Experimental Evidence on Artificial Intelligence in the Classroom

Ferman, Bruno and Lima, Lycia and Riva, Flavio

Sao Paulo School of Economics - FGV, Sao Paulo School of Business Administration - FGV, Sao Paulo School of Business Administration - FGV

4 November 2020

Experimental Evidence on Artificial Intelligence in the Classroom*

Bruno Ferman† Lycia Lima‡ Flavio Riva§

First draft: November 4th, 2020

Abstract. This paper investigates how technologies that use different combinations of artificial and human intelligence are incorporated into classroom instruction, and how they ultimately affect students' outcomes. We conducted a field experiment to study two technologies that allow teachers to outsource grading and feedback tasks on writing practices. The first technology is a fully automated evaluation system that provides instantaneous scores and feedback. The second one uses human graders as an additional resource to enhance grading and feedback quality in aspects in which the automated system arguably falls short. Both technologies significantly improved students' essay scores, and the additional inputs from human graders did not improve effectiveness. Furthermore, the technologies similarly helped teachers engage more frequently on nonroutine tasks that supported the individualization of pedagogy. Our results are informative about the potential of artificial intelligence to expand the set of tasks that can be automated, and on how advances in artificial intelligence may relocate human labor to tasks that remain out of reach of automation.

JEL Codes: I21, I25, I28, J22, J45.

*The authors would like to acknowledge helpful comments from David Autor, Erich Battistin, Leonardo Bursztyn, Guilherme Lichand, Cecilia Machado, Marcela Mello, Vítor Possebom, João Pugliese, Rodrigo Soares, Michel Szklo and Thiago Tachibana that substantially improved earlier versions of this draft. This project would not have been possible without the collaborative efforts of the Espírito Santo's Education Department (SEDU/ES). We also thank the Lemann Foundation for supporting the implementation of the interventions; the staff at Oppen Social, and specially Ana Paula Sampaio, Andressa Rosalém, Camille Possatto and Elionai Rodrigues, for carefully implementing the teachers' survey; the Centro de Políticas Públicas e Avaliação da Educação at Universidade Federal de Juiz de Fora (CAEd/UFJF), and specially Manuel Palacios and Mayra Moreira de Oliveira, for all the assistance with the implementation of the writing tests in public schools in Espírito Santo. Finally, we are deeply indebted to the implementer's staff, who made the design of the experiment possible and helped us tirelessly in the various stages of this research. We gratefully acknowledge financial support from J-PAL through the Post-Primary Education Initiative, which allowed the collection of primary data from students and teachers and the correction of the ENEM training essays and the biographical narratives. Maria Luiza Marques Abaurre and Daniele Riva provided helpful guidance on the background information necessary to understand the grading criteria for written essays and group them in the way we do in this paper. We uploaded a full pre-analysis plan at the American Economic Association Social Science Registry (AEARCTR-0003729). This research was approved by the Committee on the Use of Humans as Experimental Subjects (COUHES, Protocol #18115953228) at MIT and the ethics' committee at Fundação Getulio Vargas (FGV). The authors declare that they have no relevant material or financial interests that relate to the results described.

†Sao Paulo School of Economics — FGV, bruno.ferman@fgv.br, *corresponding author*.

‡Sao Paulo School of Business Administration — FGV, lycia.lima@fgv.br.

§Sao Paulo School of Business Administration — FGV, flaviorussoriva@gmail.com.

1 Introduction

The recent progresses in artificial intelligence (AI) changed the terms of comparative advantage between technology and human labor, shifting the limits of what can —and reviving the debate on what should— be automated. In educational policy circles, in particular, the now broad scope of applications of AI to linguistics prompted a controversy on automated writing evaluation (AWE) systems (see, for instance, the Human Readers Petition).[1] Central to the controversy on AWE is the ability of systems that are "*completely blind to meaning*" to emulate human parsing, grading and feedback behavior.[2] However, such controversy largely bypasses the fact that AWE may not be introduced in isolation. Following the rationale from, for example, Acemoglu and Autor (2011), by performing routine tasks previously thought to be out of reach from automation, AWE may induce a re-allocation of tasks between technology and human labor. In this context, AWE systems may be effective in improving even skills in which AI alone arguably still falls short in evaluating. Overall, from an economics perspective, the relevant question should not be whether AWE systems are able to perfectly emulate teacher's feedback and grading behavior, but whether such systems, when incorporated into instruction, can effectively improve students' outcomes.

This paper approaches these questions by investigating how educational technologies (*ed* techs) that use different combinations of artificial and human intelligence are incorporated into instruction, and how they affect students' outcomes. We present the results of a randomized field experiment with 178 public schools and around 19,000 students in Brazil. The 110 treated schools incorporated one of two *ed* techs designed to improve scores in the argumentative essay of the National Secondary Education Exam (ENEM). These *ed* techs differ in the way they combine artificial intelligence and external human support in order to alleviate Language teachers' time and human capital constraints. Time constraints, in particular, tend to be more binding for Language teachers handling large written essays such as the ENEM essay, which require time-intensive tasks (Grimes and Warschauer, 2010). In post-primary education, given that instruction needs to contemplate relatively advanced topics, teachers' human capital is also likely a limitation to building writing skills (Banerjee et al., 2013).

Both *ed* techs rely, to some extent, on an AWE system embedded on an online platform with low Internet requirements. The *first* ed *tech* ("enhanced AWE") uses the system's ML score to instantaneously place students on a bar with five quality levels and to provide information on syntactic text features, such as orthographic mistakes and the use of a conversational register ("writing as you speak"). The system withholds the ML score and, about three days after submitting essays, students receive a final grading elaborated by human graders hired by the implementer. This grading includes the final ENEM essay score, comments on the skills valued in the exam and a personalized comment

[1] At its core, AWE uses: (i) natural language processing to extract syntactic, semantic and rhetorical features related to essay quality, and (ii) machine learning (ML) algorithms to generate scores and allocate feedback based on these features.

[2] The quoted expression is taken from McCurry (2012) (p. 155), who also presents a rich description of the controversy on AWE. Essentially, critics argue that pure AWE systems cannot measure the essentials of good writing and might make writing unnecessarily more prolific by taking linguistic complexity for complexity of thought. The Human Readers Petition provides further criticism on the use of machine scoring in high-stakes assessment calling upon schools to "STOP using the regressive data generated by machine scoring of student essays to shape or inform instruction in the classroom" and to "STOP buying automated essay scoring services or programs in the counter-educational goal of aligning responsible classroom assessment with such irresponsible large-scale assessment".

on essay quality. Overall, the main goal of incorporating additional inputs from humans is to enhance grading and feedback quality on aspects in which AI may fall short. The *second* ed *tech* ("pure AWE") uses only AI to grade and provide feedback, without the participation of human graders. As in the enhanced AWE treatment, students are placed on the quality bar and receive information on text features right after submitting essays, but are also presented to the system's predicted score and to a suitable feedback selected from the implementers' database.

There are several features of our setting that make it interesting to study AWE-based *ed* techs. First, the essay grading criteria range from lower-level skills, such as the command over orthography, to rather complex skills, such as the ability to interpret information and sustain a coherent point of view. These tend to be the bulk of the skills considered in the evaluation of argumentative written essays (Barkaoui and Knouzi, 2012; McCurry, 2012). Importantly, the criteria encompass both skills for which pure AWE systems are arguably good at evaluating and skills that such systems would fall short in capturing. Thus, one could expect the additional inputs from human graders to differentially affect different types of skills that are evaluated in the exam. Second, ENEM is the second largest college admission exam in the world, falling shortly behind the Chinese *gāokǎo*. In 2019, the year of our study, roughly 5 million people and 70% of the total of high school seniors in the country took the exam, which acts as a key determinant of access to higher education. This speaks to the potential effects of scaling up these *ed* techs which, at least for the pure AWE system, would be relatively low cost. Finally, the gap in ENEM scores between public and private students is substantially larger for the essay when compared to the other parts of the exam. Thus, in our context, both technologies could make public school students more competitive for admission into post-secondary institutions.

The primary goal of the experiment was to describe and compare the effects of both *ed* techs on ENEM essay scores. We find that the enhanced AWE improved scores in roughly 0.1σ. The total effect is channeled by improvements in all skills evaluated in the exam: (i) *syntactic skills* (0.07σ), which comprise the command over the formal norm of written language and the skills that allow one to build a logical structure connecting the various parts of the essay; (ii) *analytical skills* (0.04σ), which comprise the skills necessary to present ideas that are related to the essay motivating elements and develop arguments to convince the reader of a particular point of view; (iii) *policy proposal skills* (0.16σ), which capture the ability of students to showcase critical-thinking by proposing a policy to the social problem that figures, in each year, as the topic of the essay. Surprisingly, the point estimates and coverage of confidence intervals for the pure AWE *ed* tech are virtually the same. Therefore, we find evidence that the additional inputs from human graders did not improve the effectiveness of the *ed* techs to improve scores that capture a broad set of writing skills.[3] Using the essay public-private achievement gap to benchmark magnitudes, we find that both *ed* techs mitigate 9% of the gap. In the policy proposal score gap, which is currently at a high 80%, the effects imply a reduction of 20%. Since language's most sensitive period of development happens before adolescence and tends to be less responsive to variation in educational inputs (see Knudsen et al., 2006), we consider that these are economically meaningful effects. From a policy perspective, the fact that we find no differential effects across arms bears relevance, as expanding the program without human graders is substantially simpler and cheaper.

[3] In the estimation, we pooled information on both the official ENEM 2019 exam and on an independently administered essay with the same structure. The effects are similar but less precisely estimated if we consider the two data sources separately.

Using primary data on students, we find that the *ed* techs improved the quantity of training and feedback. Both treatments increased the perceived quality of the feedback, with stronger effects for the enhanced AWE *ed* tech. This is perhaps expected given the semantic nuances and the complexity of most of the skills valued in the exam. Finally, we show that both technologies increased the number of ENEM training essays students discussed with their teachers. If teachers completely delegated essays' training to the *ed* tech, then we should expect that treatments would decrease the number of essays students discussed with their teachers. In contrast, we find evidence that both technologies helped teachers engage more frequently on nonroutine tasks that support the individualization of pedagogy.

Despite the similar results on impacts and mechanisms —apart from the somewhat expected leverage in feedback quality that human participation entails—, we find suggestive evidence that teachers adapted differently to the introduction of the *ed* techs using teacher-level primary data. Teachers using the enhanced AWE *ed* tech perceived themselves as less time-constrained to deliver the language curriculum material, and adjusted hours worked from home downwards. Teachers in the pure AWE *ed* tech arm were not impacted by the introduction of the *ed* tech in any of these margins. At face value, these results suggest that teachers in the pure AWE treatment arm took over some of the work that the the additional input of human graders provided in the enhanced treatment. However, we cannot rule out that these differences were due to differential attrition in the teachers' survey.

