Effect estimates using publicly available school-level data in a cluster-randomized educational experiment

Adam C Sales
Worcester Polytechnic Institute
asales@wpi.edu

Charlotte Z Mann
California Polytechnic State University
czmann@calpoly.edu

Johann Gagnon-Bartsch
University of Michigan, Ann Arbor
johanngb@umich.edu

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

ABSTRACT

Effect estimates from randomized trials, though famously free of confounding bias, may nevertheless be too noisy to be of much use. Recent work has shown that supplementing experimental analyses with observational data can improve statistical precision without contributing confounding bias. The idea is, basically, to use observational data to train an algorithm to use baseline covariates to predict experimental outcomes, and then to use predicted outcomes as an additional covariate in effect estimation. However, due to student privacy regulations, observational data is often unavailable. Moreover, outcomes for many educational field trials use bespoke learning measures that are not available in analogous observational datasets. This paper illustrates a way to circumvent these issues in the context of the Texas subset of the Cognitive Tutor Algebra 1 effectiveness trial, which features school-level randomization and a specialized posttest. We train a model using publicly available school-level covariates to predict state-test school passing rates, and use the resulting school-level predictions to reduce the standard errors in a student-level hierarchical effect model.

Keywords

Causal inference, Supervised learning, Educational effectiveness

1. INTRODUCTION

One promising avenue for enhancing the precision of RCTs lies in leveraging the vast quantities of administrative data housed within state longitudinal data systems (SLDS). These datasets, which track student performance and demographics over time, offer a rich source of information that could reduce estimation error and improve the reliability of experimental findings. Specifically, [13, 4, 14] have promoted the use of auxiliary data (what they call the "remnant" of an RCT)—data from subjects who did not participate in the RCT, but for whom covariate and outcome data are available. The idea is to use auxiliary data to train a supervised learner predicting outcomes as a function of baseline covariates, and use the trained model to generate predicted outcomes for RCT subjects. These predicted outcomes may be used as additional covariates to sharpen the causal inference.

Unfortunately, significant barriers remain. Privacy regulations often restrict access to student-level data within SLDS, limiting researchers' ability to fully harness these resources. Moreover, existing methods that utilize auxiliary data typically require the same variables to be present in both the experimental and administrative datasets. This requirement poses a challenge in educational research, where experiments frequently employ bespoke outcome measures that are not captured in administrative records.

This paper presents an initial attempt at a methodological workaround for this issue in the context of cluster-randomized field trials, where treatment is assigned at the school level but bespoke outcomes are measured at the student level. The approach leverages publicly available, school-level aggregated administrative data that includes a proxy outcome—an outcome measured consistently across the entire state and, ideally, correlated with the experimental outcome.

When treatment is randomized at the school level, improving precision largely depends on accurately predicting school-average outcomes. As such, incorporating school-level covariates derived from proxy outcomes can be particularly valuable. To illustrate this strategy, we applied it to two randomized field trials: the year-2, Texas subset of the Cognitive Tutor Algebra I (CTA1) effectiveness trial [9] and the ASSISTments efficacy trial in Maine [12]. In both cases, we used school-level passing rates on state math tests as proxy outcomes, exploiting their availability and expected correlation with the experimental measures.

Both trials employed paired school-level randomization. Our analysis revealed that when ignoring the pairing structure, incorporating predictions from the auxiliary data noticeably improved precision. However, after properly accounting for the paired randomization in the analysis, the improvements were modest and inconsistent. This outcome reflects the strength of the experimental design itself—paired random-

Adam Sales, Charlotte Mann, Johann Gagnon-Bartsch, and Neil Heffernan. Effect estimates using publicly available school-level data in a cluster-randomized educational experiment. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette (eds.) Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025, pp. 578–581. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870254

ization effectively captured much of the variance in the auxiliary predictions, leaving less room for further gains through proxy-based covariates.

These findings highlight both the potential and limitations of using publicly available administrative data to enhance RCT precision. While auxiliary data can offer substantial benefits in certain contexts, its added value may diminish in well-designed experiments that already control key sources of variability through design features like paired randomization.

