Tools for Planning and Analyzing Randomized Controlled Trials and A/B Tests
A Half-Day Workshop

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

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

Duy M. Pham
Worcester Polytechnic Institute
dmpham1@wpi.edu

ABSTRACT

With the rapid growth of education data mining as a research field in recent years and the increasing interest in measuring the effectiveness of the tools we produce, there has been an increasing number of randomized educational experiments and the amount of data gathered from them – particularly with A/B testing on education software platforms. In turn, there is an ever-present demand for robust and flexible estimation and analysis of treatment effects with these data. In this hands-on tutorial, we will demonstrate, explain, and show participants how to implement novel and powerful approaches to select experimental sample sizes, unbiasedly estimate average treatment effects, and analyze patterns of treatment effect heterogeneity. These methods are easy to implement with off-the-shelf open-source software and have been shown to produce more precise effect estimates than currently popular methods.

Keywords

A/B testing, randomized controlled trial, causal inference

1. INTRODUCTION

Education researchers are often interested in causal effects—for instance, how a particular learning or teaching strategy may affect students' abilities, knowledge, or affective states. The most sure-fire approach to measuring causation is the Randomized Controlled Trial (RCT), in which an experimenter randomizes subjects into two or more groups, exposes each group to a different condition, and then measures outcomes of interest. RCTs can range from large field trials [9, 6, 2] that evaluate a program, technology, or curriculum over the course of a full school year, to A/B tests conducted within an online learning environment [8], measuring the short-term impact of, for instance, types of hints or feedback [16], metacognitive prompts [5], or non-cognitive interventions [13].

A. C. Sales, J. A. Gagnon-Bartsch, and D. M. Pham. Tools for planning and analyzing randomized controlled trials and a/b tests. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 1026–1029, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12730043

Education researchers typically analyze RCT data with t-tests or regression models. However, those methods leave a lot of data, power, and science on the table. They neglect, or do not make full use of, the rich baseline data—for instance, prior clickstream or administrative data—available for each student participating in the RCT; they make no use of data on covariates and learning outcomes from students who were not part of the RCT—large, rich auxiliary datasets that are often available; and they only estimate overall average treatment effects, masking between-student variations in effectiveness that are sometimes present.

This tutorial will teach participants how to use modern statistical approaches to effect estimation that leverage off-the-shelf machine learning algorithms [14, 15] and incorporate auxiliary datasets [3, 11] to estimate treatment effects with no bias – but with much greater precision. These methods can identify average treatment effects that would otherwise be lost in the noise of more simple approaches, allow researchers to explore treatment effect heterogeneity, and even plan better experiments to begin with.

The statistical theory underlying some of these methods has been introduced to the EDM community in technical papers and presentations [10, 11]. In contrast, **this tutorial will focus on the practical application of the methods** using a new open-source library in `R` [12] that we have developed—**by the end of this tutorial, participants will be able to use these methods, just (or almost) as easily as running a simple linear regression.**

2. BACKGROUND

The tutorial will focus on a suite of methods based on a familiar idea in EDM: avoiding biases due to phenomena such as overfitting or the optimizer's curse by training a model on one dataset and evaluating it on another, either by sample splitting or cross-validation. When applied to the analysis of RCTs, the separation of model training and testing allows for not only unbiased but also extremely flexible causal estimation.

2.1 Leave-One-Out Cross Validation for Estimating Causal Effects

The LOOP ("Leave-One-Out Potential Outcomes") estimator [14] estimates causal effects by combining randomization and supervised learning algorithms. For each subject $i = 1, \ldots, n$ in an RCT, it trains predictive models using the other $n - 1$ subjects to impute the outcomes subjects would

potentially exhibit if assigned to, say, condition A or condition B, as a function of baseline covariate data. LOOP then uses the randomization of treatment assignment to construct an unbiased estimate of each subject's treatment effect using their imputed potential outcomes. Finally, it estimates the average treatment effect as the average of those individual effect estimates.

When the analysis model matches the design of the experiment, the causal estimators are guaranteed to be unbiased and their confidence intervals or p-values are guaranteed to be conservative. LOOP is a "design-based" estimator—these guarantees depend on the experimental design, not on any statistical model. It can use any model or algorithm to predict potential outcomes and will be unbiased even if the model is inaccurate or fits poorly.