In addition to describing effects on primary outcomes and mechanisms, we consider indirect effects of the *ed* techs on other learning topics. Specifically, we discuss whether our data are consistent with: (i) positive or negative spill-overs to the narrative textual genre, which could come, for instance, from improvements in skills that are common to all genres (like orthography) or from adverse effects of "training to the test"; (ii) positive or negative effects on subjects related to Language, which could arise from complementarities with writing skills or increases in motivation to take the ENEM essay; (iii) positive or negative effects on subjects unrelated to Language (such as Mathematics), which could arise, once again, from an increase in motivation or a crowding out in effort due to an increase in essays' training. Across all families of learning subjects, we find statistically insignificant results. Since we pool several sources of data, we are able to reject even small adverse effects in each of these families of outcomes, suggesting that the effects of the *ed* techs were restricted to their main goal of improving ENEM essay scores.

These findings add to a growing literature on *ed* techs and on the effects of technology on instruction. To the best of our knowledge, this is the first impact evaluation of a pure AWE system — a learning tool widely used in the US (McCurry, 2012) — that uses a credible research design and a large sample to illustrate how these technologies are incorporated into instruction and affect students' outcomes.[4] More broadly, in face of the somewhat mixed evidence of the *ed* tech literature (Bulman and Fairlie,

[4]In particular, we are not aware of any impact evaluation that does so in a post-primary education context. Outside post-primary education, Shermis et al. (2008), Palermo and Thomson (2018) and Wilson and Roscoe (2019) use experimental data on grades 6-10. However, we believe there are important limitations in the results presented in these papers. First, the main outcomes in Shermis et al. (2008) and Palermo and Thomson (2018) are variables generated by the automated systems system, which will introduce severe measurement error in skills if treated students have higher ability to game the system in order to receive better scores. Second, in both papers randomization was conducted at the individual level, which has important implications on the way the AWE systems are integrated into instruction and raises serious concerns about spill-overs. Most outcomes in this literature are also not economically important. Wilson and Roscoe (2019) present an evaluation of the effects *Project Essay Grade Writing* in Texas on the state English Language Arts test but treatment was randomized using a very small sample of clusters (3 teachers in 10 different classrooms) and the control group received a recommendation of using Google Docs as an alternative resource.

2016), Muralidharan et al. (2019) argue that "*realizing the potential of technology-aided instruction to improve education will require paying careful attention to the details of the specific intervention, and the extent to which it alleviates binding constraints to learning*". The *ed* techs we analyze were designed to alleviate important binding constraints in our setting — most importantly, time and human capital constraints—, and feature most of the promising channels of impact of *ed* techs discussed by Muralidharan et al. (2019).[5] The positive effects we find, and a detailed analysis of mechanisms, corroborate and illustrate the conclusion from Muralidharan et al. (2019). Finally, a comparison between the two treatment arms provides evidence that teachers' human capital was not a binding constraint for the implementation of the pure AWE technology, as we found no evidence that the additional inputs from human graders improved the effectiveness of the program.[6] This is also an important result from a policy perspective, as scaling up an *ed* tech like the enhanced treatment would necessarily entail large marginal hiring and monitoring costs.

Our attempt to understand the effects of the programs on teachers' time allocation also connects our contributions to the literature on the effects of technological change on the labor market. In a seminal paper, Autor et al. (2003) argue that computer-based technologies substitute human labor in routine tasks —i.e., those that can be expressed in systematic rules and performed by automates— and complement human labor in nonroutine abstract tasks (also, see Acemoglu and Autor, 2011). AWE systems added marking essays with a focus on syntax and identifying linguistic structures to the ever expanding set of routine tasks. The question of whether AI will eventually be able to interpret written content remains, to this day, speculative. Despite such limitation, both *ed* techs reduced the burden of routine tasks, and shifted teachers' classroom activities toward nonroutine tasks: personalized discussions on essay quality.[7] In a sense, we find contextual support and one of the first pieces of evidence for the optimistic prediction that "*AI* [...] *will serve as a catalyst for the transformation of the role of the teacher* [...] *allow*[*ing*] *teachers to devote more of their energies to the creative and very human acts that provide the ingenuity and empathy to take learning to the next level.* " (Luckin et al., 2016, p. 31).

Finally, we contribute to the small set of papers that take writing skills as outcomes of interest. While there is a large number of papers in the *ed* tech literature (and educational programs, more generally) that use Language and Mathematics multiple-choice test scores, research efforts are much

[5] On *ed* techs' mechanisms in general, the authors posit that "*[a] non-exhaustive list of posited channels of impact [of* ed-*techs] include using technology to consistently deliver high-quality content that may circumvent limitations in teachers' own knowledge; delivering engaging (often game-based) interactive that may improve student attention; delivering individually customized content for students; reducing the lag between students attempting a problem and receiving feedback; and, analyzing patterns of student errors to precisely target content to clarify specific areas of misunderstanding.*" (p. 1427, fn. 1, our emphasis). Most of the listed features are essential and ever improving features of AI-based *ed* techs. Thus, besides presenting credible evidence on the effects of AWE systems on students, our results illustrate general mechanisms that make AI promising to support teachers in tasks to foster skills.

[6] Given our research design, it is not possible to distinguish whether (i) the AWE system needs to be complemented by human intelligence and school teachers played this role in the pure AWE program, or (ii) the AWE system would have been effective regardless of teachers' inputs. Given our evidence that teachers did not completely delegate instructions, and actually increased the amount of pedagogy individualization, we believe alternative (i) is more likely.

[7] While concluding something about the educational production function would require a design that exogenously varies the amount of inputs from *ed* techs conditional on it's implementation (such as in Bettinger et al., 2020), the results we find are inconsistent with complete substitution.

rarer for writing skills.[8] This is perhaps surprising, considering the ubiquity of tasks that demand writing skills in universities and in the labor market.

The remainder of the paper is structured as follows. Section 2 provides background information on the experiment's setting and on the *ed* techs we study. Section 3 describes some anticipated mechanisms we specified in the pre-analysis plan and which guided our data collection. Section 4 discusses the research design and its validity, along with the data and the main econometric specifications. Section 5 presents the main findings. Section 6 concludes the paper.

2 Context and Experimental Arms

2.1 Background

2.1.1. **ENEM**. The National Secondary Education Exam ("Exame Nacional do Ensino Médio", ENEM) is a non-compulsory standardized high-stakes exam that acts as a key determinant of access to higher education in Brazil. The exam is currently composed of 180 multiple-choice questions, equally divided into four areas of knowledge (Mathematics, Natural Sciences, Language and Codes, Human Sciences), and one written essay. The large gap between public and private schools' quality in Brazil is salient in all ENEM tests and, in particular, in the essay. The upper graph in Figure 1 describes the private school premium using data on the universe of high school seniors in ENEM 2018. Although the achievement gap is a significant feature of all parts of the exam, it is remarkably larger in the written essay (at 43%) when compared with the multiple-choice tests (at 13-21%). When compared to the multiple-choice Portuguese Language test, which measures other dimensions of literacy, the gap in the essay is more than three orders of magnitude larger. The contribution of the essay to the total achievement gap is 39%, compared to 21% and 12% in multiple-choice tests in Mathematics and Language, respectively. All these facts highlight the importance of policy interventions that may affect ENEM essay scores and make public school students more competitive for admission into post-secondary institutions.

2.1.2. **ENEM Argumentative Essay**. The main topic of the essay varies from year to year and is always introduced by excerpts, graphs, figures or cartoons that frame an important social issue. Since its creation, ENEM has proposed several polemic topics, which typically attract broad attention from the media: for example, the use of Internet data to manipulate consumers (2018), gender-based violence (2015), the limits between public and private behavior in the 21st century (2011), the importance of labor for human dignity (2010), how to stop the Amazon deforestation (2008) and child labor (2005).

In 2019, the year of our study, the topic of the official ENEM 2019 essay was "Democratization of Access to Cinema in Brazil". The first motivating element described a public exhibition of a

[8] To the best of our knowledge, research on the topic has been restricted to early childhood and preschool interventions, settings where the measurement of these skills is obviously conducted at a very basic level. Some examples are the READY4K! text messaging program for parents (York and Loeb, 2014; Doss et al., 2018) and the well-known center-based early childhood education program Head Start (Puma et al., 2005). York and Loeb (2014) and Doss et al. (2018) measure writing skills as the ability of writing one's name and upper-case letter knowledge and Puma et al. (2005) uses the ability to write letters. In a comprehensive review of experimental research on interventions to improve learning at later ages, Fryer (2017) describes several experiments (Morrow, 1992; Pinnell et al., 1994; Mooney, 2003; Borman et al., 2008; Somers et al., 2010; Kim et al., 2011; Jones et al., 2011, among others) with treatments directly aimed at improving writing skills, but the outcomes evaluated are almost exclusively reading test scores.

movie in 1895 and the skepticism of Lumière on the potential of cinema for large-scale entertainment. The second one presented a definition of cinema as a "mirror-machine", elaborated by the French philosopher and sociologist Edgar Morin. The third one described how the last years in Brazil have witnessed a secular concentration of the movie theaters in large urban centers. Finally, the fourth and last motivating element was an info-graph presenting statistics on movie consumption on television and movie theaters. At the top of the page, as in every year since the creation of ENEM in 1998, students were instructed to write an essay following the argumentative textual genre using the motivating elements as a start point and mobilizing knowledge acquired during their formation period. We now discuss how the official graders attribute scores to students facing this task.

Measurement System for Writing Skills. A successful handwritten essay begins with an introductory paragraph, followed by two paragraphs with arguments that underlie a point of view or thesis on the social problem and a final paragraph featuring a policy proposal. The five writing competencies (INEP/MEC, 2019) evaluated by graders of the ENEM essay are:

- *syntactic skills*, which comprise:
  - exhibiting command of the formal written norm of Brazilian Portuguese [*200 points*];
  - exhibiting knowledge of the linguistic mechanisms that lead to the construction of the argument [*200 points*];
- *analytic skills*, which comprise:
  - understanding the proposed topic and applying concepts from different areas of knowledge to develop the argument following the structural limits of the dissertative-argumentative prose [*200 points*];
  - selecting, relating, organizing and interpreting information, facts, opinions and arguments in defense of a point of view, using pieces of knowledge acquired in the motivating elements and during the schooling [*200 points*];
- *critical-thinking and policy proposal*, which comprise:
  - elaborating a policy proposal that could contribute to solve the problem in question, respecting basic human rights [*200 points*];

Each of the five competencies is valued by graders on a 200 points scale with intervals of 40 so that the full score ranges from 0 to 1000.

