Laboratorija za eksperimentalnu psihologiju,
Odeljenje za psihologiju,
Univerzitet u Beogradu – Filozofski fakultet

The validity of automated essay scoring

Goran Lazendić[1,2]

[1] Australian Council for Educational Research
[2] The University of Sydney

The validity of automated essay scoring remains a crucial concern despite advances in large language models. While these models exhibit remarkable language generation capabilities, their proficiency doesn't guarantee reliable essay evaluation. Validity in automated essay scoring ensures that the assessment aligns with its intended purpose and accurately measures the skills or attributes it aims to evaluate. Firstly, automated essay scoring systems need to be fair and unbiased. Even sophisticated language models can perpetuate or inadvertently learn biases in training data. If not carefully validated, these biases could unfairly impact specific groups of learners, compromising the system's fairness and ethical standing. Secondly, the complexity of human expression poses a challenge for automated systems. Essays often involve nuanced arguments, creativity, and context-dependent interpretations that may elude even the most advanced language models. Validity concerns arise when automated systems struggle to capture the depth and subtleties of human expression, leading to inaccurate assessments. Furthermore, automated essay scoring should align with educational goals. If the purpose of assessment is to measure critical thinking, creativity, or specific content knowledge, validity ensures that the automated scoring system effectively evaluates these attributes. Otherwise, the assessment may not serve its intended educational purpose. In conclusion, while large language models represent a leap forward in natural language understanding, the validity of automated essay scoring remains a critical consideration. It ensures fairness, guards against biases, and verifies that assessments align with educational objectives, ultimately maintaining the integrity and effectiveness of automated essay evaluation in diverse educational contexts.