2nd Workshop on Educational Data Mining in Writing and Literacy Instruction

Collin Lynch
NC State University
cflynch@ncsu.edu

Paul Deane
ETS
PDeane@ets.org

Piotr Mitros
ETS
pmitros@ets.org

Zhikai Gao
NC State University
zgao9@ncsu.edu

Damilola Babalola
NC State University
djbabalo@ncsu.edu

ABSTRACT

The rapid advancement of technologies like generative AI presents both challenges and opportunities for writing and literacy education. Issues such as detecting AI-generated text and ensuring student data privacy are growing concerns, while AI-driven tools for automated essay scoring, personalized instruction, and feedback generation offer promising opportunities. Effectively addressing these challenges and leveraging AI's potential is crucial for shaping the future of writing education within the EDM community.

This workshop will explore the latest research and applications in writing and literacy education, and it will provide a platform for discussion and knowledge exchange. As part of the workshop we will introduce our developed platform design to support ethical management, real-time data collection, and analysis of student writing. Through paper presentations and an interactive tutorial session, participants will gain valuable insights, foster collaborations, and contribute to the responsible integration of AI in writing education.

Keywords

Educational Data Mining,Writing Education,Literacy Education,EdTech Tools

1. OVERVIEW AND RELEVANCE

The widespread adoption of Intelligent Tutoring Systems (ITS) and online learning platforms has led to an unprecedented accumulation of student learning data, including detailed records of their writing and literacy activities. With the rise of AI and Data Science—particularly, advancements in Large Language Models and generative AI—Educational Data Mining (EDM) is increasingly being used to personalize instruction and address students' unique writing needs. By harnessing data from writing activities, educators can provide more targeted support, enhancing student learning outcomes and engagement.

Recognizing the transformative impact of these developments, this workshop aims to bring together educators and researchers to explore key themes, including:

- Writing behavior analysis.
- Advancements in NLP for writing assessment.
- Writing style and quality.
- Plagiarism detection.
- Automatic essay scoring and feedback generation.
- Applications of Large Language Models in writing analytic or classroom practice.
- Policy, privacy, and ethical concerns with writing data.
- Psychometrics and measurement.
- Behaviors for ESL students.
- K-12 writing and literacy education.
- Quantitative analysis and case studies for writing education practice.
- Collaborate Writing and peer review.
- Data-driven supports for special education in literacy.
- Interactive writing environments and Intelligent Tutoring System (ITS).
- Writing Dataset across disciplines.

We will announce a Call for Papers related to the above themes. In addition to regular paper presentations from the participants' submissions, we will host a tutorial session to introduce our developing tool called Writing Observer. Writing Observer is an open-source platform that can handle learning process data from student writing data. This tool allow researchers and educators to collect click-by-click online writing data on Google Docs and display real-time teacher dashboards. Therefore, we are highly interested in

Collin Lynch, Paul Deane, Piotr Mitros, Zhikai Gao, and Damilola Babalola. 2nd Workshop on Educational Data Mining in Writing and Literacy Instruction. 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. 688–690. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870310

hearing any feedback or perspectives from the relevant researchers and educators. This tutorial session could also provide more collaboration opportunities for both us and the participants.

We expect around 10-30 audiences for this full-day workshop. Our target audience is researchers and educators with interests or experience in educational research, ITS, writing analytics, or any other related fields which highly overlap with the EDM community.

2. PAST AND AFFILIATED EVENT

2.1 EDM'24: 1st Educational Data Mining in Writing and Literacy Instruction Workshop

We hosted the first Educational Data Mining in Writing and Literacy Instruction Workshop (WLIEDM'24) at the EDM 2024 conference. This is a half-day tutorial and half-day workshop event, which we will follow in the same format this year as well. For the 1st WLIEDM workshop, we had 30 participants joining the discussion either by in-person or online zoom meeting. We successfully presented the potential usage of Writing Observer for actual class and research opportunities. The participants have given us valuable feedback on improving the system and any interesting research they would like to conduct with it. For the research presentations, we accepted five publications and invited all authors to present their research.

2.2 AIED'24: The Learning Observer A Prototype Platform for the Integration of Learning Data Workshop

We organized this tutorial workshop at the AIED conference in 2024. In this workshop, we introduced the Learning Observer, an open, transparently-governed consortium to manage student data in the interest of students and the general public. During this 2-day event, we discuss with the participants about the challenges in both the policy and technology when design and develop this platform. Using Writing Observer as an example, we showcased how this platform could be used pedagogically.

