Hochberg	3	11	8	10
Benjamini-Yekutieli	2	2	2	2

Nevertheless, we will press on. Table 3 gives the count of significant p-values using each of the four estimates. The first row gives a count of unadjusted p-values; if each pairwise comparison were considered in isolation, these would be the relevant counts. A researcher using T-Tests would report discoveries in 38 cases, researchers using LOOP would report one additional discovery, and those using ReLOOP or ReLOOP+ would report an additional 3 discoveries. However, since there were 227 total hypothesis tests, even if the null hypothesis were true in every case we would expect around 11 significant p-values; in other words, since we are considering the p-values as a group a multiplicity adjustment is in order. We considered two adjustment methods, both designed to limit the "false discovery rate"—the proportion of the discoveries that are, in fact, type-I errors—to 5%. The second row of Table 3 counts p-values adjusted with the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). This procedure is guaranteed to control the false discovery rate only if the tests are independent[3]. The pairwise comparisons we consider are not independent, since each A/B test may have contributed several pairwise comparisons, which share data. After Benjamini-Hochberg adjustment, a researcher using T-Tests would only discover 3 effects, while researchers using LOOP would discover 11, those using ReLOOP would discover 8, and those combining within-sample and remnant-based adjustments with ReLOOP+ would lead to two additional discoveries or 10 total.

It may be surprising that although ReLOOP+ standard errors tend to be smaller than those from LOOP, LOOP leads to one more discovery than ReLOOP+. In fact, there were two cases in which LOOP p-values were significant after Benjamini-Hochberg adjustment, but ReLOOP+ p-values were not. In both cases, ReLOOP+ standard errors were lower, but ReLOOP+ estimates were closer to zero as well. There was one case in which ReLOOP+ led to a significant p-value

[3]There are some types of dependence which are OK, too, but they are difficult to describe, much less to verify.

but LOOP did not; in this case, RELOOP+ returned a smaller standard error and an effect estimate larger magnitude than LOOP.

The third row of the table counts significant p-values adjusted by the more conservative Benjamini-Yekutieli procedure (Benjamini and Yekutieli, 2001), which controls the false discovery rate even under arbitrary dependence of tests. Researchers using any of the four estimators we've considered and adjusting with the Benjamini-Yekutieli procedure would all reject 2 null hypotheses among the 227 possibilities.

5.2. WHICH REMNANT DATA HELPS THE MOST?

Figure 4 expands on figure 3 by contrasting the performance of RELOOP and RELOOP+, relative to T-Tests and LOOP, using remnant-based imputation models trained using different types of remnant data. As described above, the "action" model uses data on each student's daily actions in ASSISTments leading up to the A/B test, the "student" model used student-aggregated performance metrics prior to the beginning of the A/B test, and the "assignment" model used student performance metrics on previous assignments or skill-builders each student had worked on. Finally, the "combined" model—also shown above, in Figure 3—was an ensemble of the action, student, and assignment models. By examining the performance of each separate model, we can get a sense of the relative contribution of each type of remnant data to RELOOP or RELOOP+'s performance.

Comparing across models fit in the remnant, the action-level model performed the worst, while the combined model was responsible for the greatest decrease in sampling variance. Interestingly, the assignment-level model performed nearly as well as the combined model, suggesting that action- and student-level data did not contribute substantially. This pattern is consistent across the three different comparisons shown, comparing RELOOP and RELOOP+ to T-Tests, and comparing RELOOP+ to LOOP.

6. RESEARCH QUESTION 2: RELOOP FOR SUBGROUP EFFECTS

To judge RELOOP's potential for improving (or worsening) precision in subgroup effect estimates, we created subgroups using each of the nine student-aggregated covariates available for each of the randomized comparisons we considered. Specifically, we first pooled each of the 9 covariates $\boldsymbol{x}_k$, $k = 1, \ldots, 9$ across all of the 227 pairwise comparisons, and calculated the 1/3 and 2/3 quantiles, $q_{1/3}(\boldsymbol{x}_k)$ and $q_{2/3}(\boldsymbol{x}_k)$. Then, for each contrast and each covariate x, we identified students with x values that were "low" ($x_{ik} < q_{1/3}(\boldsymbol{x}_k)$) or "high" ($x_{ik} > q_{2/3}(\boldsymbol{x}_k)$). Finally, using each of the four estimators described in the previous section, for each pairwise contrast and for each covariate, we estimated two effects: one for low students and one for high.

In addition to the nine within-sample covariates, we also looked for effects in subgroups defined by the remnant-based imputations themselves—that is, students with a high or low probability of completing their assignment, using the remnant-based model.

All told, this should have resulted in $227 \times 10 \times 2 = 4,540$ estimates for each of the four estimators. In practice, we did not estimate effects if either treatment arm within a subgroup had fewer than 10 subjects, which excluded 210 of these comparisons, and we encountered other estimation problems (such as the lack of variance in outcomes) in 12 others, leaving a total of 4,318 random comparisons to consider. Now, these 4,318 comparisons are by no means independent—they represent different ways to slice the data from the original 68 A/B tests.

