Causal Inference in Educational Data Mining

Anthony F. Botelho
University of Florida
a.botelho@ufl.edu

Avery H. Closser
University of Florida
avery.closser@ufl.edu

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

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

Kirk P. Vanacore
Worcester Polytechnic Institute
kpvanacore@wpi.edu

ABSTRACT

The aim of intelligent tutoring systems, educational games, MOOCs, and similar computerized learning tools is to enhance student learning. While Educational Data Mining (EDM) research primarily focuses on identifying, measuring, and predicting learner behaviors or outcomes, understanding causality is crucial. Causal inference goes beyond prediction to estimate the impacts of various factors on student behaviors or learning outcomes, providing insights into the underlying causes of educational phenomena. This approach is essential for designing effective educational technologies and policies. The field of causal inference, drawing from statistics, philosophy, economics, and computer science, offers methods to navigate complex scenarios and confounding variables, facilitating the examination of causality in digital learning platforms. This workshop will explore the integration of causal inference with EDM, addressing how to handle nested educational data, approaches to quasi-experimental data, embedding experiments in digital contexts, and large-scale randomized experimentation. It aims to feature discussions on ongoing projects, open problems, and new opportunities, aiming to foster collaborations and advance causal methods within EDM.

Keywords

Causal Inference, Treatment Effects, Causal Modeling, Data Mining, Experimental Design

1. INTRODUCTION

The goal of crafting intelligent tutoring systems, educational games, MOOCs, and similar computerized learning tools is to enhance student learning. While the primary focus of Educational Data Mining (EDM) research remains on techniques for identifying, measuring, and predicting learner behaviors or outcomes, there is a critical need to understand causality. Causal inference goes beyond mere prediction to estimate the impacts of various factors on students' behaviors or learning outcomes. It is not just about predicting who will struggle or succeed, but unraveling the underlying causes behind these phenomena. Causality plays a pivotal role both in learning science, which explores how educational interventions and computerized inputs affect educational outputs, and in policy-making, where the aim is to design and systematically deploy systems that enhance learning. Causal inference allows us to confidently make design decisions for educational technology to improve learning outcomes.

The domain of causal inference, encompassing fields of statistics, philosophy, economics, and computer science, has seen rapid advancements. This emerging science addresses the challenges involved in estimating effects amidst complex situations in which confounding variables can obscure results. Thus, it offers methods that allow for examining causality within the context of digital learning platforms which further affords opportunities that may leverage the utilization of educational data mining. Using causal methods, we can explore how impacts of learning interventions and educational technologies differ among learners and identify the mechanisms driving causal effects. Additionally, there is great potential for utilizing educational data mining methodologies to strengthen causal inference techniques within educational contexts. This workshop aims to address outstanding questions and continue discussions around topics including, but not limited to: how to account for the nested structure of educational data [1], approaches to causal inference in the case of quasi-experimental or non-experimental data [2], how to embed experiments in digital learning contexts [3], and randomized experimentation at scale using digital platforms [4].

The workshop will feature invited discussions showcasing ongoing projects addressing causal inquiries, along with brief talks on relevant work in progress, spanning all stages of development. Furthermore, it will provide a platform for EDM researchers to present open problems related to causality in their research. Researchers will have the opportunity to discuss open challenges and problems or new opportunities to leverage EDM to draw causal inferences about affective-, performance-, and learning-based outcomes. We aim for subsequent open-ended discussions among participants to generate constructive suggestions and foster potential collaborations that advance efforts to incorporate causal methods into EDM techniques. This forum aims to provide presenting researchers with valuable insights and catalyze collaborative efforts in tackling causal inquiries within the

A. F. Botelho, A. H. Closser, A. C. Sales, N. T. Heffernan, and K. P. Vanacore. Causal inference in educational data mining. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 1034–1036, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12730047

Table 1: Tentative Schedule

0:00-0:10	Introduction
0:10-1:10	Keynote Speaker (TBA)
1:10-2:00	Long Paper Presentations (15 minutes each with 10 minute discussions)
2:00-2:45	Short Paper Presentations (10 minutes each with 5 minute discussions)
2:45-3:30	Session on Open Problems and Opportunities for EDM

EDM domain while emphasizing the reciprocal relationship between examining causality within EDM and leveraging EDM to support causal inference methodologies. We aim to drive this conversation forward from previous years by working towards establishing good and best methodological practices for applying EDM techniques to better understand causality within educational data.

In sum, the workshop will be organized to stimulate discussion among participants, including, hopefully, constructive suggestions for open problems related to causal inference in the context of EDM research.

2. CONTENT AND THEMES

Our workshop aims to shed light on the prevalence and significance of causal inquiries within EDM while emphasizing the two overarching themes: 1) examining causality within EDM contexts and 2) leveraging EDM to support causal inference methodologies.

