*The summary was developed with support from Microsoft Copilot.*

**Presentation title: Proficient or Pretentious? Can we tell the difference?**

*Tamar Shalamberidze, Main specialist of Georgian Testing Division*

The growing presence of generative AI in our daily lives made me wonder whether STANAG 6001 raters could reliably distinguish authentic test scripts from those generated by AI. Inspired by research from the University of Reading[[1]](#footnote-1), I designed an experiment in collaboration with the generous help of colleagues from the BILC community and the Georgian national testing team.

The aim was to explore rater sensitivity to inauthentic scripts and answer two key questions:

* Could a test-taker using AI go undetected during exams?
* Would AI get a test-taker a better result?

To investigate, raters were split into two groups:

* **“Detectives”:** informed that AI-generated texts might be included.
* **“Blind raters”:** unaware of any AI involvement.

Both groups were asked to assign STANAG 6001 proficiency levels to a mix of 55 genuine and 8 AI-generated scripts (produced by ChatGPT and Copilot). The Georgian national testing team participated during live testing session. Blind raters received 11 scripts—5 of which were AI-generated and were only told about the possibility of AI after submitting their first ratings. They were then asked to guess which texts were fictitious.

**Key Findings:**

* Many raters noticed signs like flawless grammar, overly smooth structure, mechanical tone, or vocabulary atypical for L2 learners.
* Despite these clues, detecting AI writing was not straightforward.
* False alarms were common, and at least one AI-generated script went completely undetected.
* Surprisingly, AI-generated texts often received higher scores than genuine ones.

**Conclusion:** These findings present a challenge to the validity and fairness of writing assessments. Yet, with proper regulations, awareness, and training, we can adapt. AI should not be feared as a threat but embraced as a driver of more robust and fair testing practices.

1. https://www.reading.ac.uk/news/2024/Research-News/AI-generated-exam-answers-go-undetected-in-real-world-blind-test [↑](#footnote-ref-1)