Using Publicly Available Auxiliary Data to Improve Precision of Treatment Effect Estimation in a Randomized Efficacy Trial

Charlotte Z. Mann
University of Michigan
Department of Statistics
323 West Hall
1085 S. University Ave
Ann Arbor, MI 48103
manncz@umich.edu

Jiaying Wang
University of Southern California Viterbi School of Engineering
3650 McClintock Ave
Los Angeles, CA 90089
jwang745@usc.edu

Adam Sales
Worcester Polytechnic Institute
Mathematical Sciences Department
27 Boynton St
Worcester, MA 01609
asales@wpi.edu

Johann A. Gagnon-Bartsch
University of Michigan
Department of Statistics
323 West Hall
1085 S. University Ave
Ann Arbor, MI 48103
johanngb@umich.edu

ABSTRACT

The gold-standard for evaluating the effect of an educational intervention on student outcomes is running a randomized controlled trial (RCT). However, RCTs may often be small due to logistical considerations, and resulting treatment effect estimates may lack precision. Recent methods improve experimental precision by incorporating information from large, observational, auxiliary data sets. Specifically, predictions of the outcome of interest from a model fit on the auxiliary data can be used in covariate adjustment. Such auxiliary data, on students or schools not included in an RCT, but with similar characteristics, is often available for educational RCTs. This is the case for a trial evaluating the efficacy of the Cognitive Tutor Algebra I curriculum (CTAI), an alternative algebra curriculum that included a computerized tutoring system. The Texas Education Agency (TEA) provides publicly available data on thousands of schools across Texas, including the 44 schools randomized in the CTAI study as well as nearly 3,000 additional, auxiliary schools. We develop an auxiliary model predicting passing rates for a standardized test in mathematics, which flexibly incorporates the 5,000 covariates available in the TEA data through random forest modeling. We compare our approach, using these auxiliary model predictions, to more standard estimators of the effect of CTAI on schools' mathematics passing rates. We find that leveraging information from the auxiliary data increases precision beyond standard methods that rely on the experimental sample alone, even for this paired trial with a powerful baseline covariate. We additionally demonstrate that working with auxiliary information provides practical benefits for analysis, beyond this increased estimation precision.

Keywords

causal inference, data fusion, potential outcomes, RCT, covariate adjustment

1. INTRODUCTION

Educational policy decisions rely on the evaluation of prospective educational interventions. Randomized controlled trials (RCTs) are considered the "gold-standard" for evaluating the effect of an intervention on educational outcomes because they support unbiased estimation. However, RCTs are often small due to cost and other logistical considerations, which can result in effect estimation that is not precise enough to support strong conclusions. This issue is exacerbated in educational experiments since the effects of educational interventions on common outcomes of interest (test scores, graduation rates, etc.) are often small [10, 4].

A common approach to improving precision of experimental estimates is using covariate adjustment to account for variation in the outcome that is not explained by the treatment. The precision gained through covariate adjustment depends on how predictive the covariates are of the outcome. Recent work has shown that additional precision may be gained by integrating auxiliary information into covariate adjustment.

There are a number of approaches for integrating auxiliary information in RCT analyses to improve precision [15, 24, 1, 3, 13, 7, 5]. We focus on an approach that uses auxiliary

C. Z. Mann, J. Wang, A. Sales, and J. A. Gagnon-Bartsch. Using publicly available auxiliary data to improve precision of treatment effect estimation in a randomized efficacy trial. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 518–525, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729874

data to construct a highly predictive covariate for the outcome of interest in the RCT [5] (hereafter, "data integration approach"). Consider a setting in which there is a large, observational dataset that describes schools or students not included in an RCT, but that have similar characteristics to the RCT sample ("auxiliary data"). We assume that an outcome of interest and covariates are available for both the auxiliary and RCT samples. Then, one can fit a model with the auxiliary data predicting the outcome of interest ("auxiliary model"). The predictions from this model, applied to the experimental sample, can be a powerful covariate for adjustment. The idea is that, since the auxiliary sample is typically much larger than the experimental sample, high-dimensional covariates can be accommodated and a model of the outcome using the auxiliary data will be more predictive of the outcome than a model based on the experimental sample alone.

We apply this data integration approach to analyze a randomized trial evaluating the efficacy of the Cognitive Tutor Algebra I (CTAI) curriculum [14]. The CTAI study was a paired trial that randomized a small number of schools within different states. We focus on study schools in Texas and are able to complement information on the trial schools with extensive publicly-available data for 3,000 other schools in the state. We first develop a model predicting schools' performance on a mathematics standardized test with these auxiliary schools. We combine the modern data integration approach [5] with another recently developed design-based estimator for paired trials [26], using predictions from the auxiliary model as a covariate to estimate the average effect of CTAI on school-level mathematics performance.