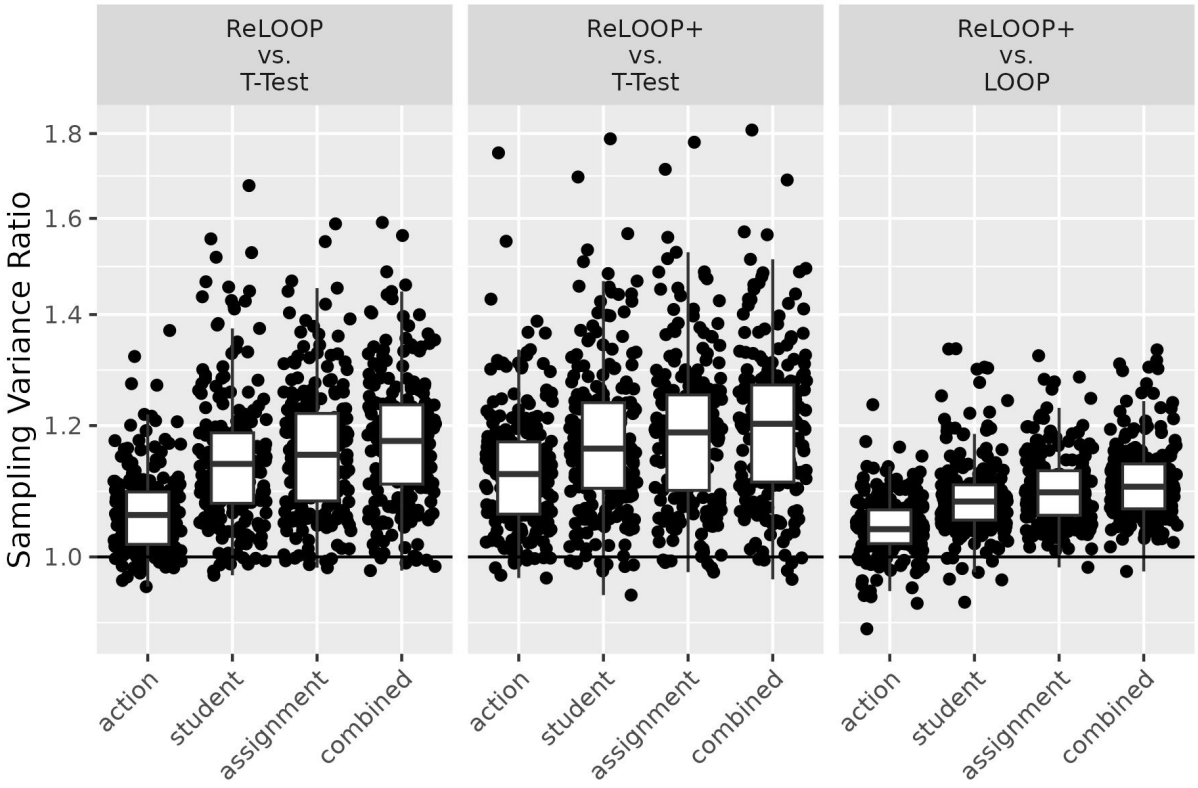

Figure 4: Boxplots and jittered scatter plots of the ratios of estimated sampling variances of $\hat{\tau}^{DM}$ (i.e. "T-Test," which includes no covariate adjustment), $\hat{\tau}_{LOOP}(\boldsymbol{x})$ ("LOOP", which adjusts for covariates within sample, but does not use the remnant), $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}})$ ("ReLOOP," which adjusts for remnant-based imputations but not within-sample covariates), and $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$ ("ReLOOP+," which adjusts for both within-sample covariates and remnant-based imputations) in 227 randomized contrasts. The Y-axis is on a logarithmic scale.

Nevertheless, by considering them all we may be able to discern some patterns in RELOOP's effectiveness in improving precision.

First, though, Figure 5 shows sampling variance ratios pooled across all A/B tests, pairwise comparisons, and subgroups. For the first time, we see some cases of covariate adjustment substantially harming the precision of effect estimates—RELOOP gave larger standard errors than T-Tests in about 11% of cases, RELOOP+ gave larger standard errors than T-Tests in around 12% of cases and RELOOP+ gave larger standard errors than LOOP in about 13% of cases. In the vast majority of these cases the effect was comparable to decreasing the sample size by less than 10%, but in about 3% of cases using RELOOP or RELOOP+ instead of T-Tests or LOOP was equivalent to decreasing the sample size by 10% or more, and in a handful of cases the decrease was even larger, up to nearly 50%.

Still, in the majority of cases remnant-based covariate adjustment improved the precision of impact estimates, sometimes by dramatic amounts. For all three comparisons shown in the figure, the median sampling variance ratio was greater than 1.1, meaning that RELOOP or RELOOP+ was equivalent to increasing the sample size by more than 10% at least half the time. Much more dramatic improvements were also common: in 25% of cases, RELOOP outperformed the T-Test by 22% or more, RELOOP+ outperformed the T-Test by 25% or more, and RELOOP+ outperformed LOOP by at least 18%. In some extreme cases, the improvement due to RELOOP or RELOOP+ was equivalent to doubling or tripling the sample size, and in one case, it was equivalent to multiplying the sample size by more than five.

Echoing the analysis in Section 5.1, Table 4 shows the number of discoveries—i.e. $p < 0.05$—a researcher would make using each of the three estimators. If p-values are not adjusted for multiple comparisons, a researcher using RELOOP or RELOOP+ would reject 16 or 18 more null hypotheses, respectively, than a researcher using LOOP, and 44 or 46 more than a researcher using T-Tests. If p-values are adjusted with the Bejamini-Hochberg procedure, a researcher using T-Tests would fail to reject every one of the 4,318 null hypotheses, while one using LOOP would reject 23, one using RELOOP would reject 17, and a researcher using RELOOP+ would reject 25, ensuring tenure and grant funding. After adjusting with the Benjamini-Yekutieli procedure, researchers using LOOP would report two discoveries, and those using RELOOP or RELOOP+ would report eight.

Table 4: The number of p-values less than $\alpha = 0.05$ using each of the four estimators. The table counts significant p-values unadjusted for multiple comparisons, and adjusted with the Benjamini-Hochberg and Benjamini-Yekutieli procedures.

	T-Test	LOOP	ReLOOP	ReLOOP+
Unadjusted	375	403	419	421
Benjamini-Hochberg	0	23	17	25
Benjamini-Yekutieli	0	2	8	8

The following two subsections dig deeper into these varying effects by looking at subgroup effects broken down by subgroup and as a function of sample size.