However, the more accurate the model's predictions, the more precise the LOOP estimate will be. [14] recommends a Random Forest algorithm and empirical results show that LOOP using Random Forests can yield much more precise effect estimates than regression.

2.2 The More (Data) the Merrier: Incorporating Auxiliary Data

Besides data on covariate values and outcomes of individuals within RCTs, researchers often also have access to those for individuals outside of the RCTs [10]. For instance, researchers analyzing A/B tests on an educational platform can access historical data from students who used the platform before the A/B test began. Researchers conducting a field experiment have data from schools and school districts that did not agree to be randomized between conditions.

Auxiliary datasets are often much larger than RCT datasets, allowing researchers to train supervised learning algorithms that are higher-dimensional, more complex, and more accurate than what would be possible using only RCT data. [3] and [11] use auxiliary data to train a model predicting the outcomes as a function of baseline covariates, use the trained model to predict outcomes for subjects in the RCT, and then use the predictions from this model as an additional covariate in LOOP. [11, 3] showed that doing so can improve the precision of causal estimates, even if auxiliary data do not closely reflect or match up with the experimental data.

2.3 Not Everyone Experiences the Same Effect

LOOP estimators are essentially averages of unbiased estimates of individual treatment effects. Researchers interested in how treatment effects differ between individuals can *model* these effect estimates instead of averaging them. For instance, researchers interested in understanding what factors are associated with higher or lower treatment effects may regress the individual effect estimates on a set of baseline predictors (which, in this case, are moderators). Researchers interested in predicting an effect for new subjects—perhaps to personalize their tutoring plans—can fit a neural net or random forest to individual effect estimates, and use the fitted model to anticipate the effect for new users. A similar approach can be used to extrapolate the result of an RCT to a new population of potential users, as long as all of the important moderators are measured and included in the model [7].

Part	Description	Timing
I	Conceptual Overview	0:00–0:30
II	Estimating Effects with RCT Data	0:30–1:15
III	Incorporating Auxiliary Data	1:15–2:00
	Break	2:00–2:30
IV	Treatment Effect Heterogeneity	2:30–3:15
V	Planning Experiments	3:15–4

Table 1: Tutorial schedule

2.4 Using Auxiliary Data to Plan Experiments

Power analysis and sample size selection are among the most vexing parts of experimental design, partly because they require experimenters to guess at a set of unknown parameter values. Auxiliary data can help here, too, serving as a basis for these guesses. Specifically, researchers planning on using an existing auxiliary dataset to boost the precision of their estimate can use the auxiliary dataset to train a model predicting outcomes as a function of covariates before the experiment even begins. Then, they can use cross-validation estimates of that model's prediction accuracy to anticipate the precision of their eventual effect estimate from the RCT.

However, the cross-validation estimate of the model's accuracy will only serve as a good prediction of the model's performance imputing potential outcomes for RCT participants if subjects who participate in the RCT are similar to the model training set. To achieve more honest predictions, we may assess a model's performance on a large number of subgroups within the auxiliary data, and use results for those subgroups to calculate a range of power estimates or require sample sizes for an upcoming experiment. Then, a researcher can make an informed decision on power and sample size, based on auxiliary data and on the uncertainty of extrapolating from the auxiliary dataset to a new, randomized dataset.

3. TUTORIAL GOALS

By the end of the tutorial, participants will be able to:

- Use modern machine learning techniques to estimate precise and unbiased effects from Bernoulli-randomized or pair-randomized experiments.
- Incorporate appropriate auxiliary data into effect estimators to further improve precision without incurring additional bias.
- Estimate and model individual, average, and heterogeneous treatment effects.
- Use auxiliary data to help plan a randomized experiment.
- Identify scenarios where or assumptions under which these techniques are appropriate or inappropriate.

4. THE PLAN

The tutorial will begin with conceptual overviews of randomized trials, principles of causal inference, and the methods

we will focus on. The bulk of the tutorial will be spent alternately discussing specific use cases and guiding participants through hands-on demonstrations using [either datasets they brought or] example EDM datasets.

4.1 Software

We have developed an open-source package to implement these methods on the `R` data analysis platform, and a Shiny [1] application—a graphical user interface—for power analysis using auxiliary data. To facilitate their use for tutorial participants, we will provide a downloadable Docker container that contains the `R` program, our software library (along with libraries it depends on), and example datasets. By opening the docker container on their laptops, participants can carry out all of the analyses we will describe without having to install `R` or download any additional files. The Docker container will run on all major operating systems, including Windows, OSX, and Linux.