[5, 18] show that the data integration approach essentially cannot harm precision, so the question becomes whether the approach is worth the additional effort, compared to the possible precision gain, in different settings. While [5, 18] have shown the efficacy of the data integration approach across a number of relevant settings, the setting in this paper has not been previously evaluated with this method and is of interest to educational researchers. First, the CTAI study is a field experiment, while previous evaluations of the data integration method focus on EdTech A/B testing. Second, the CTAI study is a paired trial, where the schools were paired based on baseline characteristics, making it a powerful design, which should already improve precision over Bernoulli randomization. Finally, a pretest score – the performance on the outcome assessment of interest, before the treatment was applied – is available. The pretest score is often a highly prognostic covariate in educational trials, so one may not expect to be able to improve precision much beyond adjusting with the pretest alone. Thus, the CTAI study is a setting for which one may not expect data integration to provide considerable precision gains, and therefore would not be worth the additional effort. However, we find that incorporating auxiliary information does considerably improve precision beyond standard adjustment methods using the trial data alone, including adjusting with the pretest score.

Beyond the possible gain in precision, an advantage of leveraging auxiliary information in effect estimation is that any kind of pre-processing or modeling can be done with the auxiliary data, as long as the experimental outcomes are not touched. Thus, the auxiliary data can be thought of as a "sandbox," where the data may be explored without the risk of undermining the experimental design. In contrast, analyses of experimental data typically must be pre-specified, and should not be changed once the experimental data is in hand [2, 12]. We demonstrate this benefit of this flexibility with the auxiliary school data – evaluating different approaches to accounting for missing values, predictive algorithms, and covariate selection. We find that random forest models performed as well as more sophisticated machine learning algorithms and have the added benefits of being easy to fit and offering model interpretation. The auxiliary model itself provides interesting information about the relationship between predictors and the outcome. Thus, working with auxiliary data offers the additional benefit of complementing the RCT analysis with insights gained from an exploratory observational analysis.

Thus, this work complements that of [5, 18, 26], making two primary contributions. First, we demonstrate the effectiveness of their data integration approach in practice, in a new and relevant setting – a paired field trial with a highly predictive pre-treatment covariate. Second, we articulate in detail the process of implementing the data integration approach given a trial and relevant, publicly available, observational data, which highlights practical benefits to the approach beyond simply a gain in statistical efficiency.

This paper is organized as follows. Section 2 describes the CTAI study and the focus of our re-analysis of the study. Section 3 presents relevant notation, mode of inference, and causal estimators used in our analysis of the CTAI study. Section 4 details how we developed auxiliary models with publicly available data for use in covariate adjustment. Section 5 discusses the results comparing different covariate adjustment strategies for treatment effect estimation with the CTAI study. Section 6 concludes.

2. COGNITIVE TUTOR ALGEBRA I STUDY

In this section, we describe the CTAI study [14] and the analytical data available in more detail. CTAI, published by Carnegie Learning, Inc., was an alternative Algebra I curriculum which incorporated individualized student instruction through computerized tutoring. The study included 147 middle and high schools across seven states, which were pair matched within states based on school characteristics. Within each pair, schools were randomly assigned to implement CTAI or to continue Algebra I instruction as usual for two years (the 2007/08 and 2008/9 school years). See [14] for a detailed discussion of the design and implementation of the trial.

For the purposes of the current analysis, we focus on the 44 study schools in Texas, which include six pairs of high schools and 16 pairs of middle schools.[1] This focus on Texas schools is due to the availability of data on thousands of middle and high schools in Texas for years preceding and after

[1] The purpose of the re-analysis is to demonstrate the methods discussed and compare them to common methods, not to make any comparisons between our results and the original study analysis.

the study period. Specifically, the Texas Education Agency (TEA) published the Academic Excellence Indicator System (AEIS) from 2003 to 2011, which includes thousands of school-level measures each year, including information on student and teacher demographics, school finances, and student outcomes [21].

We estimate the average effect of CTAI on schools' passing rates for the 8th or 9th grade mathematics section of the Texas Assessment of Knowledge and Skills (TAKS) in 2008 [23]. TAKS was a standardized exam administered each year, with school-level results reported in the AEIS. Students were considered to have "met standards" or passed the math TAKS in 8th and 9th grade if they answered 60% or more of the 50 questions correctly [21].

The AEIS provides a rich source of information on the 44 schools in the CTAI study, as well as a large set of auxiliary schools, including a relevant educational outcome. Therefore, the CTAI study provides a nice example for incorporating auxiliary data in an RCT analysis in practice.