As specified in the pre-analysis plan, we study these five competencies aggregating them into these three different categories, which we refer to as skills hereafter. The first competency is the command over the formal norm of the written language, which comprises, among other things, the precise use of vocabulary, correct orthography, verbal concordance and the use of the neutral register — as opposed to the informal of "conversational" register typical of oral or intimate communication. The second competency relates to the student's ability to build a logical and formal structure connecting the various parts of the essay. Students are thus evaluated in terms of their capacity of establishing relations using prepositions, conjunctions and adverbs building a "fluid" text within and across paragraphs. Jointly considered, these two competencies characterize the "surface" (INEP/MEC, 2019, p. 21) of the text

or aspects that linguists call *syntactic*. The next two competencies, on the other hand, are directly related to the meaning conveyed by the student essay. They require that test takers present ideas that are related to the essay topic and develop arguments in order to convince the reader of a particular point of view, displaying a set of *analytical* skills. These benefit not only students that "write well" but also students that have a more solid educational background and can leverage potential synergies with other topics covered in the post-primary education curriculum. Finally, the fifth and last writing skill evaluated is the ability of students to showcase critical and practical-thinking by elaborating a *policy proposal* in response to the point of view presented. While the grouping is based on our own interpretation of the grading criteria, it was motivated by interactions with the implementers' staff and by our reading of the specialized literature on writing assessments in Brazil and elsewhere (specially Abaurre and Abaurre, 2012; Neumann, 2012).

The private school premium helps validate the way we chose to group competencies. Above, we highlighted that differences in the essay score account for the largest share of the public-private achievement gap in ENEM. The bottom graph in Figure 1 break down this premium for each of the competencies and skills presented above. Notably, the gap seems to be increasing in the skill complexity or sophistication, starting at 23-30% for the syntactic aspects of the text (similar to Mathematics), reaching roughly 50% and a large 80% for the analytic skills and the policy proposal, respectively.[9]

ENEM Essay Grading. The graders hired by the Ministry of Education are bachelor degrees in areas related to Language and personally correct the digitized version of the handwritten essays using an online platform. Training happens a couple of months before the test and consists nowadays of a two-day course where the writing skills and the grading criteria are discussed and specific examples on how to use the writing rubric on a set of excerpts are presented. In the first step, each essay is graded by two different persons. If the two grades disagree on more than 100 points in total or on more than 80 points on at least one skill, a third grader with high agreement rates is assigned automatically and grades the essay. If the third grader agrees (in the sense described above) with at least one of the two initial graders, the final score is the simple average between the two closest scores given to the written essay. If the disagreement is not solved with the participation of a third-party, the essay is sent to a board of experienced graders which meet to correct these essays in person.

2.2 Treatment Arms

The implementer of the *ed* techs was created in 2015 and its main goal is to improve writing and address the "literacy gap" in Brazil. The main current product is an *ed* tech based on an online platform that corrects and provides feedback on ENEM written essays using an AWE system (Fonseca et al., 2018) supervised by independently hired human graders. The next paragraphs describe the main current *ed* tech developed by the implementer (hereafter, enhanced AWE *ed* tech) and an alternative treatment that removes human graders from the algorithm supervision tasks, letting the artificial intelligence "do all the work" at once (hereafter, pure AWE *ed* tech).

2.2.1. **Enhanced AWE *ed* tech**. Even though access to the online platform can be done indepen-

[9]The correlation between scores in competencies also help us validate our grouping of competencies. In ENEM 2018, the correlation between scores in the first two skills is very large (0.82), as is the one between the next two (0.94). In interactions with the implementer's staff, we learned that one of the their priors is that these skills are jointly developed in the formation of writing skills.

dently by students outside the school, the implementer tries to provide teachers with an instrument to support the development of writing skills inside the classroom. The program spans the academic year of high school seniors and consists of five ENEM writing practices elaborated by the implementer. The integration of these writing practices to the other activities in the Portuguese Language course is discretionary, but essays are scheduled to happen in predefined time intervals. In 2019, the topics of the ENEM practice essays, in order of appearance, were:

- The Challenges of Integrating Technology with Instruction in Brazilian Schools;
- Communication in the Internet Era: Freedom or Intolerance;
- The Challenges of Current Work Conditions in Brazil;
- The Escape from Hunger and Famine in Brazil;
- Art and Culture for Social Transformation.[10]

Students' Interface. During a practice, the platform saves essays automatically and frequently to prevent students from missing their work upon problems with the computers or the Internet connection. After the submission, the platform interacts instantaneously with the student providing a comprehensive set of text features used to compare the essay to "goals" that would bring the student closer to achieve a perfect score. Some examples are: number of words, number of spelling mistakes and uses of informality tones, intervention elements and social agent markers. This immediate screening of the text also allows for a quick test of plagiarism and the student is advised to work on her text by studying with the help of online materials that are elaborated internally. At this point, the student is also presented to a signal of achievement based on the AWE predicted essay score, displayed in a performance bar with 5 levels.

Human Graders. The enhanced program withholds the AWE predicted score and provides students with a rough approximation thereof to avoid introducing noise in the evaluation process. The final score shown to students is set by human graders independently hired on a task-based contract that pays 3.50 Reais (approximately 2019 US 0.85 dollars) *per* essay. The graders have access to the essay with all the information on text features shown to students and choose a value, ranging from 0 to 200 (in 40 point intervals) without seeing the grade predicted by the ML algorithm. When a human grader chooses a score for a skill, their interface suggests a randomly chosen comment taken from a database of textual interactions chosen by the implementer, which are pre-adapted to the quality of a student in a given skill. The essay can also be personally annotated and the comments' colors are associated with each of the exam's skills. Finally, the human graders must leave a general comment on the essay, the last step before submitting the final grading back to students. The whole process takes, on average, three business days. To boost engagement students receive a text message when their final grading is available in the platform.

Teachers' Interface. During a writing activity, the ongoing essays are presented along with a progress bar, where teachers can follow the progress of students on the writing task and monitor if they have

[10]Since, by chance, there was some similarity between the topic of the last writing practice and the topic of the 2019 ENEM essay, in Section 5 we will discuss the potential direct influences that these writing topics may have exerted on the performance of students.

logged in, started writing and finished the task. The system also shows descriptive statistics on common grammar mistakes made by students in real time. Each teacher also has access to a personal dashboard with shortcuts to Excel files containing the aggregate data on enrolled students and their individual scores on the essay and skills for a given activity. Teachers also have access to the full individual essay gradings and are absolutely free to base or not their students' scores on Portuguese Language on one or several of the outputs of the platform.

2.2.2. **Pure AWE *ed* tech**. In this treatment arm, the user experience is fully based on the instantaneous outputs from the AWE system. Thus, this *ed* tech explores the possibility that a pedagogical program could be based only on information that is generated by the artificial intelligence and is currently withheld for a supervision step. The students' and teachers' interface are very similar to the one in the enhanced program, but as students submit their essays, they are presented instantaneously to the full essay score predicted by the algorithm and to comments on each skill, randomly selected in the implementers' database conditional on each predicted skill score. The interest in testing the effectiveness of the pure AWE treatment was twofold. First, currently, expanding the program entails large marginal hiring and monitoring costs of human graders. Second, the implementer knows that a significant share of students do not come back for their final grading, so there is value in gauging the net effects of reducing the lag between students attempting a writing task and receiving formative feedback even if this feedback is of arguably lower quality.

3 Hypotheses and Mechanisms

The *ed* techs described in Section 2 aim to change the nature of part of writing instruction and to provide students with oriented opportunities of practicing for the ENEM essay. We now discuss the main hypotheses of the experiment and the mechanisms by which we anticipated the *ed* techs would affect students' outcomes.[11]

3.1 Main Hypotheses

As specified in the pre-analysis plan, our main hypotheses relate to whether the *ed* techs affect ENEM essay scores. This is an important outcome for public school students, as ENEM is a key mediator of access into college and the essay is responsible for the greatest share of the public-private achievement gap in the exam (Figure 1). Moreover, since this gap is unevenly distributed across writing skills and seem to be increasing in how sophisticated they are, we were also interested in the effects of the *ed* techs on scores on skills that add up to the total essay score. We test the following hypotheses:

H_a^1 : *The enhanced AWE* ed *tech has an effect different from zero on ENEM essay scores.*

H_a^2 : *The pure AWE* ed *tech has an effect different from zero on ENEM essay scores.*

H_a^3 : *The* ed *techs have different effects on ENEM essay scores.*

[11] Whenever possible, we frame such mechanisms by referencing to promising features highlighted by the recent *ed* tech literature (see the literature review in Muralidharan et al., 2019, and, specially, fn. 1) and, more broadly, by the literature on educational policy.

As we discuss below, while our priors were that the main effects would be positive, we could not *a priori* rule out that the *ed* techs would have had adverse impacts on students. We also predicted mechanisms that would favor either the pure or the enhanced AWE *ed* techs. Therefore, we also had no priors on whether the differential effects of the two *ed* techs would be positive or negative.

3.2 Potential Mechanisms

3.2.1. **Training and Feedback**. The first channels of impact we anticipated are changes on the quantity of training and the quantity/quality of feedback provided to students.

First, as they reduce the time spent preparing and grading ENEM essay practices, the *ed* techs may help circumvent important time constraints faced by Language teachers and increase the number of essays assigned. The new inputs from the *ed* techs could, however, crowd out one by one the essays teachers would assign themselves, or even reduce the total number of essays written if teachers and students take more time to conclude a writing activity using the platform. The latter possibility could arise if there are major constraints on the capacity of schools to provide access to computers and an adequate Internet connection. For a discussion on the importance of these issues in the implementation of *ed*-tech programs in primary public schools in Brazil, see Ferman et al. (2019). Anticipating this potential bottleneck, the technology was developed so that the Internet requirements for using the platform are intentionally low.

Second, the two programs may affect the quality of the feedback. It is not *a priori* obvious whether the feedback would be better than the feedback students would receive from teachers in the absence of the programs. Importantly, whether they would improve quality should depend crucially on how teachers interact with the platform, and how traditional instructional tasks are re-distributed to the AWE systems and, in the enhanced AWE system, to human graders.[12] More specifically, it is possible, for example, that teachers' feedback for more complex skills would be better than the one provided by the pure AWE *ed* tech alone. However, it may be that teachers compensate the shortcoming of a pure AWE system, especially if their time constraints become less binding due to the AWE feedback. Overall, we anticipated that the enhanced AWE *ed* tech should provide better feedback relative to the pure AWE *ed* tech, particularly for more complex skills. However, the extent of such differences, and the differences relative to the control group, depend crucially on how these technologies are incorporated in the school.

Third, the *ed* techs could put teachers in a better position so as to engage in personal interactions with their students focusing on more complex aspects of essays.[13] Finally, the programs may affect the timing of the feedback. In this case, the pure AWE *ed* tech would have an advantage, as it provides instantaneous feedback to students. As stated by Muralidharan et al. (2019), this is an important potential advantage of *ed* techs.