We will also provide instructions for installing our library and downloading datasets for participants who already have `R` on their computers.

4.2 Example Data Sets

Each discussion of a use case will be illustrated with one of two example datasets, and after each discussion participants will have the opportunity to carry out their own analysis on that dataset. We intend to use two EDM-themed datasets in the tutorial:

- Data from an A/B test run on the ASSISTments E-Trials platform designed to measure the effect of spacial features on students' demonstrated mastery of order-of-operations problems [4]. The dataset includes data on the performance of 5,732 middle school students assigned to one of two conditions, along with nine aggregated student-level covariates. The associated auxiliary dataset includes log data from ASSISTments users who worked on the same ASSISTments module before the A/B test started.
- School-level data from a field trial testing the effectiveness of an intelligent tutoring system (ITS). In the experiment, schools were paired and randomized to either use the ITS or continue business as usual. The dataset includes publicly available school-grade-level passing rates, prior achievement, and demographic variables. Associated auxiliary data includes the same school-grade-level measures for schools in the same state that did not participate in the RCT.

4.3 Tutorial Organization

Part I: Conceptual Introduction. The tutorial will begin with a conceptual introduction to causal inference from RCTs, as well as the logical underpinnings, structure, and assumptions of the methods we will be discussing. We will begin by reviewing causal estimation targets—individual, average, and conditional average treatment effects—and how RCTs facilitate their unbiased estimation. Next, we will describe common effect estimators, including the simple difference (t-test) and regression estimators. Finally, we will explain, at a high level, how researchers can use covariates, auxiliary data, and modern machine learning methods to estimate individual and average treatment effects more precisely, without bias. By the end of this part of the tutorial, participants will be able to describe different goals of RCT analysis and have a deep enough understanding of the estimation process to be able to choose the appropriate estimator for a given scenario, and to make well-informed modeling decisions.

Part II: Estimating Average Treatment Effects with RCT Data. The remainder of the tutorial will be purely practical, beginning with the most common use case, estimating average treatment effects using RCT data. We will on two common experimental designs: first, we will discuss Bernoulli randomization, in which each subject is randomized between conditions independently, such as in the ASSISTments E-trials dataset. We will use the E-trials data to demonstrate how to use our software package in `R` to precisely and unbiasedly estimate average effects for Bernoulli designs by identifying and formatting covariates, treatment indicators, and outcome measures, entering them into the appropriate function, and interpreting the results. Next, we will guide participants in using our software to estimate effects in the same dataset, which we will distribute.

We will repeat this process with the second experimental design, paired randomization, which is common in field trials such as the ITS effectiveness trial: we will demonstrate the use of our software, and guide participants as they use our software to estimate average effects using the ITS field trial data.

Finally, we will briefly describe other experimental designs covered by our software, and show participants how they can specify their own predictive functions within our software.

Part III: Incorporating Auxiliary Data. Auxiliary data—covariates and outcome measures from subjects who were not randomized between experimental conditions—can be used to improve the precision of both average and individual treatment effects without causing bias. We will use both example datasets to demonstrate to participants how to identify usable and potentially helpful auxiliary data; how to verify that pre- and post-randomization measures are correctly labeled as such; how to specify, choose, and fit predictive models using auxiliary data; and how to use those models to more precisely estimate average treatment effects. We will demonstrate the use of these methods in `R` using the example datasets, along with their associated auxiliary data, and guide participants through the same.

Part IV: Treatment Effect Heterogeneity. In this part of the tutorial, participants will learn how to use the techniques from the first half of the session to estimate individual treatment effects and subgroup or conditional average effects, and how to use models to uncover patterns in treatment effect heterogeneity. Specifically, we will show how to predict the treatment effect for a new participant—potentially allowing for personalized education—how to extrapolate average effect estimates to new populations, and how to interpretable models to individual effect estimators that may shed light on *how* effects vary between subjects. We will also explain the conditions under which these estimates are accurate. For this portion, we will use the E-trials dataset for demonstra-

tions and hands-on activities.