We will solicit work on topics including, but not limited to:

- A/B testing
- Graphical causal models/Bayesian networks
- Analyzing data from randomized experiments
- Multi-armed bandits
- Investigations of causal mechanism/mediation analysis
- Estimating EDM program impacts
- Identifying and predicting differential effects
- Connections between machine learning and causal inference
- Dynamic treatment regimes
- Principal stratification
- Causal inference in EDM without randomization

3. SCHEDULE AND FORMAT

The proposed schedule is provided in Table 1. Following each presentation, we will promote and facilitate discussion among workshop participants with the intention of expanding beyond a simple question and answer format. Toward this, in addition to presentations on the previously-listed topics, we will also hold a session focused on open problems and opportunities to leverage EDM within the context of causal inference, guided by submissions as well as issues that are raised during the workshop itself.

3.1 Submission Types

- Full Papers: 10 pages maximum
- Short Papers: 6 pages maximum
- Open Problems/Opportunities for EDM: 4 pages maximum

3.2 Reviewing Process

The workshop organizers will review papers, alongside ad hoc external reviewers whose expertise is appropriate for the submissions.

4. PREVIOUS EDITIONS AND EXPECTED PARTICIPANTS

This convening would represent the fourth edition of this workshop at EDM. With previous workshops held in 2020, 2021, and 2022, and with over 30 individuals attending each, the interest among the EDM community in topics related to causal inference have been clear. Our goal is to resume this convening at EDM 2024 and expect similar participation as in previous iterations. With the co-location and similar timing of the Learning @ Scale conference as well as events held by SEERNet[1], we anticipate similar levels of high interest and attendance for this workshop.

5. ORGANIZERS

Anthony F. Botelho (University of Florida). Anthony is an Assistant Professor of Educational Technology and computer science education in the College of Education at the University of Florida. His research seeks to impact learning through the blending of learning theory and quantitative methods. Anthony's primary lines of research include the study of student cognition, behavior, and affect, identifying effective learning interventions through causal inference, and developing human-in-the-loop systems and tools to support teachers.

Avery H. Closser (University of Florida). Avery is an incoming Assistant Professor of Emerging Technologies and Learning in the College of Education at the University of Florida. Her research aims to leverage cognitive theory to advance learning technologies and open materials for instructional practice. She specializes in experimental design in the context of learning technologies and explores best practices for methodologies related to this area of research.

Adam C. Sales (Worcester Polytechnic Institute). Adam is an Assistant Professor of Mathematical Sciences and an affiliate of the Learning Sciences and Technologies and Data Science programs at WPI. His research in applied statistics focuses on methods for causal inference using large, administrative

[1] seernet.org

datasets, primarily with applications in learning sciences and social sciences. He has developed and worked on methods combining machine learning with design-based analysis of randomized trials and matched observational studies, principal stratification and mediation analysis using log data from intelligent tutoring systems, and regression discontinuity designs.

Neil T. Heffernan (Worcester Polytechnic Institute). Neil is the William Smith Dean's Professor of Computer Science at WPI, the creator of ASSISTments, and an active researcher in the fields of 1) artificial intelligence and education, 2) educational data mining and 3) learning analytics. In order to support research in these fields, Dr. Heffernan created the E-TRIALS Testbed, a tool that allows ASSISTments to be used as a platform to do science and support evidence-based practice. He has dozens of papers in educational data mining, and 20+ papers in comparing different ways to optimize student learning.

Kirk Vanacore (Worcester Polytechnic Institute). Kirk is a Ph.D. student in Learning Sciences and Technologies. He applies statistical and machine learning models to data from computer-based learning platforms to understand learning mechanisms. His work includes using experimental designs and observational studies to understand the nuances of how educational programs and pedagogies impact learning. He focuses on the intersection of failure, struggle, and learning.

6. REFERENCES

[1] A. H. Closser, A. Sales, and A. F. Botelho. Should we account for classrooms? analyzing online experimental data with student-level randomization. *Educational technology research and development*, pages 1–30, 2024.

[2] W. L. Leite, Z. Jing, H. Kuang, D. Kim, and A. C. Huggins-Manley. Multilevel mixture modeling with propensity score weights for quasi-experimental evaluation of virtual learning environments. *Structural Equation Modeling: A Multidisciplinary Journal*, 28(6):964–982, 2021.

[3] B. A. Motz, P. F. Carvalho, J. R. de Leeuw, and R. L. Goldstone. Embedding experiments: Staking causal inference in authentic educational contexts. *Journal of Learning Analytics*, 5(2):47–59, 2018.

[4] 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.