6.1. SUBGROUP EFFECT STANDARD ERRORS BY COVARIATE

Figure 6 shows boxplots of sampling variance ratios comparing RELOOP to T-Tests and RELOOP+ to LOOP for each subgroup we considered. A few features are apparent. First, RELOOP performs

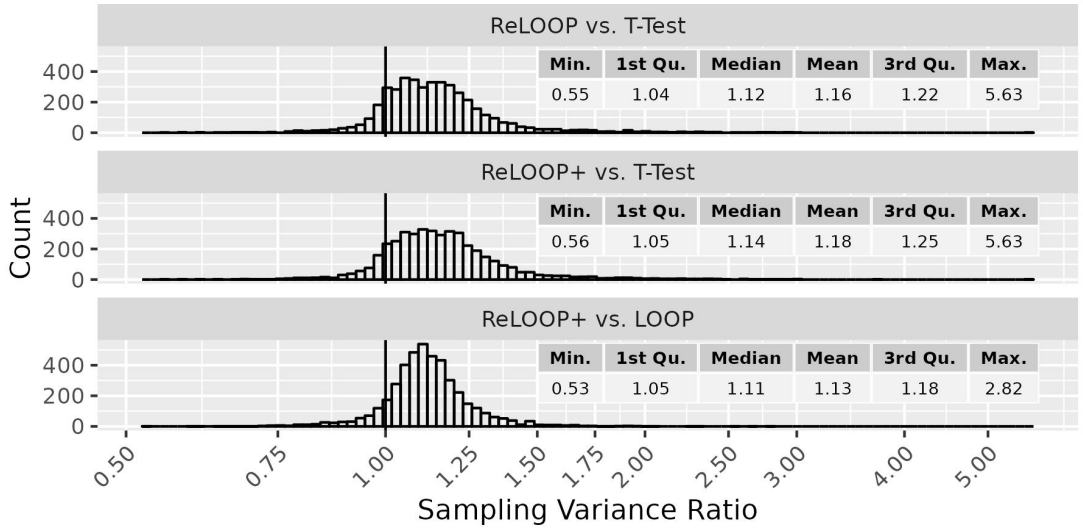

Figure 5: Histograms of the ratios of sampling variances of $\hat{\tau}^{DM}$ (T-Tests), $\hat{\tau}_{LOOP}(\boldsymbol{x})$ (LOOP), $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}})$ (ReLOOP), and $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$ (ReLOOP+) for 4,318 estimated subgroup effects. Sample statistics of the distributions of ratios are also shown. The X-axis is on logarithmic scale.

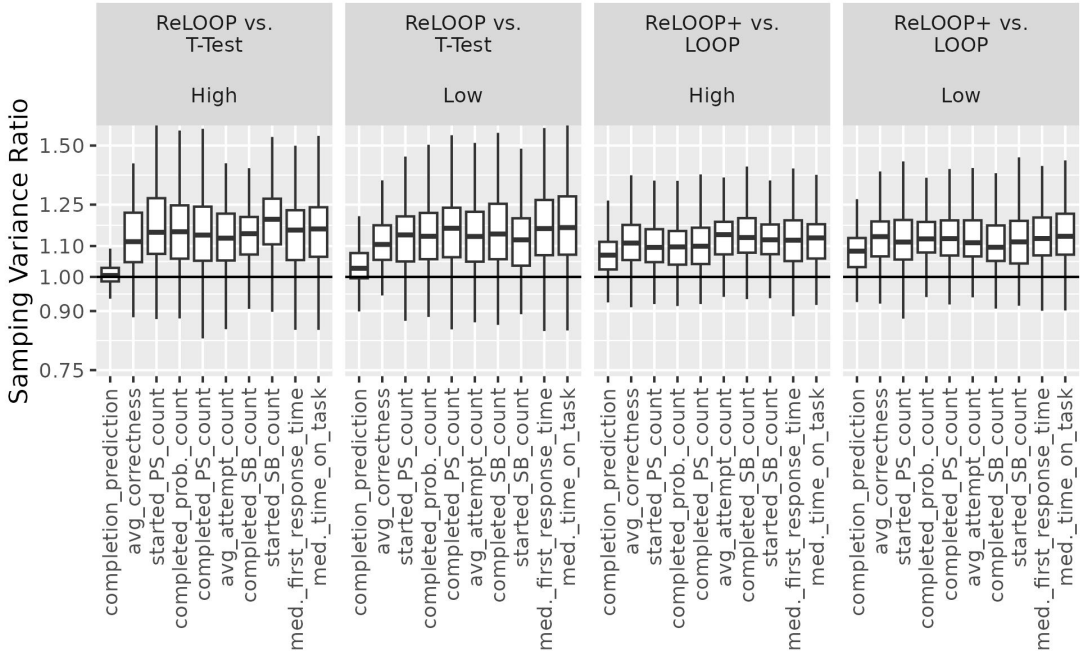

Figure 6: Boxplots of the ratios of sampling variances of $\hat{\tau}^{DM}$ (T-Tests), $\hat{\tau}_{LOOP}(\boldsymbol{x})$ (LOOP), $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}})$ (ReLOOP), and $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$ (ReLOOP+) for each subgroup considered. Outliers are omitted. The Y-axis is on a logarithmic scale.

