The Impact of Artificial Intelligence on Learning in Brazil

Researchers:

Bruno Ferman

Lycia Lima

Flávio Riva

Sector(s): Education

Location: Espírito Santo, Brazil

Sample: 178 public schools encompassing ~19,000 senior high school students

Initiative(s): Post-Primary Education Initiative (PPE)

Target group: Secondary schools Students Teachers

Outcome of interest: Student learning Technology adoption Take-up of program/social service/healthy behavior

Intervention type: Digital and mobile Online learning Pedagogical innovation Technology

AEA RCT registration number: AEARCTR-0003729

Partner organization(s): Espírito Santo's Education Department (SEDU/ES), Lemann Foundation, Google.org, Oppen Social, Fundação CAEd

Artificial intelligence (AI) has the potential to support teachers in completing time-intensive, subjective tasks, such as grading essays, and to provide individualized feedback to students on their writing. However, evidence on whether and how AI impacts students' writing skills remains limited. To fill this gap, researchers evaluated the learning impacts of using AI systems to score and comment on essays written for Brazil's national post-secondary admission exam. In schools where AI technology was introduced, teachers were able to provide more frequent individualized feedback to students, and students' essay scores improved as a result.

Policy issue

Improving learning is one of the most pressing goals for educational policy in low- and middle-income countries. A common barrier to learning is that teachers are time-constrained and often have to juggle between providing individualized assistance to students and performing routine tasks, like grading. This is especially true for the development of writing skills, which requires teachers to spend long hours grading.

Educational technologies (*ed* techs) could help alleviate this challenge by performing the operational parts of teaching and allowing teachers to reallocate time from less to more complex tasks (e.g., one-on-one feedback to students). For example, automated writing evaluation (AWE) systems can potentially help improve writing by using natural language processing and machine learning algorithms to predict scores and allocate feedback, thereby reducing the amount of time teachers need to spend on grading. However, there is little evidence on the effectiveness of *ed* techs or artificial intelligence (AI) focused on writing. Are AWE systems effective in improving students' writing skills?

Context of the evaluation

Providing high-quality education is a challenge in Brazil, especially when it comes to language skills. According to the 2018 PISA exam, a worldwide study of students' scholastic performance, the average 15-year-old Brazilian student scored 413 points on reading, compared to an average of 487 points in all OECD countries.[1]

In response to the need for higher-quality education in Brazil, the implementing partner in this evaluation was launched in 2015 with the mission of improving literacy and writing skills among school-aged youth by applying artificial intelligence to linguistics in public schools. Its main product was a pedagogical program that provided feedback on writing skills to students, using an automated writing evaluation (AWE) system combined with validation of feedback by human essay graders. The AWE system was embedded in an online platform that granted students access to in-classroom practice opportunities for the essay of the National Secondary Education Exam (*Exame Nacional do Ensino Médio,* ENEM).

ENEM is the second largest college admission exam in the world. It has been increasingly used as an entry exam by many post-secondary institutions in Brazil, and the essay portion of it accounts for the largest share of the public-private school achievement gap in ENEM scores. In allowing teachers to reallocate their time from grading to providing more tailored assistance to students, the AWE system could potentially improve students' writing skills and their performance on the exam, helping close the opportunity gap between public and private-school students.

One of the advantages of the provider's AWE technologies was that they were based on a platform that worked well with poor internet connections. Given the low cost of sharing online access to automated essay scoring, this algorithm could represent a cost-effective way of improving writing skills among school-aged youth, even in contexts of low internet connectivity.

Brazilian students with their laptops using the AWE system.

Photo: Vinicius Valpereiro | J-PAL

Details of the intervention

Researchers partnered with the implementer to measure the impacts of two AWE-based programs on students' writing skills. The evaluation took place in 178 public schools with about 19,000 students who had computer access in the state of Espírito Santo. Schools were randomly assigned to one of three groups:

1. *Enhanced AWE intervention (55 schools):* Students in this group had access to the *ed* tech initially offered by the implementer, which combined feedback generated by the AWE system and essay grades provided by human scorers. The students were assigned five ENEM training essays in the platform, which gave them instantaneous feedback on syntactic text features, such as spelling mistakes and the use of informal language, and with an estimated essay score based on a performance bar composed of five levels. About three days after submitting their essays, the students received a grade from a human scorer, as well as comments on the five skills valued in ENEM. For simplicity, the five abilities can be broadly classified into three categories: syntactic skills, analytical skills, and policy proposal skills. As in the official exam, each of the five competencies were valued by graders on a 200-point scale, adding to up to 1,000 points in a full score.
2. *Pure AWE intervention (55 schools)*: Students in this group had access to the same *ed* tech platform, but did not receive additional feedback from human graders. As in the enhanced AWE intervention, the students received instantaneous feedback on text features as well as the five-level performance assessment right after submitting their essays. However, instead of being scored by a person a few days after, the students were instantaneously presented the AWE system's predicted grade on a 1,000-point scale and were provided feedback selected from the implementers' database of standard feedback.
3. *Comparison group (68 schools):* Students did not have access to the provider's platform.

The State's Education Department selected schools to participate in the evaluation based on a 2017 survey on proneness to online technology adaptation. These 178 schools received 8,000 laptops between February and April of 2019 to ensure adequate computer availability for the implementation of the *ed* techs, regardless of whether they received the intervention.