[12] In particular, teachers could feel uncertain and end up not outsourcing some of their usual tasks to automated systems and/or to human graders they don't know personally. In other words, one could not rule out from start that we would face very low levels of compliance in the field. This seemed particularly relevant in the pure AWE treatment arm, which the school's staff may not understand, trust or, even worse, fear.

[13] Following Autor et al. (2003), we framed this anticipated mechanism as a complementarity between the system's parsing and grading tasks and teacher's nonroutine analytical (interpretation) and interactive tasks (providing in-person individualized advice). In sum, we hypothesized that the *ed* techs could support the individualization of pedagogy. Notice, however, that the *ed* techs could lead to a complete delegation of tasks to the new inputs, and even more if students use the enhanced AWE.

3.2.2. **Teachers' Time Allocation**. By altering traditional writing instruction, the treatments may change teachers' time allocation to different sets of tasks inside and outside the classroom. In particular, since the technologies allow teachers to outsource routine tasks of essay parsing and grading of syntactic features, they might induce a re-allocation of work time towards nonroutine tasks (such as preparing classes, correcting other homework or providing one-to-one tutoring). However, the freed up time outside the classroom may only be compensated by adjustments in extra-hours worked outside schools, which are common for Language teachers in our setting. Apart from changing time allocation to different tasks, the technologies could also alleviate teachers' feelings on being time constrained. Both sets of changes (behavioral and psychological) could end up affecting teachers' ability to help students improve ENEM essay scores.

3.2.3. **Expectations and Aspirations**. The integration of the *ed* techs may also affect teachers' expectations about their students' educational prospects and shift students' aspirations towards academic tracks after leaving high school. Both sets of changes, in turn, may affect teachers' instructional efforts and/or students' training effort for ENEM essay and other parts of the exam. First, teachers may consider that the *ed* techs are, indeed, working. As discussed in detail in our section on results, this is consistent with the fact that almost the entirety of teachers complied with the treatments in all activities. Second, over the year, teachers and students receive different information about writing quality than they would receive in the absence of the *ed* techs.

3.2.4. **Knowledge About Students**. The online platform provides teachers with summary statistics on their students (mainly, average score and evolution in each activity and skill) and with gradings on individual essays. If this information is a better or more engaging approximation to the real quality of essays than the one they would acquire themselves over time, the *ed* techs will accurately update teachers' beliefs about the "average" student, while at the same time highlighting important heterogeneities across students. The former process could affect the optimal targeting level of collective instruction (Duflo et al., 2011) and/or help teachers address the problem of facing various levels of writing quality.[14] Finally, improved knowledge about students may trigger important synergies between the content in classrooms dedicated directly to Grammar and Literature, generating positive spill-overs on other aspects of language more commonly captured in multiple-choice tests.

Overall, while some of the anticipated changes are consistent with the *ed* techs leading to improvements in ENEM essay scores (H_a^1 and H_a^2), one cannot rule out that they wouldn't be able to do so. Also, given the large differences in structure between *ed* techs discussed above, we didn't have strong priors on the the sign of differential effects (H_a^3). For these reasons, we worked with two-sided hypothesis tests in our analysis of primary outcomes.

3.3 Secondary Hypotheses

While our primary goal was to estimate the effects of the *ed* techs on ENEM essay scores and to identify their most important channels of impact, we also considered that they could have positive or negative spill-over effects on a broader set of outcomes. More specifically, we also test:

[14] We find support for the fact that the latter process may be very important in the case of writing: not only the variance of the distribution of ENEM essays in 2019 is two to three larger than the dispersion of multiple-choice Portuguese exam, the dispersion of residuals of performance in ENEM 2008 after absorbing school fixed effects.

H_a^4 : *The* ed *techs have an effect different from zero on scores capturing writing skills in other textual genres.*

H_a^5 : *The* ed *techs have an effect different from zero on scores in topics related to literacy.*

H_a^6 : *The* ed *techs have an effect different from zero on scores in topics not related to literacy.*

Once again, there are reasons why one may find that the effects in each family of outcomes are ambiguous. First, we consider effects on writing scores or skills in other textual genres (H_a^4). On the one hand, when training for the ENEM essay, students may practice more writing and receive more feedback generating spill overs. This would arguably be more important in the ones that are not genre-specific, such as the command over the formal written norm, and less so in the ones that are genre-specific. On the other hand, treatments may hinder the development of these skills if the feedback the *ed* techs provide is too specific for the ENEM essay and students end up "training to the test" and worsening their performance in different writing tasks. Considering this possibility is important both from a general perspective, but also because different textual genders are used by other post-secondary institutions as criteria to admit students.

It is also difficult to pin down the effects on language-related or non-related skills and test scores, such as Mathematics and Language (H_a^5 and H_a^6). On the one hand, the *ed* techs may crowd out time and effort used to study other subjects inside and outside the classroom. On the other hand, improvements in writing skills may be complementary to other subjects, like reading, which could possibly reflect in multiple-choice Language scores. For language (non-writing) skills, the program can also positively affect students' scores if it allows Portuguese teachers to better allocate their time to teaching other subjects, such as Grammar.

4 Research Design

This section describes our research design, emphasizing aspects that allow us to draw conclusions using the experimental data. We also discuss how we measured outcomes and gauged the mechanisms discussed in the last paragraphs and the main econometric specifications used in the analysis of results.

4.1 Sample Selection and Assignment Mechanism

4.1.1. **School Sample**. In March 2019, we received a list of public schools in Espírito Santo selected to participate in the experiment by the State's Education Department ("Secretaria de Estado da Educação", SEDU/ES). At that point, we learnt that the selection of schools used information from a 2017 survey on proneness to online technology adaptation. These schools received 8,000 notebooks between February and April of 2019 to ascertain that computer availability would not be a first-order concern for the implementation of the *ed* techs. Importantly, schools received notebooks regardless of treatment status.

Columns 1 to 3 in Appendix Table A.1 present comparisons between the universe of schools in Espírito Santo and the experimental sample of schools. As expected, considering the technology requirements used to build the experiment list, we find that 93% of schools in our sample have access to broadband Internet, against 80% in the whole state. In the microdata from the 2018 ENEM essay,

students in our sample also have slightly higher test scores.[15] All these characteristics are consistent with an important difference: there is only one rural school in our sample, while rural schools comprise around 7% of schools in Espírito Santo.

While the list of schools was not constructed to be representative, it comprises 68% of the urban public schools and around 84% of the students in urban public schools of the state. Therefore, we see our school sample as a good approximation to the population of urban school students in Espírito Santo.

4.1.2. **Randomization**. The final number of treated units in the first arm of experiment was chosen based on constraints in the implementer capacity of providing the enhanced AWE *ed* tech to more than 55 schools in 2019. The randomization used the following strata: (i) a geographical criterion, the 11 regional administrative units in the Espírito Santo state; (ii) the average score in the ENEM 2017 essay;[16] (iii) participation on an implementation pilot in 2018.[17] We used the median or quartiles of the average score in ENEM 2017 to split schools within an administrative unit and generated an independent stratum for the 6 schools that had no students taking the 2017 ENEM test.

The whole process of sample selection and randomization led to a total study sample size of 178 schools divided in 33 strata (of sizes 2 to 8), with 110 schools equally divided in treatment arms and 68 schools assigned to the control group.

4.2 Data

4.2.1. **Primary Outcome: ENEM Essay Scores**. To study our primary outcome of interest, we use de-identified administrative microdata on the 2019 ENEM essay scores. In addition, we partnered with the state's educational authority to collect an essay with the same textual genre and grading criteria of ENEM. One of Brazil's leading education testing firms ("Centro de Políticas Públicas e Avaliação da Educação", CAEd/UFJF) elaborated the proposal and graded the essays. Such essay was an additional part of the state's standardized exam, and was presented to teachers and students as an initiative of the state called *Writing Day*, not related to the experiment. As an incentive for students, all teachers in Espírito Santo were instructed to use the grade in this essay as a share of the final grade in Portuguese Language (8%). Hereafter, we refer to this set of primary data as "nonofficial" ENEM essay.[18]

The decision to collect the complementary data was based on several reasons. First, due to recent changes in the political landscape of Brazil and the dismantlement of part of the leadership of the

[15] Interestingly, we find much smaller differences in the policy proposal scores which is the most sophisticated writing skill captured by the ENEM essay scores.

[16] Information on the ENEM 2018 exam was not available at the time of the randomization.

[17] Only 5 schools in our sample were part of this pilot, which happened in the two last months of the second semester of 2018 (two writing activities). Our main intention was to understand better the behavior of the pure AWE *ed* tech and check whether it could sustain engagement over time. We created one independent stratum for these schools and kept their random treatment assignments. This decision was taken jointly with the implementer and SEDU/ES to minimize transitions that would lead to dissatisfaction among treated schools and awareness about the experiment.

[18] The topic of the essay was "The Construction of a National Brazilian Identity for the Portuguese Language". The first motivating text described spoken language as a cultural phenomena that dynamically builds the identity of a nation. The second one presented an argument from a Brazilian linguist in favor of the recognition of Brazilian Portuguese as a distinct language from the Portuguese spoken in Portugal. Some differences between the two languages are illustrated in the third motivating element. Finally, the fourth and last motivating element briefly argued that knowing how to write is an essential part of the civic duties of individuals.

autarchy in charge of the exam, we believed that there was a chance that microdata from the exam would not be available for research purposes. Second, we thought it was important to have at least some control over the theme of one of the essays, to guarantee that (by chance) students in the treatment group would not have better scores simply by writing about a topic that they had just trained in one of the program-related writing practices. This turned out to be important, as we discuss while presenting our main results. Third, we anticipated that we would be able to include better individual-level controls in the related regressions because, for these data, we can match students' outcomes with more controls that are highly predictive of achievement (such as the Portuguese Language multiple-choice scores in the state's standardized exams). Finally, we anticipated that participation in the *ed* techs could lead some students to enroll in ENEM, generating differential attrition. As discussed in the pre-analysis plan, following Ferman and Ponczek (2017), we pre-specified the following strategy in case we found significant differential attrition for at least one dataset: if considering both datasets led to larger confidence sets while using bounds to account for differential attrition, we would focus on the results from the data with less attrition problems, and present the results with the other data in the appendix.

4.2.2. **Mechanisms**. In order to provide a rich understanding of the channels of impact of both *ed* techs and try to explain potential differences in treatment effects, we collected primary data on students and teachers at the end of 2019. We partnered with state's educational authority and included multiple-choice questions in the state's standardized exam ("Programa de Avaliação da Educação Básica do Espírito Santo", PAEBES) questionnaire, which happened three weeks before ENEM and independently collected data through phone surveys with teachers in November and December (after ENEM).