2. TWO SCHOOL-RANDOMIZED RCTS, AND AUXILIARY DATA

2.1 The CTA1 Effectiveness Trial

In the CTA1 effectiveness trial [9], researchers randomized 73 high schools and 74 middle schools between two conditions: algebra I students in schools randomized to the control arm learned algebra with traditional textbooks and pencil-and-paper assignments, and students in schools randomized to treatment were given access to an online Algebra I intelligent tutoring system, paired with a complementary classroom curriculum. Prior to randomization, schools were blocked into pairs and triples based on a set of baseline school-level characteristics, and were then randomized within those blocks. Algebra I students in participating schools took a pretest and a posttest from the the CTB/ McGraw-Hill Acuity series. This test was not administered to students who were not enrolled in Algebra I, nor to students in schools that did not participate in the RCT.

The study included schools in seven states and spanned two years, but for this analysis we used a subsample—students in Texas middle and high schools in the second year of implementation. We limited the study to Texas due to the ready availability of rich school level administrative data. We limited our analysis to the second year of implementation because in the first year there was a significant and substantial imbalance of pretest scores between students in treatment and control schools. To partially offset the reduction in sample size due to these limitations, we pooled data across middle and high schools, yielding an analysis sample of 2,842 students in 20 pairs of schools (40 schools total); each pair included one school randomized to treatment and another randonmized to control.

Our analysis of the CTA1 Texas data builds heavily on prior work reported in [8], who gathered thousands of rich basline school-level features on available for schools across the state. These included data on enrollment, demographics, and prior achievent over several years. [8] used that dataset to train a random forest [2] algorithm to predict school passing rates on 8th or 9th grade mathematics section of the Texas Assessment of Knowledge and Skills (TAKS) in 2008. In the present study, we will use the same school-level TAKS predictions from [8] to adjust estimates using student-level posttest outcomes.

2.2 The ASSISTments Efficacy Trial

In the ASSISTments efficacy trial, reported in [12], reseachers constructed 22 pairs of middle schools in Maine and randomized one school from each pair to the treatment condition, where teachers were given access, training, and support for the ASSISTments online homework program, or to a business-as-usual condition. The initial plan of the study was to use the Maine state test as the primary outcome, but during the study period the state first shifted to an alternative testing system, the Smarter Balanced test, and then shifted to yet another standardized testing framework. For this reason the researchers instead administerd the TerraNova assessment to students in participating schools.

Unfortunately, we do not have access to almost any of the student-level baseline covariates from the study (the one exception being an indicator for special education status) but we do have design variables, i.e. pair and school identifiers and an indicator for treatment status, and TerraNova test scores. In addition, we gathered 745 school-level measures of for several years before the onset of the study, along with school-average Smarter Balanced passing rates. We used this data to train a SuperLearner [15], an ensamble learner, combining random forest and regularized linear regression models to predict Smarter Balanced passing rates. We used the ensuing predictions, calculated for the RCT participants, as a covariate in our causal estimation.

3. METHODS FOR CLUSTER-RANDOMIZED STUDIES

There is a surprisingly wide variety of methods, and little consensus among statisticians, on how best to estimate average treatment effects from a cluster randomized study such the CTA1 or ASSISTments studies [7].

For this study, we considered three approachs. First, we estimated hierarchical linear models (HLMs [11]; sometimes called mixed-effects or a multilevel models [5]). These are probably the most popular method for cluster RCTs in education research, and were the models of choice for both [9] and [12]. The HLMs had the following form:

$$Y_i = \beta_0 + \beta_1 Z_{s[i]} + \alpha_p + \gamma X_i + \delta \widehat{pred}_{s[i]} + \eta_{s[i]} + \epsilon_i \quad (1)$$

where $i = 1, \ldots, N$ indexes students, and each student i is nested within a school $s[i]$, $s = 1, \ldots, S$. Y_i is the posttest, measured at the student level: the CTB/McGraw-Hill test for the CTA1 study and the TerraNova exam for the ASSISTments study. α_p, $p = 2, \ldots, P$ are fixed effects for the randomization pairs (since there is a global intercept β_0 in the model, we force $\alpha_1 = 0$ for identifiability). X_i is a student level baseline covariate—special education status in the ASSISTments study and pretest in the CTA1 study—$\widehat{pred}_s$ is the school-level prediction from the model trained on auxiliary data, $\eta_s \sim N(0, \sigma_\eta)$ is a random intercept for school, and $\epsilon_i \sim N(0, \sigma_\epsilon)$ is a random student-level regression error. Finally, $Z_{s[i]}$ is the randomized treatment assignment for student i's school: $Z_s = 1$ if school s is randomized to the treatment arm and 0 otherwise. We take its coefficient β_1 as the treatment effect. We fit model (1) in `R` using the `lme4` package [10, 1].