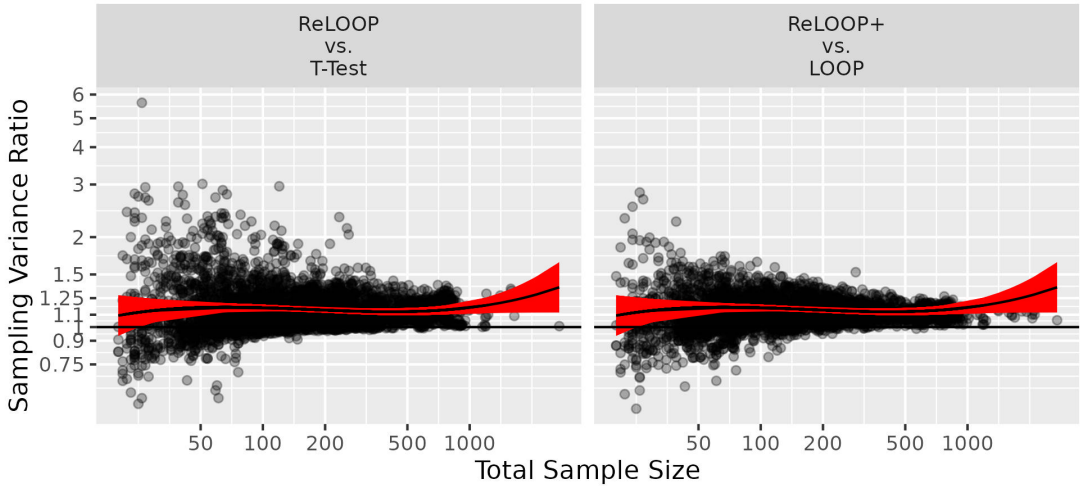

Figure 7: The ratios of sampling variances of $\hat{\tau}^{DM}$ (T-Tests) to $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}})$ (ReLOOP) and $\hat{\tau}_{LOOP}(\boldsymbol{x})$ (LOOP) to $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$ (ReLOOP+) as the total sample size of the subgroup varies. The X- and Y-axes are on a logarithmic scale.

no better than T-Tests for the high `completion_prediction` subgroup, and little better than T-Tests for the low `completion_prediction` subgroup. These are the subgroups defined based on $\hat{y}_i^r$; since the variance of $\hat{y}_i^r$ is, by definition, lower in these subgroups than in the sample as a whole, there is less opportunity to use it for variance reduction.

Aside from those defined based on `completion_prediction`, there was little difference in ReLOOP's effectiveness between subgroups. In every case, the lower quartile was greater than 1, though the lower tail reached below 1. For comparisons between ReLOOP and T-Tests, the median ratio was between 1.1 and 1.2, while for ReLOOP+/LOOP comparisons, the medians were somewhat lower.

Figure 7 plots the sampling variance ratios comparing ReLOOP to T-Tests and ReLOOP+ to LOOP against each subgroup's sample size. A semi-parametric regression fit (the natural logarithm of the sampling variance ratio regressed on a b-spline of the log of sample size with four degrees of freedom) is plotted over the points. The standard error shown is adjusted for the correlation of ratios from the same experiment. There is little evidence of a trend in the mean improvement due to ReLOOP—instead, it appears fairly constant as the sample size varies. On the other hand, the range and spread of ratios decrease markedly as sample sizes increase. Every case in which ReLOOP hurt the precision relative to T-Tests by more than 10% was in a subgroup with $n < 80$, as were all but one of the cases when ReLOOP adjustment was equivalent to multiplying the sample size by 2.5 or higher, relative to T-Tests. Apparently ReLOOP's greatest potential for radically improving statistical precision occurs in relatively small samples. On the other hand, in relatively small samples the asymptotic guarantee that ReLOOP cannot increase estimated sampling variance apparently does not hold consistently.

7. Research Question 3: RELOOP with an Unrepresentative Remnant

Previous sections illustrated the potential for a model fit in the remnant to improve the precision of treatment effect estimates in A/B tests, without assuming that both datasets were drawn from the same population. However, in previous examples, it was not always entirely clear in what way the data from the remnant may or may not have been representative of RCT data. In this section, we examine a case where the remnant is primarily composed of one demographic subgroup, while the RCT is a mix of subgroups.

In particular, we describe an experiment in which we intentionally designed the remnant to differ from the RCT, in order to investigate the impact remnant unrepresentativeness may have on RELOOP or RELOOP+'s ability to improve statistical precision.

The experiment builds on the analyses of previous sections. However, to illustrate the effects of a remnant that is not representative of the RCT, we re-trained $\hat{y}^{REM}(0)(\cdot;\boldsymbol{\beta})$ using a subset composed disproportionally (though not entirely) of white and Asian males, and examined the estimated sampling variance of the RELOOP+ estimator for the entire RCT, for a similarly-composed subset, and for that subset's complement.

7.1. "Inferred Gender"

To help maintain students' privacy, ASSISTments does not gather data on student demographics. However, the ASSISTments Foundation gathers (but does not publish) students' names, to facilitate classroom instruction (teachers need to know which student's assignment they are grading). For some analyses on ASSISTments data, analysts will attempt to guess a student's gender identification based on that student's name. To do so, the Python package "gender-guesser"[4] was given each student's first name. The gender-guesser package uses a library of names and a script released by the German tech magazine, Heise, to determine which gender a name is associated with based on input from native speakers of various European and Asian languages. The script categorizes a name as being male, female, mostly male, mostly female, androgynous, or unknown if the name is not in the library. Clearly, this process is faulty and inexact. That being said, there is good reason to believe that most students who are inferred to be male or mostly male are male, and most inferred to be female or mostly female are female.

There is also reason to believe that the "unknown" category has a higher proportion of non-Asian racial or ethnic minorities or immigrants than the inferred male or female categories. This claim follows from the assumption that names that are not in the library are uncommon and that uncommon names are probably most common among populations with non-European or non-Asian language traditions (including immigrants and native speakers with non-European or non-Asian cultural traditions) and African Americans since there is a long tradition of distinctive naming in the African American community (Cook et al., 2014).