Training and Feedback. For students, the variables collected on training, individual feedback and interactions with their teachers were: (**s**.i) the number of essays written to train for the ENEM in 2019; (**s**.ii) the number of ENEM essays that received individualized comments and/or annotations; (**s**.iii) their perception on the usefulness of the comments and/or annotations —not useful at all, somewhat useful, very useful; (**s**.iv) the number of essays that received a grade; (**s**.v) the number of essays graded that were followed by a personal discussion with the teacher. All student variables were top-coded at 10. The variables on essay assignment and collective feedback behavior of teachers during 2019 were: (**t**.i) number of essays assigned to train for the ENEM; (**t**.ii) number of essays assigned inside the classroom; (**t**.iii) number of essays graded; (**t**.iv) number of essays assigned that were followed by a discussion about common mistakes; (**t**.v) number of essays assigned that were followed by a discussion about good writing patterns. Most of the questions of the teacher survey were open-ended so they tend to provide very large and implausible values for some individuals. As specified in the pre-analysis plan, we winsorize these data at the top 1% and, for our main results, we also investigate whether this choice of winsorizing parameter is relevant.

Teachers' Time Allocation. We asked teachers about their perceptions on the the time available during the year to improve students' knowledge on each subject of the high school senior year curricula using a Likert Scale, where 1 meant "Time is very insufficient" and 5 meant "Time is more than sufficient".

We also asked the number of hours in a typical week allocated to: (i) marking essays, (ii) correcting classwork and homework related to Grammar or Literature, (iii) preparing lectures and materials for

activities, including home assignments, (iv) giving individual support to students (one-on-one tutoring), guiding and counseling those with academic problems or special interests, and (v) extra-hours of work (inside or outside schools).[19]

Teachers' Expectations and Students' Aspirations. We pre-specified the analysis of the following outcomes: (**t**.i) teachers' perceptions about the proportion of their students that will succeed in the ENEM test and be admitted in a college (either public or private) in 2020; (**t**.ii) teachers' expectation on grades of students on the ENEM 2019 essay; (**s**.i) students' plans for 2020 (work, college, or both), which we will use to understand whether the programs shift students' aspirations towards attaining post-secondary education.

Knowledge About Students. We measure teachers' knowledge using the following variables: (**t**.i) teachers' perceptions on how much they know about the strengths and weaknesses of their students in writing essays, and on Grammar and Literature, in a scale of 1 to 10; (**t**.ii) difference between the actual average grade in the exam's essay and the teachers' predicted average grade of their students in public schools in the written essay of ENEM 2019.

4.2.3. **Secondary Outcomes**. To understand whether effects on test-specific writing skills would spill-over to different textual genres, we asked students to write a narrative essay in the same day we collected the ENEM training essays in schools. The essay proposal and the grading criteria were also developed and corrected independently by CAEd/UFJF. The proposal presented the student with three motivating elements. The first one was a definition of biography as a textual genre. The second and third ones were excerpts from biographies. At the end of the page, students were instructed to write a narrative telling the reader about a special moment in her life-story.

To study learning in other subjects during the year, we use administrative data on all other 2019 multiple-choice ENEM test scores and on the multiple choice state's standardized exams. We combine information from the ENEM Mathematics, Natural Sciences, Human Sciences tests, and the PAEBES Mathematics, Physics and Chemistry standardized exams to test our main hypothesis on indirect effects on the accumulation of skills in subjects non-related to Language. We proceed similarly for the subjects that are related to Language, using the ENEM Language and Codes test and the PAEBES Language (reading) exam administered by SEDU/ES.

4.3 Econometric Framework

4.3.1. **Identification and Estimation**. Given the randomized nature of the assignment mechanism, the causal impact of being offered a chance to use the *ed* techs can be studied by comparing outcomes in schools selected for the treatment conditions and outcomes in schools selected to form the control group. Since we have data on two different exams for our primary outcome, we append the two scores to maximize the power of our experiment and estimate the intention-to-treat (ITT) effects using the following regression:

$$Y_{ise} = \tau_{\texttt{ITT}}^{\text{Enhanced AWE}} W_s^{\text{Enhanced AWE}} + \tau_{\texttt{ITT}}^{\text{Pure AWE}} W_s^{\text{Pure AWE}} + \mathbf{X}'_{ise}\mathbf{\Pi} + \epsilon_{ise} \quad (1)$$

[19]The division of tasks into the 4 first groups was built from the O*NET list of typical tasks of high school teachers, and maps well into the Teaching and Learning International Survey (TALIS) of OCDE countries.

where Y_{ise} is the essay score of student i, in school s, for exam e, which can be the score in the 2019 ENEM essay or in the argumentative essay that was included in the state's standardized test and $W_s^{\text{Enhanced AWE}}$ ($W_s^{\text{Pure AWE}}$) is an indicator that takes value 1 if school s was randomly assigned to the version of the *ed* tech with(out) human graders. In equation (1), the vector $\mathbf{X}_{ise}$ contains strata fixed effects, an indicator variable of the exam, and the school- and individual-level covariates specified in the pre-analysis plan.[20] The differential ITT effect between the two *ed* techs is estimated using the regression:

$$Y_{ise} = \tau^{\Delta} W_s^{\text{Enhanced AWE}} + \tilde{\mathbf{X}}'_{ise}\mathbf{\Gamma} + \nu_{ise}, \tag{2}$$

where we include only students from the two treatment arms. In this regression, the vector of covariates $\tilde{\mathbf{X}}'_{ise}$ includes the artificial intelligence score from the first essay of the program, in addition to all covariates in $\mathbf{X}_{ise}$.[21] Since both treatment arms are indistinguishable prior to the feedback students received from this first essay, this variable can be used as a covariate. Of course, this cannot be used in the regression model (1), because this information is not available for the control students. The idea of estimating the differential effect from regression (2) instead of using regression (1) is that we expect this variable to be highly correlated with the follow-up essay scores, which will potentially improve the power in this comparison. In the case of other individual and teacher-level regressions, we estimate:

$$Y_{is} = \tau_{\text{ITT}}^{\text{Enhanced AWE}} W_s^{\text{Enhanced AWE}} + \tau_{\text{ITT}}^{\text{Pure AWE}} W_s^{\text{Pure AWE}} + \mathbf{X}'_{is}\mathbf{\Lambda} + \nu_{is}, \tag{3}$$

where Y_{is} is an outcome of interest — for instance, the number of ENEM training essays written or assigned — for student or teacher i and the other variables have been defined above. In the teacher regressions we only add our school-level covariates. In the regressions using student data we add the same controls added in specification (1).

4.3.2. **Inference**. Inference is based on the inspection of three different sets of p-values. First, we present p-values based on standard errors clustered at the strata level. As reported by de Chaisemartin and Ramirez-Cuellar (2019), standard errors clustered at the school level would be downward biased in this setting. This is confirmed by the inference assessment proposed by Ferman (2019), which shows that clustering at the school level would lead to over-rejection, while clustering at the strata level is reliable. Note that this way we take into account in our specification for the primary outcome that we may have information on more than one essay for each student. Second, we present randomization inference p-values using the randomization protocol and 1,000 placebos that maintain the stratified

[20] The school-level covariates, which we can merge with data from both exams, are the 2018 average ENEM essay score in the full score or in each skill. We add as school-level covariates the school average ENEM score in 2018 or, for each skill group subject, the school average in the group or subject. The individual-level covariates are: (i) female indicator; (ii) age dummies ranging from 17 or less to 23 or more; (iii) educational and occupational characteristics of the mother and father of the students; (iv) household income category; (v) baseline Language and Mathematics proficiency scores using data from another state's standardized exam that happened right before the treatments were implemented. These covariates are interacted with the exam indicator to take into account that the set of covariates available for observations from the 2019 ENEM are different from the other exam (we cannot identify students in the ENEM essay in order to observe baseline achievement for these students). We also replace missing school-level and individual-level continuous covariate values with the control group mean and included an indicator for missing in this covariate in the regression. For discrete covariates we created a complementary category for missing variables.

[21] We are not be able to match students on the ENEM 2019 microdata. Therefore, this variable will only be included as covariate for the other essay score. We will interact this variable with an indicator variable for the ENEM essay.

structure of the original assignment mechanism. The inference tests use the coefficient estimate as the randomization test statistic.[22] Third, we present p-values adjusted for multiple hypothesis testing (MHT) based on the procedure proposed by Holm (1979).[23] There are two possible margins of adjustment: multiple treatments and multiple outcomes. Thus, for instance, when we consider the main effects of the treatments on the three ENEM groups of skills, we will correct for the fact that we are testing six hypotheses (three outcomes and two treatments). To simplify the interpretation of the main findings and maximize the power of our tests on mechanisms we also condense variables within a family a mechanisms discussed in Section 3 following the hypothesis testing procedure of Anderson (2008), unless otherwise specified.

4.4 Design Validity

4.4.1. **Balance**. In Table 1, we consider if the assignment mechanism generated balanced groups across treatment arms. Columns 1, 2 and 3 present estimates of each treatment indicator and their difference from ordinary least squares regression with strata indicators. Standard errors clustered at the strata level are in parentheses and p-values from inference tests are in columns 4, 5 and 6.

Panel A uses standardized individual level covariates from ENEM 2018. Overall, we find that the experimental sample of schools is balanced according to this set of observables. If anything, treated schools' students fared slightly worse in the ENEM 2018 written essay and other exams, but such differences tend to be small in size and are never statistically significant.

Panel B uses student-level information from a standardized exam that was implemented by the State's Education Department in all public schools in Espírito Santo in April 16th 2019. Treated schools were informed by the State's Education Department about the additional inputs from the pure and the enhanced AWE on April 11th, and teachers' training started only after the standardized exam (end of April). Therefore, it is safe to assume that treatment assignment did not meaningfully affect the results in this exam. These data provide valuable information because it is based on the students that actually participated in the experiment, as opposed to the variables discussed above. Consistent with the results shown in Panel A, we find that students in the treatment arms had slightly worse test scores in Portuguese Language and Mathematics at baseline, but once again these differences are not statistically significant. Also consistent with the randomized assignment mechanism, the joint p-values (Young, 2018) in the bottom rows of Table 1 are greater than 0.701 for all comparisons.

The comparison between experiment arms for a wide range of covariates thus provides compelling evidence that the randomization generated statistically comparable groups of students at baseline. Notice, however, that Table 1 does not contain all the variables we use as covariates in specifications

[22] Specifically, we present, for each relevant estimate, the proportion of placebo estimates that are larger (in absolute value, in the case of one-sided test) than the "actual" estimate. This procedure has the advantage of providing inference with correct size regardless of sample size and are particularly important for the sample of teachers, for which we can't rely on a large-sample for inference purposes. To test the hypothesis of no treatment effect in each arm, we use two separate sets of permutations. For instance, to test whether the standard program had no effect, we keep the assignment of schools in the pure AWE treatment arm and generate 1,000 alternative draws under the original randomization protocol for units in the control and the enhanced treatment, proceeding analogously when testing whether the pure AWE program had no effect.