3. CAUSAL FRAMEWORK AND ESTIMATION

The CTAI study is a pair randomized experiment. In this trial design, units are matched into pairs, and within each pair one unit is randomly assigned to treatment and then the other is automatically assigned to control. Consider such a randomized experiment with N pairs of schools, indexed $i = 1, ..., N$ (resulting in $2N$ total schools in the experiment). Within each pair, we arbitrarily label the schools as the "first" or "second" school, indexed by $k = 1, 2$. We define a binary treatment assignment as $T_i = 1$ if the first school in pair i was assigned treatment (and therefore, the second control), and $T_i = 0$ if the first school was assigned control, where $P(T_i = 1) = \frac{1}{2}$.

The inference in this paper operates under the Neyman-Rubin model [19, 17] (also called the "potential outcomes framework"), a common, non-parametric causal model. The only source of stochasticity in this model comes from the randomized treatment assignment, whose distribution is known in an RCT. The analysis follows the approach of [26] and [5], which we summarize briefly in the remainder of this section.

We assume that each school has two, fixed potential outcomes, y^t_{ik} and y^c_{ik}, which would be observed if the school was assigned to the treatment or control group, respectively. Therefore, the observed outcome for the first school in pair i is $Y_{i1} = T_i y^t_{i1} + (1 - T_i) y^c_{i1}$, and for the second is $Y_{i2} = (1 - T_i) y^t_{i2} + T_i y^c_{i2}$. Define the average, within pair treatment effect as $\tau_i = \frac{1}{2}(y^t_{i1} - y^c_{i1} + y^t_{i2} - y^c_{i2})$. Our target estimand is the average treatment effect over all schools in the experimental sample (ATE), defined as:

$$\bar{\tau} = \frac{1}{N} \sum_{i=1}^{N} \tau_i$$

It is also useful to define two quantities: $v^{(1)}_i = y^t_{i1} - y^c_{i2}$ and $v^{(2)}_i = y^t_{i2} - y^c_{i1}$, so that $\tau_i = \frac{1}{2}(v^{(1)}_i + v^{(2)}_i)$. Also note that $v^{(1)}_i$ is observed if $T_i = 1$ and $v^{(2)}_i$ is observed if $T_i = 0$.

A common estimator for the ATE is the difference-in-means, or the difference in the mean outcome between the treatment and the control group. In a paired experiment the difference-in-means estimator can also be written as: $\hat{\tau}^{DM} = \frac{1}{N} \sum_{i=1}^{N} (2T_i - 1)(Y_{i1} - Y_{i2})$. [26] propose a design-based estimator for paired trials that incorporates covariate adjustment, typically improving precision over the difference-in-means estimator, without adding bias. Their estimator, which is related to previous work such as [1, 25, 16], is defined as follows:

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} (2T_i - 1)[(Y_{i1} - Y_{i2}) - \hat{d}_i],$$

where $d_i = \frac{1}{2}(v^{(1)}_i - v^{(2)}_i)$, and $\hat{d}_i$ is any estimate of that value. As long as $\hat{d}_i \perp T_i$, $\hat{\tau}$ is an unbiased estimator of $\bar{\tau}$. Therefore, the authors estimate $\hat{d}_i$ by sample-splitting [26, 1]. Specifically, they suggest fitting an imputation algorithm for d_i with all of the other pairs, excluding pair i. We will refer to $\hat{\tau}$ as the "sample-splitting estimator" for the remainder of the paper. Let $\boldsymbol{x}_i$ represent a vector of covariates for the two schools in pair i.[2] Denote the leave-one-pair-out imputation algorithm for pair i as $\hat{d}_{-i}(\cdot)$, so the estimate $\hat{d}_i = \hat{d}_{-i}(\boldsymbol{x}_i)$. While we represent $\hat{d}_{-i}(\cdot)$ as a single imputation algorithm, in practice it involves combining predictions of $v^{(1)}_i$ and $v^{(2)}_i$ (see [26]), which becomes relevant for variance estimation as described in Section 5.

The precision of the estimator depends on the mean squared error (MSE) of $\hat{d}_i$ – the smaller the MSE, the greater the precision. Incorporating information from a large auxiliary data source to fit $\hat{d}_{-i}(\cdot)$ can decrease the MSE of $\hat{d}_i$ beyond using the trial sample and covariates alone. Let $\hat{y}^r(\cdot)$ denote some prediction algorithm developed on the auxiliary data ("auxiliary model"), predicting the outcome of interest in the RCT with predictors that are available in the RCT. Then, one can generate predictions based on this auxiliary model, for the schools in the RCT, and treat these predictions $x^r_i = \hat{y}^r(\boldsymbol{x}_i)$ as a covariate to impute $\hat{d}_i$ [5]. Because $\hat{y}^r(\cdot)$ is fit on schools outside of the RCT sample and $\boldsymbol{x}_i$ are measured pre-treatment, x^r_i is also a pre-treatment covariate. Therefore, the auxiliary predictions can be used to generate imputation models, $\hat{d}_{-i}(\cdot)$, without introducing bias.