It follows that while the set of students labeled "Male" or "mostly male" includes students with diverse genders, ethnicities, and linguistic traditions, it includes a disproportionate number of white and Asian males. In this way, this set of students follows an unfortunate, though common, pattern of disproportionately white male training sets for machine learning algorithms (Denton et al., 2020).

[4] https://pypi.org/project/gender-guesser/

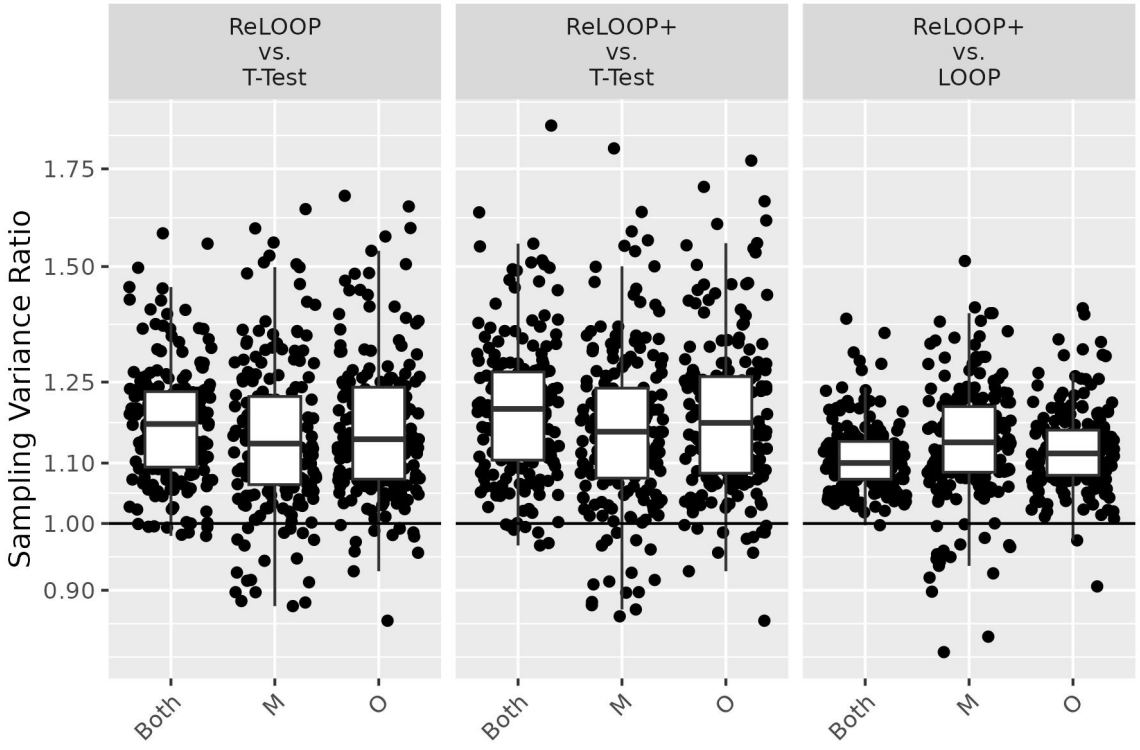

Figure 8: Results comparing estimators using imputations from the remnant, $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}})$ or ReLOOP+ (with or without other covariates), to estimators that do not, $\hat{\tau}^{DM}$ and $\hat{\tau}_{LOOP}(\boldsymbol{x})$. For all analyses, the remnant was composed of only students whose inferred gender was male; imputations from a model trained on the male remnant were used to analyze A/B tests including all participants ("Both"), or just inferred male ("M") or inferred non-male ("O").

To demonstrate the ability of the ReLOOP+ estimator to estimate internally-unbiased causal effects, even when the remnant reflects common biases in training datasets, we artificially limited the remnant to students labeled "Male" or "mostly male". Then, we estimated three sets of effects: one in which the RCT was limited in the same way as the remnant—i.e. to students labeled as male—another in which only the students who would be excluded from the remnant—those not labeled male—and the complete RCT data.

7.1.1. Results

Using the predictions from the model described above, we estimated $\overline{\tau}_{RCT}$ for each experimental contrast in the same four ways as in the previous sections: with T-Tests $\hat{\tau}^{DM}$ (i.e. no covariate adjustment), with LOOP $\hat{\tau}_{LOOP}(\boldsymbol{x})$ (i.e. only within-RCT adjustment), with ReLOOP $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}})$, an estimator using *only* predictions from the remnant for covariate adjustment, and with $\hat{\tau}_{LOOP}(\hat{\boldsymbol{y}}^{\boldsymbol{r}}, \boldsymbol{x})$ ReLOOP+, which uses both aggregated student-level covariates and the predictions from the remnant.

Figure 8 shows the results comparing estimators that use imputations from the remnant to those that do not. Both estimators ReLOOP and ReLOOP+ are almost always similar to or more precise than the T-Test estimator $\hat{\tau}^{DM}$. The only exception is a handful of cases in which including remnant imputations is equivalent to decreasing the sample size by 15% or less. This