Part V: Planning an Experiment. The final part of the tutorial will focus on one of the first stages of experimentation—using auxiliary data to choose an appropriate sample size. We have developed an interactive graphical interface, using the `R` web application framework Shiny, to perform power analyses and select a sample size using auxiliary data, and in this part of the tutorial, we will demonstrate to tutorial participants how to use the application, as well as providing an opportunity for them to try it themselves using the field trial or ASSISTments E-trials auxiliary dataset, or another dataset of their choosing.

5. ORGANIZERS AND PRESENTERS

Adam C. Sales[1] is an Assistant Professor of Statistics and a member of the Learning Sciences and Technologies faculty at Worcester Polytechnic Institute. He works on incorporating log data from computer-based learning applications into causal models to better understand what works in education and why.

Johann A. Gagnon-Bartsch[2] is an Associate Professor of Statistics at the University of Michigan. His research focuses on causal inference, machine learning, and nonparametric methods with applications in the biological and social sciences.

Duy M. Pham is a graduate student in the Department of Data Science at Worcester Polytechnic Institute, working with Adam Sales. He has led the development of treatment effect heterogeneity estimators using LOOP and auxiliary datasets.

Charlotte Z. Mann is a recently minted PhD in Statistics from the University of Michigan, where she worked with Johann Gagnon-Bartsch. She has led the development of the `R` package for LOOP and the application of LOOP for pair-randomized experiments.

Jaylin Lowe is a PhD student in the Department of Statistics at the University of Michigan, working with Johann Gagnon-Bartsch. She is leading the development of the Shiny application and methods to estimate power and select sample sizes using auxiliary data.

6. REFERENCES

[1] W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPherson. Package 'shiny'. *See http://citeseerx. ist. psu. edu/viewdoc/download*, 2015.

[2] L. E. Decker-Woodrow, C. A. Mason, J.-E. Lee, J. Y.-C. Chan, A. Sales, A. Liu, and S. Tu. The impacts of three educational technologies on algebraic understanding in the context of covid-19. *AERA open*, 9:23328584231165919, 2023.

[3] 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, 2023.

[4] A. Harrison, H. Smith, T. Hulse, and E. R. Ottmar. Spacing out! manipulating spatial features in mathematical expressions affects performance. *Journal of Numerical Cognition*, 6(2):186–203, 2020.

[5] C. Lang, N. Heffernan, K. Ostrow, and Y. Wang. The impact of incorporating student confidence items into an intelligent tutor: A randomized controlled trial. *International Educational Data Mining Society*, 2015.

[6] 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.

[7] J. Pearl and E. Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 247–254, 2011.

[8] E. Prihar, M. Syed, K. Ostrow, S. Shaw, A. Sales, and N. Heffernan. Exploring common trends in online educational experiments. In *Proceedings of the 15th International Conference on Educational Data Mining*, 2022.

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

[10] A. Sales, A. Botelho, T. Patikorn, and N. T. Heffernan. Using big data to sharpen design-based inference in a/b tests. In *Proceedings of the eleventh international conference on educational data mining*, 2018.

[11] A. C. Sales, E. B. Prihar, J. A. Gagnon-Bartsch, N. T. Heffernan, et al. 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.

[12] R. C. Team. R language definition. *Vienna, Austria: R foundation for statistical computing*, 3(1), 2000.

[13] K. Vanacore, A. Gurung, A. Mcreynolds, A. Liu, S. Shaw, and N. Heffernan. Impact of non-cognitive interventions on student learning behaviors and outcomes: An analysis of seven large-scale experimental inventions. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 165–174, 2023.

[14] E. Wu and J. A. Gagnon-Bartsch. The loop estimator: Adjusting for covariates in randomized experiments. *Evaluation review*, 42(4):458–488, 2018.

[15] E. Wu and J. A. Gagnon-Bartsch. Design-based covariate adjustments in paired experiments. *Journal of Educational and Behavioral Statistics*, 46(1):109–132, 2021.

[16] Y. Zhou, J. M. Andres-Bray, S. Hutt, K. Ostrow, and R. S. Baker. A comparison of hints vs. scaffolding in a mooc with adult learners. In *International Conference on Artificial Intelligence in Education*, pages 427–432. Springer, 2021.

[1] https://www.wpi.edu/people/faculty/asales

[2] https://dept.stat.lsa.umich.edu/ johanngb/