The second approach we considered used the new `R` package `propertee` to estimate treatment effects [3]. Rather than modeling outcomes, as in HLMs, the `propertee` builds directly off of the experimental design, using clusters and blocks to compute observation-level weights that are explic-

itly targeted at average effects for well-defined samples of subjects. In our analyses, we used average treatment effect weighting, that targets the average effect among all students in the study. `propertee` allows researchers to specify a separate outcome regression to increase the precision of their estimates. The predictions or fitted values from this model enter into the `propertee` estimator as offsets, subtracted from Y prior to model fitting. In some of our analyses, we computed offsets via linear regressions of the outcome on $\widehat{pred}$, a student-level covariate, or both, along with an indicator for treatment status (which improves model fit, but is automatically excluded from the offset). Finally, the software package uses sandwich-style semi-parametric standard error estimators [6] that account for the study design—including clustered treatment assignment and randomization blocks or pairs—and propegates the uncertainty from the covariance adjustment outcome regression.

A third appoach, dRCT, is described in [8]. It uses school-averaged data, weights schools based on their sample sizes, and uses a leave-one-out approach to covariate adjustment that allows for any model predicting outcomes as a function of covariates, including modern machine learning. In our examples, due to the small number of schools involved, we used ordinary least squares for covariate adjustment. Currently, the dRCT package does not include methods for cluster randomized trials without pair matching, so in this paper we always account for pair matching in dRCT analyses.

4. RESULTS

4.1 CTA1 Study

Despite being predictions of a different outcome altogether, the predicted TAKS passing rates were still predictive, with a roughly 0.5 correlation with school-average outcomes.

Table 1 shows estimated treatment effects and standard errors for eight HLMs: the results on the left do not include $\widehat{pred}$ as a predictor, while those on the right do. The top two rows reflect models that ignore the blocking or pairing structure of the data, while the bottom two rows include pair fixed effects. Finally, rows two and four include pretest prediction in the model, while rows one and three do not.

When the model ignores the paired structure of the randomization, the estimated standard errors after adjusting for $\widehat{pred}$ are slightly smaller than the standard errors before adjusting; unsurprisingly, student-level pretest scores have a larger effect on standard errors than $\widehat{pred}$. Compared with the unconditional model, with a standard error of 0.211, adding just $\widehat{pred}$ decreases standard errors to 0.182, while adding pretest decreases standard errors to 0.168. Nevertheless, even when pretest is already part of the model, including $\widehat{pred}$ decreased standard errors very slightly.

When pair fixed effects—α_p in (1)—are included in the model to account for the randomization pairs, including $\widehat{pred}$ increases standard errors.

Table 3 shows analogous results using the `propertee` estimator. In this case, when randomization pairs are ignored and there are no other baseline covariates in the model, including $\widehat{pred}$ decreases standard errors by approximately 1/3, a greater decrease even than inclusion of pretest scores. When randomization pairs are ignored but pretest is included in the model, also including $\widehat{pred}$ decreases standard errors less dramatically. The lower two rows show estimates when randomization pairs are included. Interestingly, whenever an offset (i.e. covariate adjustment) is present in the model, the estimates and standard errors are identical to those when randomization pairs were ignored. The only difference between the top and bottom half of the table is the case in which randomization pairs, but not covariates, are included in the analysis—this model estimates the lowest standard error (and the lowst treatment effect) of any of the other estimators.

Table 1: HLM estimated average effects and standard errors with and without $\widehat{pred}$, adjusting for pretest, or accounting for randomization blocks (i.e. pairs)

	Without Aux. Pred.		With Aux. Pred.	
cov	est	se	est	se
Ignoring Blocks				
No Pretest	0.010	0.211	0.111	0.182
Pretest Adj.	0.087	0.168	0.142	0.160
Including Blocks				
No Pretest	0.013	0.146	0.008	0.161
Pretest Adj.	0.090	0.123	0.063	0.134

Table 2: dRCT estimated average effects and standard errors with and without $\widehat{pred}$, and/or adjusting for pretest. All estimates account for randomization blocks.