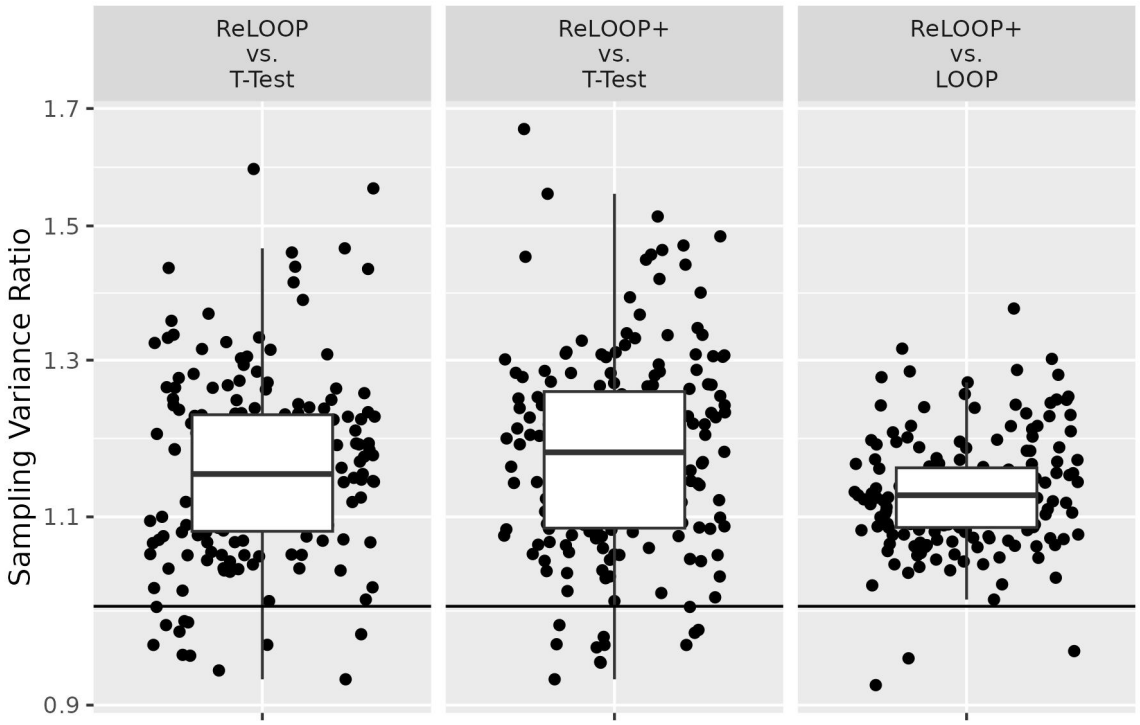

Figure 9: Results comparing post-stratification estimators using imputations from the remnants ReLOOP or ReLOOP+ (with or without other covariates) to estimators that do not, $\hat{\tau}^{DM}$ and $\hat{\tau}_{LOOP}(\boldsymbol{x})$.

is mostly due to very small samples in some RCTs. On the other hand, in most of the RCTs, the improvement was 10% or more, and in many it was upwards of 30%. Comparing ReLOOP+ to LOOP, including imputations from the biased remnant led to a 10% or higher increase in precision in most cases. Most surprisingly, the estimators performed as well or better in the non-Male sets and the full RCTs than in the Male subset. That is, using a model trained in a demographically distinct population did not reduce the method's effectiveness.

8. Research Question 4: reloop for Population Average Effects

Previous sections have focused on estimating $\overline{\tau}_{RCT}$, the average effect of a treatment for subjects in an RCT. However, often researchers are interested in $\mathbb{E}_{POP}[\tau]$, average effects across a wider population, POP.

To attempt to estimate $\mathbb{E}_{POP}[\tau]$, we conducted a post-stratification estimator (3) using the guessed gender predictor. While we do not observe the true distribution of guessed gender among all middle school ASSISTments users, we may estimate it from the remnant. When we do so, we find that roughly a third are labeled "Male."

We calculated four post-stratified estimators for each treatment contrast, using the four sets of $\overline{\tau}_G$ estimates. Then, as in $\overline{\tau}_{RCT}$ estimation, we gauged whether ReLOOP or ReLOOP+ improve the statistical precision of $\hat{\tau}^{DM}$ or $\hat{\tau}_{LOOP}(\boldsymbol{x})$.

Figure 9 contrasts sampling variances between post-stratification using auxiliary data versus only using RCT data. Indeed, including imputations from the remnant improves the precision of these estimators greatly.

While ReLOOP and ReLOOP+ are guaranteed to be internally unbiased for overall average effects $\overline{\tau}_{RCT}$ or subgroup effects $\overline{\tau}_{G=k}$, these guarantees do not extend to external bias (unless, of course, RCT is a random sample from POP). In particular, we cannot claim that the post-stratification estimates whose standard errors are represented in Figure 9 are unbiased for $\mathbb{E}_{POP}[\tau]$. However, if the effects of any of the interventions in the study vary by inferred gender, then the post-stratified estimates are likely to be less externally biased than $\overline{\tau}_{RCT}$ estimates. We have shown here that ReLOOP and ReLOOP+ can improve their precision, as well.

9. Discussion

Using remnant-trained models to predict A/B test outcomes, then using those predictions to estimate effects, has the potential to boost the precision of average effect estimators in education research. For typical analysis of A/B testing results, the use of remnant-based imputations could be equivalent to increasing the sample size by as much as 40-50% relative to t-tests and as much as 30% relative to state-of-the-art unbiased, covariate-adjusted effect estimators. Further, in the A/B tests we analyzed, incorporating remnant-based imputations never noticeably harmed precision.

The benefits of remnant-based predictions were even more pronounced in estimating subgroup effects and could be roughly equivalent to increasing the sample size by factors of 2, 3, or more. On the other hand, for subgroups with fewer than 100 students, there was a small risk that incorporating remnant-based predictions could harm precision instead of improving it.

The benefits of using the remnant appear to extend to cases in which the remnant does not resemble data from A/B tests on demographic characteristics. In fact, counterintuitively, we found greater benefits in the subgroup that was least represented in the remnant.

Finally, we found that incorporating remnant-based predictions into a post-stratification model can substantially improve post-stratified estimates, and hence help researchers generalize their findings to broader populations.

9.1. Limitations and Future Work

The methods discussed here are not a panacea. First of all, they do not apply in every randomized trial—in particular, large datasets including covariate and outcome data for non-participants are not always available. Furthermore, LOOP methods are only currently available for Bernoulli or pair-randomized RCTs (including cases in which subjects are randomized with different probabilities) but not for completely randomized or cluster-randomized designs, or for general blocked designs. We are currently working on extending LOOP—and hence ReLOOP—to these more complicated experimental designs, as well as to observational studies.