[23] The Holm (1979) MHT adjustment works by ordering the p-values from smallest to largest, with their corresponding null hypotheses. Then the smallest p-value is multiplied by 6, the second one is multiplied by 5 and so on. Formally, we set the Holm adjusted p-values $\hat{p}_i^h = \min(k_i \hat{p}_i, 1)$, where k_i is the number of p-values at least as large as $\hat{p}_i$ within a family of hypothesis.

(1) and (2). The other covariates were collected at the student's questionnaire (for example, age, parents' education, and household wealth), so we do not have information for all students at baseline. We consider balance with respect to these covariates in the next paragraphs by conditioning our samples to non-attritors.

4.4.2. **Attrition**. The first rows in Table 2 presents estimates and inference tests for attrition in our main analytical samples. Column 1 presents attrition rates in the control group. Columns 2, 3 and 4 present estimates from specification (3), an ordinary least squares regression with indicators for each of the two experiment arms and strata indicators. In columns 5 to 7, we add to this regression the school-level and individual-level controls available in the beginning of the year that we use in our main regressions.

For the analysis of student-level data, we start with the baseline list of 19,516 students in experimental schools using the same data on the April standardized exam we used for balance. For the nonofficial ENEM essay we administered, we find an attrition rate of 22% for students in the control schools, with no statistically significant differences among students in the treated schools. We reach the same conclusion by considering attrition in the students' questionnaire we used to collect information on the mechanisms, where the proportion of attriters was 17%.

For the official ENEM essay, we do not have identified information at the student level. For this reason, we are only able to identify the students' school and whether the student was a high school senior in 2019. Thus, for each school, we contrast the number of students with information on the ENEM essay with the number of students enrolled in April 2019 to investigate attrition problems. In these data, we also find that the share of students that are present in the ENEM essay is not significantly different across the experimental groups.

In Appendix Tables A.2 and A.3 we also consider balance on all our covariates conditional on being a non-attriter, respectively for the nonofficial and the official ENEM essays. We find no evidence that the experimental groups are different even when we condition on being observed. Considering the three treatment arms and the two datasets, we have six pairwise comparisons, with joint p-values of equality (Young, 2018) for all covariates ranging from 0.161 to 0.910. This provides further evidence that student-level attrition is not a problem in our analysis.

The fourth row in Table 2 describes attrition in the teacher-level data. We collected information on 84.6% (274) of the 324 teachers assigned to teach high school senior classes in schools in the experimental sample as of April 2019. The estimates of attrition indicate that teachers working in schools that adopted the enhanced AWE system were more likely to attrite (p-value=0.080). This conclusion holds whether we control only for strata fixed effects (column 2) or when we also add the ENEM 2018 average essay score in the school (column 5). We discuss robustness tests on our teacher results while presenting the results.

4.4.3. **Mobility**. A potential threat to the validity of our experiment would be students switching to different schools because of the treatment. This could happen if, for instance, more motivated students moved to treated schools to get access to the *ed* techs. In the nonofficial ENEM essay, we are able to identify individual students. Therefore, this does not pose a significant problem, as we would able to consider the initial allocation as an instrument for treatment status. However, for the official ENEM essay, students' mobility could be a more serious problem, as we are only able to identify students'

schools and whether they were graduating that year.

We expected such movements to be extremely unlikely because the randomization and disclosure of the treated schools were made in the middle of April 2019, a couple of months after the school year began. Nevertheless, we use administrative data from SEDU/ES initial allocation and transfers to check if this is a relevant concern. we contrast the enrollment list of students in the PAEBES exam, which took place in October 2019, with the same data on the April standardized exam we used to assess balance and attrition. The results are shown in the last row of Table 2. We find that, in control schools, only 1.2% of the students enrolled at the end of the year were not originally enrolled in the same schools in April 2019. Again, these proportions are not significantly different for students in the treated schools.

Overall, the absence of patterns in student mobility related to treatment assignment, combined with the evidence above that there is no differential attrition, provides evidence that the set of students at the end of the year in experimental schools is representative of the set of the students in those schools at the time of the randomization. Moreover, the results we present in Section 5 show that the treatments significantly affected essay scores, but did not have significant effects in other exams. This provides evidence that students' mobility is not an issue when we consider the data from the official ENEM essay. For the nonofficial ENEM essay, in which we can identify individual students, we consider the initial school enrollment to define the treatment variables in equations (1) and (2).[24]

5 Main Results

5.1 Implementation and Compliance

We start by describing the timing of the experiment and the compliance behavior of teachers and students in treated schools using engagement data available from the implementer.

5.1.1. **Teachers**. Teachers were not aware of being part of a randomized trial with two treatment arms. In spite of the meaningful differences between *ed* techs, they complied very similarly with the experiment.[25] Figure 2 shows that more than 95% of teachers used the *ed* techs to assign and collect essays in each of the five writing activities. This is somewhat surprising, given that the use of the technologies was enthusiastically supported but not set as mandatory by the educational authority of the state. We observe little or no variation between writing activities and across treatment arms and,

[24]Since mobility is very low, however, results are virtually the same if we consider the end-of-the-year allocation of students for both exams.

[25]The implementation started in mid April 2019 with itinerant presentations of both *ed* techs across the educational administrative units of Espírito Santo. The academic year starts in February, but the state educational authority postponed this step of the intervention until all the laptops were distributed to schools in the experimental sample. The presentations were scheduled and enforced by the State's Education Department through direct and *online* communication with the treated schools' principals. In each administrative unit, the implementers' staff divided schools according to randomization status to two different rooms, one for each *ed*-tech. These presentations consisted of 2-hour informal lectures on how to use the platform and on which type of individual and aggregate information it would store during the year. In order to standardize the presentations and minimize the likelihood of suggesting that there would be two different AWE-based systems being used across schools, the presenter were either only in charge of presenting the enhanced or the pure AWE treatment. These presentations were attended by 257 individuals representing 101 schools (92%). These individuals were not all teachers (one third were). Consistent with the randomization and blinded nature of the experiment, there is no difference in the probability that a teacher was sent as a representative by treatment arm (p-value = 0.469). To boost engagement and circumvent implementation problems, teachers that were not present in the training sessions were also invited to online training sessions.

in fact, we cannot reject the null hypothesis of no difference in the evolution of compliance throughout 2019 (p-value = 0.245).[26] In the pure AWE arm, in particular, the high compliance is inconsistent with teachers avoiding to use a system that they don't perfectly understand —or, even worse, fear.[27]

5.1.2. **Students**. We also observed relatively high and similar levels of student compliance. At each writing activity, 75 to 80% of students enrolled in treated schools submitted essays through the platform. Again, we cannot reject the null hypothesis that compliance throughout the year was equal in both treatment arms (p-value = 0.464).

As discussed in Section 2, the feedback enhancement by human graders took, on average, three business days. To investigate whether such lag had meaningful effects on compliance in the enhanced AWE *ed* tech arm, the trend in the bottom of Figure 2 depicts the share of students submitting essays that entered the platform to check the enhanced grading. The share starts at 70%, falls slightly in the following three activities and then in the last one, when one in every two students who submitted essays came back for their grading.[28] While these figures corroborate the importance of receiving immediate feedback (as highlighted by Muralidharan et al., 2019), they also allow us to consider that differences in effects should not simply be a result of students not fully complying with the enhanced grading. As we will show, we also find that students perceived a higher quality on the feedback in the enhanced AWE *ed* tech, which also suggests that compliance was large enough to generate meaningful differences in treatment arms.

5.2 Primary Outcome: ENEM Essay Scores

Table 3 presents the main results of the experiment, which are also depicted graphically in Figure 3. In Table 3, column 1 documents that the enhanced and the pure AWE *ed* techs had almost identical effects on ENEM essay scores, at 0.095σ. Columns 2 to 4 show that the total effects are channeled by very similar positive effects in scores that measure each group of writing skills valued in the essay. The results we find in the pooled data are very similar to the ones we obtain by considering each one of the essay scores separately (Appendix Figure A.1). Since the topics of the last writing activity and the 2019 ENEM essay were about the social role of art, the fact that we find similar results considering only the nonofficial ENEM essay minimizes concerns on the external validity of the results found in the pooled data.

For both *ed* techs and outcomes, we are able reject the null hypothesis of no treatment effects (Panel A) and unable to reject the null hypothesis of no differential effects (Panel B). Thus, the additional inputs from human graders did not affect the extent to which the *ed* techs were able to improve scores capturing a broad set of writing skills.

Column 2 presents effects on the first group of skills valued in the ENEM essay, which are related to

[26] We test this hypothesis by running a regression of the teachers' indicator of compliance at the extensive margin — measured by an indicator of assigning and having students submit essays through the platform — on treatment arm indicators, writing activity indicators and their interactions and testing that the interaction terms are jointly significant.

[27] Additionally, the fact that compliance was sustained rules out the possibility that teachers learnt and became disappointed with the quality of the feedback from both *ed* techs. Since differences in mechanisms on the teacher-side will not be driven by large differences in compliance, we can interpret the estimated ITT effects as good approximations of the average effect on the treated parameter. These measures of compliance also contribute to the educational research on the topic, which so far dealt only with subjective measures of social acceptance of AWE systems (see Wilson and Roscoe, 2019, for instance).

[28] We can reject the null hypothesis that compliance was stable throughout the year (p-value < 0.001).

syntax. In ENEM, scores capturing syntactic skills measure both the ability of students to correctly use the formal norm of written language and their ability to build a sound linguistic structure connecting the various parts of the essay. Panel A documents that the enhanced AWE *ed* tech increased scores in syntactic skills in 0.066σ and that the pure AWE *ed* tech increased scores in 0.056σ. In Panel B, we show that these absolute effects do not translate into significant differential effects.

Notice that syntactic skills are the ones in which both *ed* techs fare similarly in capturing and fostering, since both are able to instantaneously flag deviations from the formal written norm and identify whether the essays have well-built linguistic structures. Thus, it is perhaps not surprising that the additional inputs from human graders did not matter much for scores in syntactic skills. As we now show, we also found indistinguishable effects on scores measuring writing skills that one might consider as skills in which the AI system would fall short in capturing.

The second group of writing skills, which we refer to as analytic, are related to the ability of students to develop a globally coherent thesis on the topic. The development of this thesis in a successful essay allows students to mobilize elements from different areas of knowledge (for instance, history, philosophy and arts). High scores in analytical skills thus benefit not only students that "write well" but also students that have a more solid educational background and can leverage potential synergies with other topics covered in the post-primary education curriculum. In fact, at least part of this is not even supposed to be built in schools, as a perfect score is only given to students that showcase a "vast socio-cultural background". Despite the intuitive leverage that human participation would entail in helping students to develop such a complex set of skills, we find very similar effects of both *ed* techs. In Panel A, column 3, we show that the enhanced AWE *ed* tech increased scores in syntactic skills in 0.042σ and that the pure AWE *ed* tech increased scores in 0.061σ (the first estimate is only marginally significant, MHT adjusted p-value $= 0.152$). Once again, Panel B documents that these effects do not translate into significant differential effects.