The primary goal of the evaluation was to document the impacts of the two *ed* tech systems on ENEM essay scores. By comparing the two interventions, researchers assessed if incorporating additional inputs from humans improved grading and feedback quality on aspects in which AI may fall short. Given that the human grading component is expensive and challenging to scale up, it was important to understand if the potential benefits of the enhanced AWE system relative to the pure intervention were worth their cost.

Results and policy lessons

Teachers shifted their work hours from routine (e.g., searching for orthographic mistakes) toward nonroutine tasks (e.g., providing individual assistance on essay consistency) with the adoption of both the enhanced and pure AWE systems, and students' essay scores improved as a result. Having human graders as a resource to improve grading and feedback quality did not lead to larger improvements in essay scores than the AWE system alone, despite increasing perceived feedback quality.

Take-up: Take-up was high and similar across the two *ed* tech systems for both teachers and students. In the two intervention groups, more than 95 percent of teachers used the *ed* techs to assign and collect essays in each of the five essay practices. Student compliance was also similar across interventions and relatively high. For each writing activity, 75 to 80 percent of students enrolled in the intervention schools submitted essays through the platform.

Impact on essay practice and quantity and quality of feedback received: Students in both intervention groups wrote more training essays, received more feedback on their essays, and perceived the feedback received as higher -quality. Students increased their essay practice and wrote 1.4 more training essays when using the enhanced AWE *ed* tech (a 29 percent increase relative to the baseline of 4.9 essays) and 1.6 more essays when using the pure AWE *ed* tech (a 32 percent increase) relative to the comparison

group. In addition to practicing more, students in both interventions benefited from comments and notes in about 1.3 more essays (a 40 percent increase relative to 3.4 essays on average in the comparison group) and received a grade on an additional 1.6-1.7 essays (an increase of about 45 percent).

Students exposed to either of the two *ed* techs were also 6-7 percentage points more likely to find comments and annotations *somewhat* useful (from an average of 81 percent in the comparison group), but only those using the enhanced AWE *ed* tech were more likely to find the comments on their essays *very* useful (a 6 percentage point increase relative to a base of 44 percent in the comparison group). The difference in effects on perceived feedback quality between the two groups was meaningful, suggesting that the human graders in the enhanced AWE system did contribute to higher feedback quality.

Impact on teachers' pedagogy and time allocation: Students using either of the two *ed* techs discussed roughly 35 percent more essays individually with their teachers after they received grades. However, only teachers in the enhanced AWE group reported having more time to cover the topics of writing, grammar, and literature in school. In line with these results, the share of teachers who felt their time was very insufficient dropped from 23 percent in the comparison group to 9 percent for teachers using the enhanced AWE system, but remained roughly unchanged for teachers using the pure AWE *ed* tech. Taken together, these findings suggest that teachers in the enhanced AWE group delegated part of their gradings tasks to human graders, while teachers in the pure AWE arm were able to keep pace by taking over some of these tasks, without increasing their usual workload.

Impact on essay scores and writing skills: The enhanced and the pure AWE interventions had positive and almost identical effects on the full ENEM essay score and led to similar improvements in scores on the writing skills evaluated by official graders, except for analytic skills. Both interventions increased the full essay score by about 0.09 standard deviations, mitigating an estimated 9 percent of the public-private school gap in essay scores. Specifically, syntactic skills scores improved by 0.07 and 0.06 standard deviations for students in the enhanced and pure AWE interventions, respectively; the policy proposal grade increased by 0.16 standard deviations with the enhanced AWE *ed* tech and 0.14 standard deviations with the pure AWE *ed* tech; and the analytic skills scores increased by 0.06 standard deviations with the pure AWE *ed* tech, but was not affected by the enhanced AWE intervention.

The similarity in effects across interventions suggests that the additional inputs from human graders did not change the extent to which the *ed* techs were able to improve scores. While this was expected for syntactic skills, which are graded based on standardized processes (e.g., parsing content and catching errors), it was not anticipated for the more complex parts of the exam (e.g., analytical skills and the policy proposal grade). Likely, the shifts in time allocation allowed teachers to fill in some of the gaps or limitations of the pure AWE *ed* tech relative to the enhanced platform.

In short, the study presents evidence that artificial intelligence can help overcome bottlenecks that prevent the accumulation of writing skills; in particular, teachers' time constraints to provide individualized assistance to students. Pure and enhanced AWE *ed* techs led to similar effects on essay scores and writing skills, suggesting that including human scorers was not necessary to attain the same results. This makes the intervention less costly and easier to scale.

From 2020 onwards, the results from this study were used by the State Education Department of Espírito Santo to justify procuring the pure AWE tool. The program is currently a public policy for senior high school students in the state, potentially benefiting close to 30,000 senior high school students every year.

Ferman, Bruno, Lycia Lima, and Flavio Riva. "Artificial Intelligence, Teacher Tasks and Individualized Pedagogy". Working Paper, March 2021.

---

1. "PISA 2018 Results: Combined Executive Summaries, Volume I, II and III", p. 18 (Table I.1, 2/2) Available at: https://www.oecd.org/pisa/Combined_Executive_Summaries_PISA_2018.pdf