Secondly, the methods may require considerable resources to implement—specifically, gathering high-dimensional covariate data from the remnant and RCT participants and formulating, tuning, and training a predictive model are all tasks that can require time, computational resources, and expertise (that said, in other work, we have seen decent precision gains from out-of-the-box random forest models). These issues suggest the need for guidelines as to when

ReLOOP is likely to boost precision so that the gains in gathering and modeling remnant data will be worth the effort.

In particular, we suspect that the RCT sample size may play an important role in ReLOOP's effectiveness. Our results here suggest that the most dramatic gains from ReLOOP occur when the RCT sample size is below 100; however, in a handful of these cases, ReLOOP adjustment led to substantially higher standard errors than t-tests. Even when ReLOOP adjustment hurts precision, it does not cause bias; its associated statistical inference, such as confidence intervals and p-values, remain valid. Still, a method with lower chances of hurting precision, even when RCT sample sizes are small, may be desirable. Future research may show that simpler adjustment methods, such as ANCOVA, may pose lower risks in small-sample settings.

Prior theoretical results have shown that, regardless of the properties of an RCT, its remnant, or the imputation model, ReLOOP and ReLOOP+ cannot harm precision in large samples and that they are unbiased regardless of sample size—that is, they are unlikely to hurt an analysis. What this paper adds is that ReLOOP and ReLOOP+ can dramatically improve some analyses. However, these new results are necessarily limited to the analysis of A/B tests conducted on a computer-assisted learning program, which is far from the only causal analysis in educational data mining. The only way to conclusively demonstrate the broad applicability and usefulness of ReLOOP and ReLOOP+ is to implement them in a wide array of contexts, perhaps alongside other causal estimators.

10. Acknowledgements

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.

References

Aronow, P. M. and Middleton, J. A. 2013. A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference 1,* 1, 135–154.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological) 57,* 1, 289–300.

Benjamini, Y. and Yekutieli, D. 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics 29,* 4, 1165 – 1188.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21,* 1, C1–C68.

Cook, L. D., Logan, T. D., and Parman, J. M. 2014. Distinctively black names in the american past. *Explorations in Economic History 53*, 64–82.

Denton, E., Hanna, A., Amironesei, R., Smart, A., Nicole, H., and Scheuerman, M. K. 2020. Bringing the people back in: Contesting benchmark machine learning datasets. *Proceedings of ICML Workshop on Participatory Approaches to Machine Learning*.

Ding, P., Li, X., and Miratrix, L. W. 2017. Bridging finite and super population causal inference. *Journal of Causal Inference 5,* 2.

FISHER, R. A. 1935. *Design of experiments*. Oliver and Boyd, Edinburgh.

FREEDMAN, D. A. 2008. On regression adjustments to experimental data. *Advances in Applied Mathematics 40,* 2, 180–193.

GAGNON-BARTSCH, J. A., SALES, A. C., WU, E., BOTELHO, A. F., ERICKSON, J. A., MIRATRIX, L. W., AND HEFFERNAN, N. T. Forthcoming. Precise unbiased estimation in randomized experiments using auxiliary observational data. *Journal of Causal Inference*. https://arxiv.org/abs/2105.03529.

GERS, F. A., SCHMIDHUBER, J., AND CUMMINS, F. 2000. Learning to forget: Continual prediction with lstm. *Neural Computation 12,* 10, 2451–2471.

HARRISON, A., SMITH, H., HULSE, T., AND OTTMAR, E. R. 2020. Spacing out! manipulating spatial features in mathematical expressions affects performance. *Journal of Numerical Cognition 6,* 2, 186–203.

HEFFERNAN, N. T. AND HEFFERNAN, C. L. 2014. The assistments ecosystem: building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education 24,* 4, 470–497.

HELLER, R., ROSENBAUM, P. R., AND SMALL, D. S. 2009. Split samples and design sensitivity in observational studies. *Journal of the American Statistical Association 104,* 487, 1090–1101.

IMBENS, G. W. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics 86,* 1, 4–29.

KINGMA, D. P. AND BA, J. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds.

MCDERMOTT, R. 2011. Internal and external validity. In *Cambridge Handbook of Experimental Political Science*, J. N. Druckman, D. P. Greene, J. H. Kuklinski, and A. Lupia, Eds. Cambridge University Press, 27–40.

MIRATRIX, L. W., SEKHON, J. S., AND YU, B. 2013. Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75,* 2, 369–396.

NEYMAN, J. 1923. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science 5*, 463–480. 1990; transl. by D.M. Dabrowska and T.P. Speed.

OSTROW, K. S., SELENT, D., WANG, Y., VAN INWEGEN, E. G., HEFFERNAN, N. T., AND WILLIAMS, J. J. 2016. 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*. LAK '16. Association for Computing Machinery, New York, NY, USA, 279–288.

RUBIN, D. B. 1978. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics 6,* 1, 34 – 58.

SALES, A. C., BOTELHO, A., PATIKORN, T. M., AND HEFFERNAN, N. T. 2018. Using big data to sharpen design-based inference in a/b tests. In *Proceedings of the 11th International Conference on Educational Data Mining. (EDM 2018)*, K. E. Boyer and M. Yudelson, Eds. International Educational Data Mining Society, 479–486.

SALES, A. C., PRIHAR, E., GAGNON-BARTSCH, J., GURUNG, A., AND HEFFERNAN, N. T. 2022. More powerful a/b testing using auxiliary data and deep learning. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium: 23rd International Conference, AIED 2022, Durham, UK,*

July 27–31, 2022, Proceedings, Part II (AIED 2022), M. M. Rodrigo, N. Matsuda, A. I. Cristea, and V. Dimitrova, Eds. Springer Cham, 524–527.