	Without Aux. Pred.		With Aux. Pred.	
cov	est	se	est	se
No Pretest	0.019	0.189	0.066	0.144
Pretest Adj.	0.057	0.174	0.098	0.145

5. ASSISTMENTS STUDY

Results in the ASSISTments study mirror those in the CTA1 study. When ignoring blocking, including school-level auxilliary predictions can significantly improve HLM estimation precision, even when also including a student-level covariate, in this case special education status. However, when

Table 3: `propertee` estimated average effects and standard errors with and without $\widehat{pred}$, adjusting for pretest, or accounting for randomization blocks (i.e. pairs)

	Without Aux. Pred.		With Aux. Pred.	
	est	se	est	se
Ignoring Blocks				
No Pretest	0.013	0.301	0.088	0.199
Pretest Adj.	0.051	0.227	0.085	0.193
Including Blocks				
No Pretest	0.013	0.067	0.088	0.199
Pretest Adj.	0.051	0.227	0.085	0.193

Table 4: HLM Results for ASSISTments in Maine

	Without Aux. Pred.		With Aux. Pred.	
cov	est	se	est	se
Ignoring Blocks				
FALSE	10.553	5.960	10.085	4.989
TRUE	11.095	5.686	10.712	4.894
Including Blocks				
FALSE	10.129	4.905	10.118	4.877
TRUE	10.811	4.780	10.803	4.845

Table 5: dRCT results for ASSISTments in Maine

	Without Aux. Pred.		With Aux. Pred.	
cov	est	se	est	se
FALSE	11.050	4.032	8.192	4.722
TRUE	12.618	3.867	9.592	4.818

including blocks in the the analysis, auxilliary predictions make little difference in the analysis.

In dRCT analysis accounting for pair matching, auxilliary predictions hurt estimation precision.

6. CONCLUSION

We had initially been optimistic that auxilliary predictions using school-level data and an alternative outcome measure would circumvent some of the major limitations of earlier approaches to using auxilliary data in RCT analysis. It turns out that the benefit of our approach depens on the quality of the experimental design. In a high-quality design including well-chosen matches, such as the two studies we considered, our school-level adjustments had nothing to add. However, if researchers did not initially match the schools before randomization, adjustment using auxilliary data can help recover most of benefits of matching.

7. REFERENCES

[1] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[2] L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[3] J. Errickson, J. Wasserman, and B. Hansen. *propertee: Standardization-Based Effect Estimation with Optional Prior Covariance Adjustment*, 2025. R package version 0.6.1, commit 9e10a6a759d61c8d4b5603651b71ad2f2f8664b8.

[4] 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, Jan. 2023. Publisher: De Gruyter.

[5] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2007.

[6] G. Kauermann and R. J. Carroll. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456):1387–1396, 2001.

[7] C. Z. Mann, A. C. Sales, and J. A. Gagnon-Bartsch. A general framework for design-based treatment effect estimation in paired cluster-randomized experiments, 2024.

[8] 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. PaaÃŸen 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.

[9] 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, 2014. Publisher: [American Educational Research Association, Sage Publications, Inc.].

[10] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024.

[11] S. W. Raudenbush and A. S. Bryk. *Hierarchical linear models: Applications and data analysis methods*, volume 1. sage, 2002.

[12] J. Roschelle, M. Feng, R. F. Murphy, and C. A. Mason. Online mathematics homework increases student achievement. *AERA Open*, 2(4):2332858416673968, 2016.

[13] A. C. Sales, B. B. Hansen, and B. Rowan. Rebar: Reinforcing a Matching Estimator With Predictions From High-Dimensional Covariates. *Journal of Educational and Behavioral Statistics*, 43(1):3–31, Feb. 2018. Publisher: American Educational Research Association.

[14] A. C. Sales, E. B. Prihar, J. A. Gagnon-Bartsch, and N. T. Heffernan. Using auxiliary data to boost precision in the analysis of a/b tests on an online educational platform: New data and new results. *Journal of Educational Data Mining*, 15(2):53–85, 2023.

[15] M. J. Van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.