Most surprisingly, we document particularly large and, once again, very similar, effects on the policy proposal skill in column 4. The point estimates are 0.161σ and 0.143σ for enhanced and the pure AWE *ed* tech, respectively. In Section 2, we argued that the policy proposal skill is the most sophisticated skill in the exam. The reasons are twofold. First, the ability to present a consistent policy contribution makes this a "global" property of the essay. That is, only in reference to the thesis presented in the starting paragraphs will a policy conclusion be interpreted and graded. Second, policy proposals are the means by which students showcase creativity and problem-solving skills.

Given the complexity and semantic nuances in providing grades and feedback on analytical and policy proposal skills, the absence of differences between positive effects of both *ed* techs suggests that teachers ended up filling in some of the gaps or limitations of the pure AWE *ed* tech. We present more evidence in this direction in Section 5.3.

Even though the effects we find could be considered small when compared to the ones reported in the program evaluation literature, there are relevant facts one should consider when interpreting magnitudes in this setting. First, language's most sensitive period of development happens before adolescence and tends to be less responsive to variation in educational inputs (see Knudsen et al., 2006). To the extent that writing ability is a higher-order dimension of linguistic development, this feature should also apply to writing among 17 to 19 year-old students, and perhaps even more so. Second, students in our setting are much more heterogeneous in terms of their writing scores than in

terms of their scores in multiple-choice tests.[29] Thus, it is questionable whether benchmarking sizes using policy effects on the distribution of multiple-choice test scores would be a meaningful exercise.

With that in mind, and since public school students compete with private school students in college admission, we benchmark magnitudes using the national argumentative essay public-private achievement gap and find that the *ed* techs mitigate 9% of the gap. In the policy proposal, the effects we find imply a reduction of 20% in the skill-specific gap, which is currently at a high 80%. Overall, we consider that these are economically meaningful effects that bear policy relevance, specially in a setting with a schooling system that is sharply segmented and reflects various differences in socio-economic backgrounds.

5.3 Mechanisms

5.3.1. **Training and Feedback**. We start our discussion on mechanisms by considering how individual training and feedback in its many potential forms —comments on essays, grades and individual discussions with teachers— responded to the incorporation of the *ed* techs. The results are in Table 4.

On average, students in the control group wrote 4.9 ENEM training essays during the year of the experiment. In column 1, we report that the enhanced AWE *ed* tech increased this number by 1.4 essays or 28%. Students using the pure AWE *ed* tech wrote, in turn, 1.6 more essays than students in the control group, or 32%. Both estimates are not only large in magnitude, but also precisely estimated. Using the lower bounds on confidence intervals allow us to reject effects of less than 20% in both treatment arms. In Panel B, we show that the difference between the effects of the two programs is insignificant. Hence, both *ed* techs induced similar increases in ENEM-oriented training. These results are informative about how instruction reacts to the new inputs provided by the *ed* techs, suggesting that they did not crowd out one by one the essays teachers would assign themselves. In addition, the results are inconsistent with the *ed* techs inducing writing tasks to take much more time, either because of problems with the Internet, lack of computers, or difficulties with the use of the technologies.

During the year, students may have received comments or annotations on essays written, either handwritten by teachers or sent through the platform in the treatment conditions. In the students' questionnaire, we asked students how many of the ENEM training essays they wrote were commented or annotated. The results are shown in column 2 and indicate, once again, similar and highly significant positive effects of 37-38% (1.3 essays on an average of 3.4 essays) in both treatment arms. In column 5, we document that the number of essays that were graded increased —and not differentially so— with the *ed* techs. Students on both arms observe grades on the essays submitted to the teachers or to the online platform in 1.6-1.7 more essays, a highly significant increase of 43-45%.

Regarding the quality of the feedback, 81% of the students in control schools claimed that they found the comments and annotations on essays *somewhat or very* useful to improve ENEM writing skills. The participation in either of the treatment arms raises this probability to approximately 87-88% (10%), as shown in column 3, Panel A. In Panel B, we confirm statistically that both *ed* techs had similar impacts on this margin. Column 4 presents the same estimates using a more stringent criteria on feedback quality. Confirming the intuition that external human graders improve the outputs

[29] Taking ENEM 2018 as an example, the dispersion of this distribution is almost two times the one of the Mathematics scores and almost three times the one of the Language scores. At a very general level, this suggests that there is some degree of heterogeneity being captured by scores on essays that is is not captured in multiple-choice exams.

generated by the pure AWE system, we find that students using the enhanced AWE *ed* tech were 6 percentage points or 14% more likely to claim that the comments or annotations on their essays were *very useful*. The results in the pure AWE, in turn, are negligible in size (1 p.p. or 2%) and statistically insignificant. As shown in Panel B, the results we find on the difference between *ed* techs are individually significant and marginally significant after the MHT adjustments (p-value=0.120, for the latter). All in all, these results indicate that the incorporation of the *ed* techs into instruction improved feedback quality and alleviated at least some human capital constraints of teachers with respect to the counterfactual feedback. Moreover, the differential effects in perceived quality show that the lack of difference in treatment effects on essay scores does not come from students failing to fully comply with the enhanced AWE *ed* tech by not checking the final feedback.

Finally, we asked students about the number of graded essays that ended up being discussed with teachers. This question was added to the questionnaire to understand if integrating both technologies could support the individualization of writing pedagogy. Following Autor et al. (2003), we included this question to capture a potential complementarity be between the system's parsing and grading tasks and teacher's nonroutine analytical (interpretation) and interactive tasks (providing in-person individualized advice). If teachers completely delegated essays' training tasks to the *ed* techs, then we should expect no effect, or even a negative effect on this variable. In contrast, we should expect an increase if teachers complemented the *ed* techs.

The results are presented in column 6. Students in both treatment arms discussed roughy 35% more essays individually with teachers. Once again, marginal results are highly robust to MHT adjustments, and we do not find evidence of differential effects. This result helps reconcile the lack of differential effects between the two treatment arms on skills that pure AWE systems would have more difficulty in assessing and fostering. Notice, additionally, that these results are inconsistent with the *ed* techs leading to a complete delegation of tasks to the new inputs.[30]

5.3.2. **Teachers' Time Allocation**. We asked teachers to describe the time available to cover the topics in each subject of the high school senior curriculum in 2019. The possible answers for all subjects they typically cover (writing, Grammar and Literature) were on a 5-point Likert scale and ranged from "Time very insufficient" to "Time more than sufficient". As shown in Panel A of Table 5, columns 1 to 3, the enhanced AWE *ed* tech improved these indicators by roughly 0.3 or 12% across teaching subjects. Column 4 documents that these changes translated into a significant improvement of 0.26σ in a summary index (Anderson, 2008). In turn, the pure AWE *ed* tech had a negligible impact on the summary index and on each of its components. Appendix Figure A.2 further investigates this margin of change by plotting distributions of answers for writing. The bar graphs show that 23% of teachers in control schools said that they felt that the time was very insufficient. This proportion drops to 9% for teachers using the enhanced AWE (p-value=0.008, result not shown), but is roughly unchanged for teachers using the pure AWE *ed* tech. Taken together, Table 5 and Appendix Figure A.2 present suggestive evidence that the enhanced AWE —but not the pure AWE— was able to alleviate time

[30] We also collected data on essays' assignment in the teachers' survey. The results are in Appendix Table A.4. Overall, we do not find evidence that these variables were affected by the introduction of the *ed* techs. Our interpretation (specially, giving what we found using student-level data) is that this information ended up being much less informative than the results we found using student-level data. First, the teacher survey asked about assignment, which may not have been complied by students. Moreover, even after winsorizing answers as specified in the pre-analysis plan, we were left with observations that led us to think that the numbers of essays were implausibly large. Finally, we also had differential attrition in the teachers' survey.

constraints, at least for some teachers. The fact that teachers that used the enhanced AWE *ed* tech felt less time constrained is consistent with evidence shown in Table 6. Panel A, column 1, shows that teachers in this arm worked 1.2 hours less from home in a typical week. This effect amounts to -20% of the control group mean (5.9 hours), is marginally significant (p-value $= 0.112$). At face value, these results suggest that teachers adjusted labor supply downwards in response to the incorporation of the enhanced AWE *ed* tech. However, we consider these results with caution, given that, as described in Section 4, we found differential attrition in the teachers' survey. In Appendix Table A.5, we compute lower and upper bounds associated with the estimates discussed above (Lee, 2009).

Turning to columns 2 to 5 in Table 6, we discuss the effects of the *ed* techs on hours allocated to each group of tasks. Overall, we find little support for changes along this margin. The estimates tend to be imprecisely estimated, being insignificant but also including meaningful values of treatment effects. To provide a more direct test in time allocation changes and maximize the power of our comparisons we also present p-values from chi-squared tests. As the individual estimates on tasks, this exercise provides little support for changes in time allocation in both treatment arms. Finally, in column 6, we sum the hours allocated to tasks in columns 4 and 5 to compute the share of time spent in nonroutine tasks. The results suggest little role for adjustments in time outside the classroom playing an important role in our results.

5.3.3. **Other Mechanisms**. In Table 7, we investigate whether the *ed* techs shifted teachers' expectations and students' aspirations toward academic tracks after high school. Overall, we find little support for these mechanisms playing a large role in boosting teachers' instructional efforts and/or students' training and ultimately shaping the effects we find on ENEM essay scores. Additionally, in columns 1-3 of Table 8, we document that the *ed* techs did not affect teachers' perceptions on knowing their students' strengths and weaknesses. Using the difference between a teachers' average student guessed score and the actual score as outcomes in column 4, we do not find strong evidence that the information from the *ed* techs' (individual grades and feedback or "average" indicators of performance) made teachers more accurate about their students' future ENEM achievement. Thus, the results on all outcomes suggest little role for changes in perceived or objective knowledge about students playing an important role in our results.

5.4 Secondary Outcomes: Learning in Other Topics

We now consider whether the shifts in educational inputs oriented at the ENEM essay described above had indirect effects on scores capturing other skills, either related or unrelated to writing and literacy. The results are in Table 9 and provide strong evidence that the effects of the *ed* techs were restricted to their main goal of improving ENEM essay scores. In particular, they are inconsistent with meaningful complementarities and crowd-out in effort from students or in time dedicated by teachers to other topics.