SCHOCHET, P. Z. 2015. Statistical theory for the RCT-YES software: Design-based causal inference for RCTs. *U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance*.

VAN DER LAAN, M. J. AND ROSE, S. 2011. *Targeted learning: causal inference for observational and experimental data*. Springer Series in Statistics. Springer Science & Business Media, New York, NY.

WAGER, S., DU, W., TAYLOR, J., AND TIBSHIRANI, R. J. 2016. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences 113,* 45, 12673–12678.

WU, E. AND GAGNON-BARTSCH, J. A. 2018. The LOOP estimator: Adjusting for covariates in randomized experiments. *Evaluation Review 42,* 4, 458–488.

A. VARIABLES USED IN REMNANT IMPUTATION MODEL

Table 5: Prior Student Statistics Features.

Name	Description
target_sequence	The ID of the experimental skill builder
has_due_date	Whether the skill builder had a due date
assignments_started	The number of assignments previously started by the student
assignments_percent_completed	The number of assignments previously completed by the student
median_ln_assignment_time_on_task	The median of the log of the time between starting and finishing an assignment for all the students completed prior assignments
average_problems_per_assignment	The average number of problems completed by the student across all their previous assignments
median_ln_problem_time_on_task	The median of the log of the time the student took between starting and finished all their completed prior problems
median_ln_problem_first_response_time	The median of the log of the time the student took to submit their first answer or request tutoring across all their completed prior problems
average_problem_correctness	The fraction of previously completed problems the student got correct on their first attempt without tutoring
average_problem_attempt_count	The average number of attempts for all problems previously completed by the student
average_answer_first	The fraction of times the student submitted an answer before requesting tutoring for all problems previously completed by the student
average_problem_hint_count	The average number of hints requested for all problems previously completed by the student
skill_average_problems_per_assignment skill_median_ln_problem_time_on_task skill_median_ln_problem_first_response_time skill_average_problem_correctness skill_average_problem_attempt_count skill_average_answer_first	These features are the same as the features above with a similar name, but only calculate statistics across problems with the same skills as the problems in the experimental skill builder

Table 6: Prior Assignment Statistics Features.

Name	Description
id	The ID of the student
assignment_start_time	The UNIX time of when the assignment was started
directory_1	The highest level directory of the assignment location, usually an indication of curriculum
directory_2	The second level directory of the assignment location, usually an indication of grade level
directory_3	The third level directory of the assignment location, usually an indication of unit
sequence_id	The unique ID of the skill builder assignment, or the corresponding normal skill builder ID for experiments
is_skill_builder	Boolean flag for whether or not this assignment is a skill builder or a normal problem set
has_due_date	Boolean flag for if the assignment has a due date
assignment_completed	Boolean flag for if the student completed the assignment
time_since_last_assignment_start	The time between the student starting this assignment and starting their prior assignment
All Following Features	In addition to the raw value, a value z-scored across all students who completed the assignment previously, and a percentile across students in the same class who completed the assignment previously was included in the model as well.
session_count	How many times the student left and rejoined the assignment
day_count	How many days the student worked on the assignment for
completed_problem_count	How many problems the student completed in the assignment
median_ln_problem_time_on_task	The median of the log of the time between the student starting and finishing problems in the assignment
median_ln_problem_first_response	The median of the log of the time it took for the student to submit their first answer or request tutoring on the problems they started in the assignment
average_problem_attempt_count	The average number of attempts the student made on the problems in the assignment
average_problem_answer_first	The fraction of times the student made an attempt before requesting tutoring on all the problems in the assignment
average_problem_correctness	The fraction of times the student got the problem correct on their first try on all the problems in the assignment
average_problem_hint_count	The average number of hints used by the student on all the problems in the assignment
average_problem_answer_given	The fraction of times the student was given the answer on all the problems in the assignment

Table 7: Prior Daily Actions Features.

Name	Description
id	The ID of the student
timestamp	The UNIX time at 00:00:00 of the day the action counts apply to
ln_action_1_count	Log of the count of assignment started actions taken
ln_action_2_count	Log of the count of assignment resumed actions taken
ln_action_3_count	Log of the count of assignment finished actions taken
ln_action_4_count	Log of the count of problem set started actions taken
ln_action_5_count	Log of the count of problem set resumed actions taken
ln_action_6_count	Log of the count of problem set finished actions taken
ln_action_7_count	Log of the count of problem set mastered actions taken
ln_action_8_count	Log of the count of problem set exhausted actions taken
ln_action_9_count	Log of the count of problem limit exceeded actions taken
ln_action_10_count	Log of the count of problem started actions taken
ln_action_11_count	Log of the count of problem resumed actions taken
ln_action_12_count	Log of the count of problem finished actions taken
ln_action_13_count	Log of the count of tutoring set started actions taken
ln_action_15_count	Log of the count of tutoring set finished actions taken
ln_action_16_count	Log of the count of hint requested actions taken
ln_action_17_count	Log of the count of scaffolding requested actions taken
ln_action_19_count	Log of the count of explanation requested actions taken
ln_action_20a_count	Log of the count of student correct response actions taken
ln_action_20b_count	Log of the count of student incorrect response actions taken
ln_action_21_count	Log of the count of open response submission actions taken
ln_action_25_count	Log of the count of answer requested actions taken
ln_action_26_count	Log of the count of continue selected actions taken
ln_action_30_count	Log of the count of help requested actions taken
ln_action_31_count	Log of the count of timer started actions taken
ln_action_32_count	Log of the count of timer resumed actions taken
ln_action_33_count	Log of the count of timer paused actions taken
ln_action_34_count	Log of the count of timer finished actions taken
ln_action_35_count	Log of the count of live tutoring requested actions taken
Other Actions	Artifacts of the database, always 0