This data integration approach provides a researcher considerable flexibility. First, the properties of the sample-splitting estimator discussed do not depend on what kind of imputation algorithm is used for $\hat{d}_{-i}(\cdot)$ nor the auxiliary model, $\hat{y}^r(\cdot)$. It does not even require that the either model is correctly specified. Additionally, all that is required to develop an auxiliary model is the outcome of interest and covariates that are available for the RCT. Notably, the treatment need not be present in the auxiliary data. We take advantage of the flexibility allowed for auxiliary model development in our analysis of the CTAI study, as described in the following section.

4. DEVELOPING AN AUXILIARY MODEL

[2] $\boldsymbol{x}_i$ could be a vector that appends the covariates for both schools in a pair or represent a pair-level summary of the covariates. See [26] for a discussion.

The first step of the data integration approach [5] is developing a predictive model with the auxiliary data. Here, we discuss this process in detail, highlighting insights we expect will be useful to others developing similar models, and also presenting some interesting empirical findings. As mentioned previously, [5] make no assumptions about the auxiliary model, so we are free to do essentially anything we want with the auxiliary data, the only goal being to predict the outcome of interest accurately.

4.1 Auxiliary Texas School Data

The CTAI study included both middle and high schools because students take Algebra I across a range of grades. For the sake of this analysis, we assume that students take Algebra I in 8th grade (middle school) or 9th grade (high school). Therefore, we define high schools as those whose school type is labeled as "Secondary" or "Both" (meaning middle and high school) in the AEIS data and are not missing the 2008 9th grade math TAKS passing rate. We define all other schools for which the 2008 8th grade math TAKS passing rate is available as middle schools.[3] This results in 1,436 middle and 1,467 high schools for model development.

We use campus-level AEIS data from the 2003/4 through 2006/7 school years, including campus finance, staff, student, TAKS, and other performance data.[4] We additionally include variables that we believe were measured at baseline (pre-treatment) in the 2007/2008 school year including financial, staff, and student demographic data. We remove columns for which there is little variation between schools or for which more than 60% of the values are missing. There are 3,745 and 4,440 possible predictors for middle and high schools, respectively, after removing these columns.

Our outcome of interest is the 8th grade math TAKS passing rate for middle schools and the 9th grade math TAKS passing rate for high schools, which we will refer to collectively as the "math TAKS passing rate" for the remainder of the paper. There are relevant distinctions between the outcome for the auxiliary middle schools and high schools – the average 2008 TAKS passing rate for middle schools is around 79%, while it is 63% for high schools.

4.2 Model Development

As mentioned, our goal is to develop an auxiliary model that will predict the outcome of interest with high accuracy. Therefore, we evaluate the model using MSE and R^2 for development. Additionally, given that there are relevant distinctions between middle and high schools outcomes in the AEIS data, we develop an auxiliary model based on the performance of prediction algorithms fit separately on middle and high schools.

A common feature of publicly available data is the presence of data suppression or masking in order to preserve individual privacy. Indeed, the AEIS masks school-level assessment outcomes such as passing rates that are either based on fewer than 5 students, reveal that nearly all or no students passed, or are outside of a reasonable range for the measure [22]. We replace the masked indicators that either all or no students passed an assessment with 100 or 0, respectively. However, this masking still results in a substantial number of missing values. As mentioned previously, we only include schools which are not missing the 2008 math TAKS passing rate. We explore two routes for addressing missing data in the predictors: 1) simple mean imputation and 2) imputation using random forests (with the `missForest` package in `R` [20]). For both options, we additionally generate variables indicating whether a school's value was originally missing for each column. We evaluate these two approaches based on the MSE and R^2 of predictions resulting from fitting both random forests and neural nets for middle and high schools. We find that using random forest missing value imputation did not improve model performance meaningfully over simple mean imputation, while taking substantially more computing power and time. Therefore, after centering and scaling all variables (separately for middle and high schools), we replace all missing values with 0. Including the unique set of missing value indicators for each school type results in a total of 4,787 and 5,711 predictors for middle and high schools, respectively.

Table 1: Cross-validation (10-fold) mean squared error (MSE) and R^2 for models predicting the 2008 TAKS passing rate, fit on the auxiliary Texas schools with either (1) OLS with only the 2007 TAKS passing rate (pretest) or random forest with all available predictors. Models were fit separately on middle schools and high schools.

