Contribution ID: **322** Type: **Oral Presentation**

Automated Essay Scoring Using Generative Artificial Intelligence: Illustration of a Systematic Evaluation Framework

Wednesday 23 July 2025 18:30 (15 minutes)

Automated essay scoring systems can support teachers by providing rapid, cost-effective verbal and numerical feedback on student writing. In recent years, these systems have improved significantly with the rise of generative artificial intelligence models based on the transformer architecture. Research consistently shows that these models outperform traditional machine learning approaches across a wide range of natural language processing tasks, including essay scoring.

Despite these advancements, the application of this technology in psychology and education presents several risks, including: a) biased, inconsistent, or inappropriate verbal feedback, b) numerical scores that are highly arbitrary or systematically deviate from human ratings, and c) potential discrimination against specific groups in scoring and feedback.

In this talk, we illustrate these risks using empirical findings from an ongoing pilot study on automated essay scoring in Switzerland, drawing on examples from several widely used generative models. We then introduce a framework for evaluating the reliability, validity, and fairness of automated essay scoring systems. This framework integrates psychometric principles, such as item response theory and probabilistic test theory, with benchmarking standards from computer science to systematically identify problematic model behaviour. We demonstrate the framework's practical application using anonymized data from our pilot study and summarize main takeaways and challenges.

Primary authors: MATOBA, Kyle; STAHLHUT, Laura; DEBELAK, Rudolf (University of Zurich, EPFL)

Presenters: MATOBA, Kyle; STAHLHUT, Laura; DEBELAK, Rudolf (University of Zurich, EPFL)

Session Classification: Symposium : "Statistical Learning Approaches to Psychometric Modeling Challenges"