3. WORKSHOP ORGANIZERS

Collin F. Lynch is an Associate Professor in the Department of Computer Science at North Carolina State University. His primary research is focused on developing robust ITS and adaptive educational systems for Ill-Defined domains such as scientific writing, law, and software development. His current research includes work on: argument mining and natural language processing, real-time support for classroom orchestration and writing to learn tasks, advances in student modeling, the development of embodied cognitive agents for collaborative learning, and scaffolding for CS education.

Paul Deane is a principal research scientist in the Research & Development division at ETS. He is the author of Grammar in Mind and Brain, a study of the interaction of cognitive structures in syntax and semantics, and the second author of Vocabulary Assessment to Support Instruction. His current research interests include formative assessment design in the English language arts, cognitive models of writing skills, automated essay scoring, and vocabulary assessment. During his career at ETS, he has worked on a variety of natural language processing (NLP) and assessment projects, including automated item generation, tools to support verbal test development, scoring of collocation errors, reading and vocabulary assessment, and automated essay scoring.

Piotr Mitros is a Senior Research Scientist at ETS. He is also the original author of the popular Open edX learning platform and the original founder as well as the Chief Scientist for more than five years. He have spent the past few years exploring issues around why educational initiatives go south, and evidence-based practices aren't adopted and converged on issues around governance, transparency, and incentive structures. His current work focuses on how we develop educational measurements that incentivize and support rich classroom instruction supporting diverse (rather than standardized) students.

Zhikai Gao is a Ph.D. student at North Carolina State University. His current research focuses on understanding students' learning behaviors through traceable log data from ITS. He has been studying CS education, help-seeking behavior, and LLM usage in education across disciplines.

Damilola Babalola is a second-year Computer Science Ph.D. student at North Carolina State University with a research focus on using Artificial Intelligence (Educational Data Mining and Natural Language Processing) to improve Education. His current work involves research, software development, data mining, and data visualizations aimed at assisting middle-school and high-school students in improving their essay-writing skills. The core of his research centers around the extraction and classification of student essay revisions based on their edit intention, followed by the visualization of student clusters exhibiting similar revision patterns.

4. PROGRAM COMMITTEE

- Effat Farhana: Auburn University
- Bradley Erickson: Educational Testing Service
- Caleb Scott: North Carolina State Univeristy
- Yiqi Wu: North Carolina State Unvieristy

5. WORKSHOP ORGANIZATION

We will organize this workshop as a full-day event. Half of the day will be devoted to a tutorial introduction to the Writing Observer toolkit under development at NCSU and ETS. We will first introduce the purpose of the tool and the development updates since last year's workshop. We then will work through the steps to install the system, having participants build a few dashboards and listening to their feedback. We would like to find participants; assessment of the tool, as well as potential improvement for incorporating their own work and possible research collaborations.

After lunch, we will first present two of our research conducting though the Writing Observer. Then we will move

on for the paper presentations on writing and literacy instruction. We will invite accepted submissions of full papers which describe mature work, short papers with in-progress work or student projects, and poster/demo submissions for those presenting available data, tools, and methods. This last category is particularly targeted at researchers who have data or methods available and are seeking to identify potential collaborators, and potentially would be interested in collecting or analyzing their data with Writing Observer. Based on last year's submission status, we plan to accept 5-10 submissions.

5.1 Tentative Schedule

- 9:00AM-9:15AM: Opening, introducing Writing Observer and report current updates.
- 9:15AM-10:30AM: Installation of Writing Observer in small groups.
- 10:30AM-10:45AM: Refreshment break.
- 10:45AM-12:30PM: Exploration of the system, focus group for collaboration.
- 12:30PM-1:30PM: Lunch break.
- 1:30PM-1:50PM: Keynote speech: Great Job! But why?: Identifying stable textual features for grade prediction in middle-school writing
- 1:50PM-2:10PM: Keynote speech: Tracing the Writing Journey: Using Keystroke Data to Predict Student Performance
- 2:10PM-2:30PM: Refreshment break.
- 2:30PM-4:00PM: Research presentation for accepted papers.
- 4:00PM-5:00PM: Networking and Wrap up.

6. TENTATIVE ORGANIZING TIMELINE

Following is a rough schedule of our timeline. We will adjust the timeline base on different circumstances.

- March 22nd - March 29th : Website release and announcing call for paper.
- May 26th: Submission deadline for papers from all tracks.
- June 16th: Notification for acceptance.
- June 23th: Camera-ready paper deadline.
- July 21st: Workshop day at EDM.