	Middle Schools		High Schools	
Model	**MSE**	$\boldsymbol{R^2}$	**MSE**	$\boldsymbol{R^2}$
Pretest (OLS)	85.3	0.54	206.5	0.56
All Predictors (RF)	65.6	0.65	128.2	0.72

Any modeling approach can be used for the auxiliary model, allowing for any kind of sophisticated, black-box modeling. We take advantage of this flexibility to evaluate different algorithms and subsets of predictors to predict the 2008 math TAKS passing rate.

First, we evaluate different sets of predictors to include in the auxiliary model. Commonly with educational outcomes, one would expect that a pretest score, or the performance on the outcome measure before the treatment was applied, could explain much of the variance in the post-treatment outcome. We use the 2007 math TAKS passing rate (8th grade for middle schools and 9th grade for high schools) as a pretest, and will refer to it as such for the remainder of the paper. This pretest score is thus accounting for the scores of previous students taking the same test, with the same teachers, in the same school. Table 1 shows that the pretest alone explains more than 50% of the variance in the 2008 TAKS passing rates for middle and high schools, using OLS. However, including the full set of predictors in a random forest model reduces the unexplained variance by an additional 10-15%. We also evaluate other sets of predictors in between these two extremes including all possible TAKS outcomes from previous years and and only the top 20 pre-

[3] Applying this definition to the schools in the CTAI study resulted in the same classifications of middle versus high schools as the original study.

[4] See reference code at `https://github.com/mannczz/aeis-aux-rct` for a full list of AEIS data sets used.

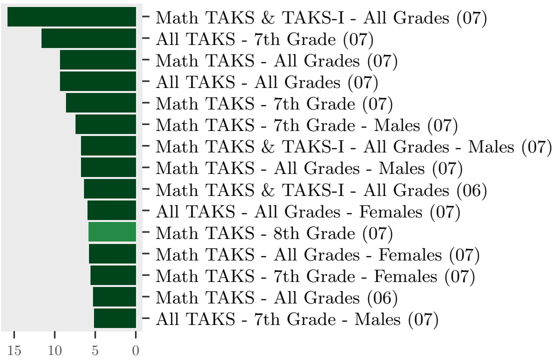

Figure 1: Top 15 predictors, ranked by variable importance, for the 2008 8th grade math TAKS passing rate in auxiliary *middle* schools according to a random forest algorithm fit with all possible predictors. Bars represent the % increase in MSE for each predictor. The pretest score is highlighted with lighter green. Any "TAKS" variable represents the passing rate, and TAKS-I is a version of the test with accommodations. Numbers in parentheses indicate the year.

dictors in terms of variable importance. We find that using all predictors out-performed other options.

We additionally explore using random forests versus neural nets to predict passing rates [11, 9]. Based on the MSE and R^2 of the resulting predictions, we find that the random forest models consistently out-perform the neural nets, for different model specifications and sets of predictors. Also, the default parameters in the `randomForest` package in `R` perform essentially as well as any further tuning.

Thus, we conducted extensive model development, evaluating difference approaches to accommodating missing values, covariate selection, and algorithm selection. Based on these steps for model development, the final auxiliary model uses random forests and all available predictors, fit separately on middle and high schools.

4.3 Model Interpretation

In general, random forests have the practical advantages of offering model interpretation in the form of variable importance and requiring little tuning, so we recommend them for fitting an auxiliary model. The ability to interpret which variables most influence the algorithm's predictive performance allows the researcher to evaluate a model on its predictive properties as well as whether it makes some sense from an educational standpoint. The goal is to develop an auxiliary model that will predict the outcomes in the *RCT sample* well. For this to be the case, there needs to be some overlap in the characteristics of the auxiliary and RCT schools and the relationship between the predictors and outcome needs to be similar between the auxiliary and RCT schools. Therefore, the variable importance could give an idea of whether the auxiliary model is relevant for the RCT schools.

We complete such an investigation with the auxiliary Texas schools in the AEIS data. Figures 1 and 2 display the top 15 predictors, in terms of variable importance for random forest models fit on the auxiliary middle schools and high schools, respectively. The variable importance is based on a random forest algorithm fit with all possible predictors on the 2008 TAKS passing rate. The bars represent the percent increase in MSE for each predictor (a common measure of variable importance [6]). The pretest score is highlighted with a lighter colored bar.

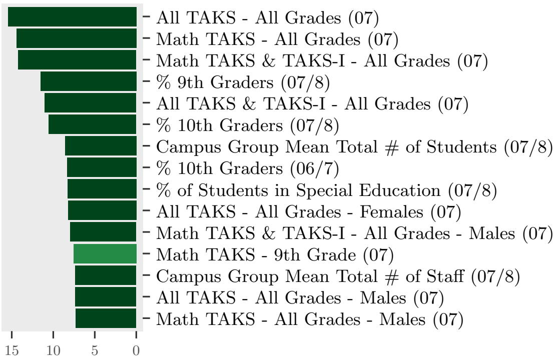

Figure 2: Top 15 predictors, ranked by variable importance, for the 2008 9th grade math TAKS passing rate in auxiliary *high* schools according to a random forest algorithm fit with all possible predictors. Bars represent the % increase in MSE for each predictor. The pretest score is highlighted with lighter green. Any "TAKS" variable represents the passing rate and TAKS-I is a version of the test with accommodations. Numbers in parentheses indicate the year of an exam or school year. "Campus Group" is a group of 40 comparison schools with similar demographic characteristics.

The variable importance validates that the auxiliary models are picking up on expected predictors for the math TAKS passing rate. The top predictors for both types of schools include passing rates on the previous years' TAKS, both for the mathematics section and for the test overall (which also includes reading, writing, science, and social studies, depending on the grade). The overall school performance was important for both middle and high schools, indicating that a general school environment or resources are relevant even for performance within a specific grade. Additionally, even though the overall TAKS performance is important, indicators of the reading TAKS do not show up among the top predictors, which further validates that the model is picking up on mathematical performance.

We additionally find interesting differences between predicting the 8th and 9th grade math TAKS passing rates. For high schools, the top predictors included demographic variables from the 2007/8 school year, such as the percent of 9th or 10th graders in a school, the percent of students in special education programs, and the number of students and staff, in addition to TAKS passing rates (Figure 2). For middle schools, the top 15 predictors are primarily different TAKS passing rates (math or overall), either for all grades or for 7th graders from the previous year (Figure 1). The previous year's passing rate for 8th graders (pretest) is a strong predictor as well, but these findings indicate that the actual cohort of students (who where in 7th grade in 2007 and 8th grade in 2008) was highly important. On the other hand, it is not possible to follow the cohort of 8th graders in 2007 who became 9th graders in 2008 with the given data, so other types of predictors rise in importance when predicting the outcome for high schools.

We gained interesting insights from the auxiliary model development, which would not be possible with the small trial data. The variable importance, in particular, provides observational results that can frame or complement treatment effect estimation results from the RCT.

5. TREATMENT EFFECT ESTIMATION

We compare the estimated precision of point estimates for the effect of CTAI on the 2008 school-level math TAKS passing rate, with different approaches to covariate adjustment. We estimate the average effect for all 44 Texas schools in the study (not separating middle and high schools). As a baseline, we estimate the ATE using no covariate adjustment, or in other words, with the difference-in-means estimator $\hat{\tau}^{DM}$.[5] We compare this to the sample-splitting estimator using three different imputation models for $\hat{d}_{-i}(\cdot)$: (1) OLS with the pretest score for the RCT sample only; (2) random forests with all covariates available from the AEIS data for the RCT sample only; (3) OLS with the auxiliary predictions. The first two imputation algorithms use only information from the RCT, while the last incorporates information from the auxiliary Texas schools using the auxiliary model described in the previous section.

To estimate the variance of the difference-in-means point estimator, we use the typical variance estimator for a paired t-test [8]: $\hat{\mathbb{V}}(\hat{\tau}^{DM}) = \frac{1}{N(N-1)}\sum_{i=1}^{N}(\hat{\tau}_i - \hat{\tau}^{DM})^2$, where $\hat{\tau}_i = (2T_i - 1)(Y_{i1} - Y_{i2})$. We use the variance estimator proposed in [26] to estimate the variance of the sample-splitting estimator: $\hat{\mathbb{V}}(\hat{\tau}) = \frac{1}{N}\sum_{i=1}^{N}(V_i - \hat{V}_i)^2$, where $V_i = T_i v_i^{(1)} + (1 - T_i)v_i^{(2)}$ and $\hat{V}_i = T_i \hat{v}_i^{(1)} + (1 - T_i)\hat{v}_i^{(2)}$. We compare the precision of point estimators in terms of their estimated relative efficiency, the ratio of their estimated variances.

Tables 2 and 3 display the results, with Table 2 showing the point and variance estimates and Table 3 showing the relative efficiency compared to two estimators. Table 3 includes the relative efficiency of each point estimator, as compared to the baseline, difference-in-means estimator ("vs. None") and to the estimator that only includes the pretest as a covariate ("vs. Pretest"). A relative efficiency greater than one indicates that the point estimator is more precise than the comparison estimator, while a relative efficiency less than one indicates that it is less precise. The estimated relative efficiency can also be interpreted as an estimated proportional change in effective sample size.

Table 2: Point and variance estimates of the ATE for the CTAI study, using different covariate adjustment approaches. "None" is the difference-in-means estimator ($\hat{\tau}^{DM}$). The rest use the sample-splitting estimator with different models to impute d_i. "Pretest" uses OLS with the 2007 math TAKS passing rate. "All Covs" uses random forests with all covariates in the trial sample only. "Auxiliary Prediction" uses OLS with the auxiliary prediction as the only covariate.

Covariates Used For Adjustment	Point Est.	Var Est.
None	-6.82	9.82
Pretest	-2.04	5.66
All Covs (RCT)	-5.90	8.03
Auxiliary Prediction	-3.77	4.13

Table 3: Relative efficiency of the point estimates, compared to using no covariate adjustment, or only the pretest in the RCT sample for covariate adjustment. Relative efficiency is calculated as $\hat{\mathbb{V}}(\hat{\tau}_A)/\hat{\mathbb{V}}(\hat{\tau}_B)$, where $\hat{\tau}_A$ is is the estimator indicated in the column header, and $\hat{\tau}_B$ is indicated by the row.

Covariates Used	Relative Efficiency	
For Adjustment	vs. None	vs. Pretest
None	1.00	0.58
Pretest	1.74	1.00
All Covs (RCT)	1.22	0.70
Auxiliary Prediction	2.38	1.37

As shown in Table 3, we find that adjusting for the pretest alone greatly improves precision – using the sample-splitting estimator with the pretest alone ("Pretest") instead of the difference-in-means estimator ("None") is equivalent to increasing the sample size by 74%. On the other hand, attempting to improve on this using only the RCT data can actually be counter-productive. When we use all covariates and random forest model within the RCT sample alone to impute d_i ("All Covs (RCT)"), the resulting effect estimate is less precise than the estimate adjusting for only the pretest. Although including all possible covariates decreased prediction MSE in the *auxiliary* data, the trial data is so small that the imputation model with all covariates is noisy, and therefore hurts precision. However, using the auxiliary predictions, which encode information from the full set of covariates, in effect estimation ("Auxiliary Prediction") far outperforms the difference-in-means estimator and improves even beyond adjusting for the pretest. We find that leveraging information from the large, publicly available, auxiliary data, rather than adjusting with the pretest score in the experimental sample alone is equivalent to increasing the sample size by 37%.

6. DISCUSSION

In this paper, we apply a recent approach to integrating auxiliary information in the analysis of an RCT to an educational trial, using publicly available data [5]. We found that working with large data on almost 3,000 auxiliary Texas schools allowed us to uncover patterns that would be difficult to find with the Texas CTAI trial sample alone (44 schools). We were able to thoroughly explore different pre-processing and prediction strategies with the auxiliary data. We prefer

[5] [26] show that if leave-one-out mean imputation is used as $\hat{d}_{-i}(\cdot)$, then the sample splitting estimator is equivalent to the difference-in-means estimator, so $\hat{\tau}^{DM}$ can also be thought of as the sample splitting estimator that uses no covariates.

using the random forest algorithm because it is simple to implement in R and provides model interpretation via variable importance. Using the resulting auxiliary predictions for the trial sample in covariate adjustment resulted in a more precise treatment effect estimate than covariate adjustment using information from the trial schools alone. Even with a powerful experimental design (paired), and a powerful covariate (pretest), we still find considerable gains to including auxiliary information in estimation.

Developing an auxiliary model is additional work. However, even beyond the gain in precision, we find that what one can learn about the outcome and covariates more than makes up for the added effort. First, we were able to uncover the importance of the cohort of students for predicting future performance in middle schools, which was not possible in the trial sample alone given the small size. We were also able to determine an approach to handle missing values, which are common in publicly available data, based only on the predictive performance of the auxiliary model. Being able to fit a models with high dimensional covariates using the auxiliary schools also had implications for covariate selection. If researchers had to pre-specify covariates for adjustment, it would only be reasonable to choose a small number to avoid model over-specification. Selecting an ideal set of highly predictive covariates a-priori could be *more difficult* than obtaining auxiliary data and fitting a random forest model on it, as we did. If one does not select covariates a-priori but instead tries to select covariates based on the trial data alone, we find that precision can actually suffer. On the other hand, leveraging the thousands of auxiliary schools, we were able to reliably evaluate different sets of covariates with cross validation. Therefore, in this example, we find considerable practical gains from modeling with auxiliary data.

This work evaluates the data integration approach on one trial. However, one may expect that there would not be much room to improve efficiency for treatment effect estimation with the CTAI study, given the paired design and powerful available baseline covariate. Therefore, the fact that we find considerable efficiency gains from the data integration approach in this trail is promising, indicating that the approach could be efficacious in other educational field trails as well, especially ones with less powerful designs or available pre-treatment covariates. We hope these promising results can encourage educational researchers to employ the methods in their trail analyses, which will continue to develop evidence regarding the efficacy of the data integration approach.

7. SUPPORTING CODE

All code used in auxiliary model development and exploration can be found on GitHub at `https://github.com/ashhhleywang/Stats-Research-Project`. Code for processing AEIS data, fitting a final auxiliary model, and all chapter results can be found on GitHub at `https://github.com/manncz/aeis-aux-rct`. Full replication data is not provided to protect the identities of the schools included in the CTAI study.

8. ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D210031. C.Z.M. was additionally supported by the National Science Foundation RTG grant DMS-1646108. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education nor other funders.

9. REFERENCES

[1] P. M. Aronow and J. A. Middleton. A Class of Unbiased Estimators of the Average Treatment Effect in Randomized Experiments. *Journal of Causal Inference*, 1(1):135–154, May 2013.

[2] F. Center for Drug Evaluation. Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products, May 2023.

[3] A. Deng, Y. Xu, R. Kohavi, and T. Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*, page 123, Rome, Italy, 2013. ACM Press.

[4] D. K. Evans and F. Yuan. How Big Are Effect Sizes in International Education Studies? *Educational Evaluation and Policy Analysis*, 44(3):532–540, Sept. 2022.

[5] J. A. Gagnon-Bartsch, A. C. Sales, E. Wu, A. F. Botelho, J. A. Erickson, L. W. Miratrix, and N. T. Heffernan. Precise unbiased estimation in randomized experiments using auxiliary observational data. *Journal of Causal Inference*, 11(1):20220011, Aug. 2023.

[6] U. Grömping. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician*, 63(4):308–319, Nov. 2009.

[7] G. Z. Gui. Combining Observational and Experimental Data to Improve Efficiency Using Imperfect Instruments. *Marketing Science*, 43(2):378–391, Jan. 2024.

[8] K. Imai. Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in Medicine*, 27(24):4857–4873, Oct. 2008.

[9] T. Kalinowski, D. Falbel, J. Allaire, F. Chollet, RStudio, Google, Y. Tang, W. V. D. Bijl, M. Studer, and S. Keydana. keras: R Interface to “Keras”, Aug. 2023.

[10] M. A. Kraft. Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, 49(4):241–253, May 2020.

[11] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[12] I. of Educational Sciences. Standards for Excellence in Education Research. Access at: https://ies.ed.gov/seer/preregistration.asp.

[13] I. M. Opper. Improving Average Treatment Effect Estimates in Small-Scale Randomized Controlled Trials. Technical report, Annenberg Institute at Brown University, Jan. 2021.

[14] J. F. Pane, B. A. Griffin, D. F. McCaffrey, and R. Karam. Effectiveness of Cognitive Tutor Algebra I at Scale. *Educational Evaluation and Policy Analysis*,

36(2):127–144, June 2014.

[15] S. J. Pocock. The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases*, 29(3):175–188, Mar. 1976.

[16] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866, Sept. 1994.

[17] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[18] A. C. Sales, E. Prihar, J. Gagnon-Bartsch, A. Gurung, and N. T. Heffernan. More Powerful A/B Testing Using Auxiliary Data and Deep Learning. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, and V. Dimitrova, editors, *Artificial Intelligence in Education*, Lecture Notes in Computer Science, pages 524–527, Cham, 2022. Springer International Publishing.

[19] J. Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4):465–472, 1990.

[20] D. J. Stekhoven. missForest: Nonparametric Missing Value Imputation using Random Forest, Apr. 2022.

[21] Texas Education Agency - Academic Excellence Indicator System. Access at: https://rptsvr1.tea.texas.gov/perfreport/aeis/index.html.

[22] AEIS Explanation of Masking Rules. Access at: https://rptsvr1.tea.texas.gov/perfreport/aeis/2008/masking.html.

[23] Texas Assessment of Knowledge and Skills (TAKS), 2017. Access at: https://tea.texas.gov/student-assessment/testing/student-assessment-overview/2017-ig-taks.pdf.

[24] K. Viele, S. Berry, B. Neuenschwander, B. Amzal, F. Chen, N. Enas, B. Hobbs, J. G. Ibrahim, N. Kinnersley, S. Lindborg, S. Micallef, S. Roychoudhury, and L. Thompson. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical statistics*, 13(1):41–54, 2014.

[25] E. Wu and J. A. Gagnon-Bartsch. The LOOP Estimator: Adjusting for Covariates in Randomized Experiments. *Evaluation Review*, 42(4):458–488, Aug. 2018.

[26] E. Wu and J. A. Gagnon-Bartsch. Design-Based Covariate Adjustments in Paired Experiments. *Journal of Educational and Behavioral Statistics*, 46(1):109–132, Feb. 2021.