Column 1 shows that there is little support in the data for changes in the writing skills used to write essays following another textual genre (narrative). The estimates are not only insignificant, but also negligible in size and, again, there is no evidence of differential effects. These are meaningful results from a general perspective on human capital formation, since one might consider that more general writing skills are valuable in future tasks students face in post-secondary education and in the labor market. The results are also relevant from a direct policy perspective on college admission, since other

exams use scores on the narrative genre as criteria. In column 2, we investigate whether the changes we described in students' training and Language teachers' behavior ended up having downstream effects on topics related to reading and literacy. Again, we find little evidence of any absolute or differential effects. In column 3, we reach similar conclusions by pooling data on multiple-choice tests capturing skills that are not related to literacy. Since we pool several sources of data, we are able to reject even small negative spillover adverse effects in each of these families of outcomes, suggesting that the effects of the *ed* techs were restricted to their main goal of improving ENEM essay scores.

5.5 Heterogeneity

In Appendix Table A.6 we investigate whether treatment effects are heterogeneous with respect to students' and teachers' characteristics, using students' socio-demographic variables (gender, race and household income), full and partial school shift, and baseline Language achievement. There is little support for the hypothesis that sub-samples of girls, blacks or *pardos*, poor (below median household income) and full-shift students were differentially affected by the *ed* techs. We do find some evidence that effects on students in the bottom quartile of the baseline Language distribution are smaller and, for the pure AWE arm, that effects are channeled by students in classes where teachers have above median workload, as measured by the number of classes taught. However, when one adjusts the marginal p-values for the number of heterogeneity margins, these results become insignificant. Overall, we conclude that there is no evidence of stark effect heterogeneity across sub-samples of the data.

6 Final Remarks

Elbow (1981) provides an insightful description of a male Language teacher doing extra-hours:

> "*He sits at his desk reading student papers. He is half done with a batch, the unread stack neatly piled to his left, each paper tightly folded long-wise; the graded pile a bit helter-skelter to his right. It is late and he stops for another cup of tea, annoyed he didn't start earlier in the evening. If he is a conscientious teacher he assigns a paper every week to every student he has. But he also kicks himself as he sits there sipping tea because he is acutely aware of how it is* he *who brought this job down on his own head.* [...] *If he isn't so conscientious he assigns writing every few weeks but he feels guilty because he knows this doesn't give his students enough practice and it means that his comment and advice on a student's paper this time will probably have no useful effect at all on what the student writes next time.*" (p. 255)

The excerpt highlights important features of the work of Language teachers, which are supported by anecdotal evidence we found in the field. Teachers know that essay assignments and practice should be frequent. Nevertheless, marking essays and providing careful feedback takes time, specially if the most rudimentary writing skills —such as syntax— are not well-developed by the writer. Thus, even for highly-motivated or "conscientious" teachers, time constraints will predictably bind. In developing countries, the downstream effects of these time constraints will potentially be reinforced by human capital constraints (for a discussion, see Banerjee et al., 2013). This paper provides evidence that AWE systems can help overcome bottlenecks that prevent accumulation of writing skills.

We show that, despite large differences in structure and costs, the two *ed* techs we study positively impacted each one of the dimensions of writing valued in the argumentative essay of a nation-wide college admission exam. The most robust evidence on mechanisms indicates that these impacts were channeled by large increases in training, improvements in feedback and more frequent personal interactions with teachers. The latter mechanism suggests that AI —be it "supervised" or "unsupervised"— will not simply substitute what are arguably teachers' most valuable inputs. The new wave of *ed* tech may instead open space for more individualized pedagogy and a division of labor that dynamically takes into account what are the tasks in which humans and machines hold comparative advantage.

More generally, the absence of differential effects between the two *ed* techs is informative about the potentials and limitations of applications of artificial intelligence. Putting ourselves in the shoes of the implementer and interpreting their actions in the light of the citation above is useful to frame this discussion. The implementer saw in one of the first Brazilian Portuguese AWE systems the potential of quickly providing *some* feedback to students and allowing teachers to outsource some of the "heavy lifting" of essay parsing and grading to an automate. The grading would then be enhanced by human graders, at the cost of a lag between students attempting a problem and receiving the complete feedback. In this sense, human graders were hired to circumvent the AWE systems most salient limitations: its lack of ability to contemplate semantic nuances (interpret) and accurately adjust communication to be useful to highly heterogeneous students (interact). Essentially, this enhancement tried to push the AWE system towards being more like a human in order to deliver high-quality, individually customized and interactive content. In the field, we found that *ed*-techs generated very similar compliance, induced the same increases in student effort and in the amount of feedback. Most importantly, we found that teachers did not simply delegate their tasks and that both *ed* techs highly supported the individualization of pedagogy —which is essentially a nonroutine interpretative and interactive task. We hope that this case-study on how AI was incorporated by teachers and students end up having consequences for the policy discussion on AWE in developing countries as these systems are developed to encompass more languages. More broadly, our results may help alleviate the "technological anxiety" (Mokyr et al., 2015) that seems to surround the debate in policy circles.

Our results also inform the debate on whether pure AWE systems should be abandoned because they take linguistic complexity for complexity of thought by simply "*counting words*" (Perelman, 2014). We show that the discussion on whether whether AWE systems are able to take all the complexity of writing into account largely bypasses the fact that these inputs interact with other inputs, such as teachers' instructional efforts for a given level of human capital. In light of that, and considering our experimental design, it is not obvious whether the introduction of AWE systems without being incorporated as classroom instruction, with the support of school teachers, would generate the same positive results. We see that as an interesting avenue for future research. Still, our results show promising prospects of using AWE systems to improve learning.

References

Abaurre, M. L. M. and Abaurre, M. B. M. (2012). Um olhar objetivo para a produção escrita: analisar, avaliar, comentar.

Acemoglu, D. and Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of labor economics*, volume 4, pages 1043–1171. Elsevier.

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American statistical Association*, 103(484):1481–1495.

Autor, D. H., Levy, F., and Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly journal of economics*, 118(4):1279–1333.

Banerjee, A., Glewwe, P., Powers, S., and Wasserman, M. (2013). Expanding access and increasing student learning in post-primary education in developing countries: A review of the evidence. *Cambridge, MA: Massashusetts Institute of Technology*.

Barkaoui, K. and Knouzi, I. (2012). Combining score and text analyses to examine task equivalence in writing assessments. In *Measuring Writing: Recent Insights into Theory, Methodology and Practice*, pages 83–115. Brill.

Bettinger, E., Fairlie, R. W., Kapuza, A., Kardanova, E., Loyalka, P., and Zakharov, A. (2020). Does edtech substitute for traditional learning? experimental estimates of the educational production function. Technical report, National Bureau of Economic Research.

Borman, G. D., Dowling, N. M., and Schneck, C. (2008). A multisite cluster randomized field trial of open court reading. *Educational Evaluation and Policy Analysis*, 30(4):389–407.

Bulman, G. and Fairlie, R. W. (2016). Technology and education: Computers, software, and the internet. In *Handbook of the Economics of Education*, volume 5, pages 239–280. Elsevier.

de Chaisemartin, C. and Ramirez-Cuellar, J. (2019). At what level should one cluster standard errors in paired experiments? *arXiv preprint arXiv:1906.00288*.

Doss, C. J., Fahle, E. M., Loeb, S., and York, B. N. (2018). More than just a nudge: Supporting kindergarten parents with differentiated and personalized text-messages. Technical report, National Bureau of Economic Research.

Duflo, E., Dupas, P., and Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review*, 101(5):1739–74.

Elbow, P. (1981). *Writing with power: Techniques for mastering the writing process*. Oxford University Press.

Ferman, B. (2019). A simple way to assess inference methods. *arXiv preprint arXiv:1912.08772*.

Ferman, B., Finamor, L., and Lima, L. (2019). Are public schools ready to integrate math classes with khan academy?

Ferman, B. and Ponczek, V. (2017). Should we drop covariate cells with attrition problems? Mpra paper, University Library of Munich, Germany.

Fonseca, E., Medeiros, I., Kamikawachi, D., and Bokan, A. (2018). Automatically grading brazilian student essays. In *Computational Processing of the Portuguese Language (Lecture Notes in Computer Science, vol 11122)*. Springer.

Fryer, R. G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experimentsa. In *Handbook of Economic Field Experiments*, volume 2, pages 95–322. Elsevier.

Grimes, D. and Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *The Journal of Technology, Learning and Assessment*, 8(6).

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

INEP/MEC (2018). Redação do enem 2018. *Cartilha do Participante*.

INEP/MEC (2019). Redação do enem 2019. *Cartilha do Participante*.

Jones, S. M., Brown, J. L., and Lawrence Aber, J. (2011). Two-year impacts of a universal school-based social-emotional and literacy intervention: An experiment in translational developmental research. *Child Development*, 82(2):533–554.

Kim, J. S., Olson, C. B., Scarcella, R., Kramer, J., Pearson, M., van Dyk, D., Collins, P., and Land, R. E. (2011). A randomized experiment of a cognitive strategies approach to text-based analytical writing for mainstreamed latino english language learners in grades 6 to 12. *Journal of Research on Educational Effectiveness*, 4(3):231–263.

Knudsen, E. I., Heckman, J. J., Cameron, J. L., and Shonkoff, J. P. (2006). Economic, neurobiological, and behavioral perspectives on building america's future workforce. *Proceedings of the national Academy of Sciences*, 103(27):10155–10162.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, 76(3):1071–1102.

Luckin, R., Holmes, W., Griffiths, M., and Forcier, L. B. (2016). Intelligence unleashed: An argument for ai in education.

McCurry, D. (2012). Computer scoring and quality of thought in assessing writing.

Mokyr, J., Vickers, C., and Ziebarth, N. L. (2015). The history of technological anxiety and the future of economic growth: Is this time different? *Journal of economic perspectives*, 29(3):31–50.

Mooney, P. J. (2003). An investigation of the effects of a comprehensive reading intervention on the beginning reading skills of first graders at risk for emotional and behavioral disorders.

Morrow, L. M. (1992). The impact of a literature-based program on literacy achievement, use of literature, and attitudes of children from minority backgrounds. *Reading Research Quarterly*, pages 251–275.

Muralidharan, K., Singh, A., and Ganimian, A. J. (2019). Disrupting education? experimental evidence on technology-aided instruction in india. *American Economic Review*.

Neumann, A. (2012). Advantages and disadvantages of different text coding procedures for research and practice in a school context. In *Measuring Writing: Recent Insights into Theory, Methodology and Practice*, pages 33–54. Brill.

Palermo, C. and Thomson, M. M. (2018). Teacher implementation of self-regulated strategy development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students. *Contemporary Educational Psychology*, 54:255–270.

Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing*, 21:104–111.

Pinnell, G. S., Lyons, C. A., Deford, D. E., Bryk, A. S., and Seltzer, M. (1994). Comparing instructional models for the literacy education of high-risk first graders. *Reading Research Quarterly*, pages 9–39.

Puma, M., Bell, S., Cook, R., Heid, C., and Lopez, M. (2005). Head start impact study: First year findings. *Administration for Children & Families*.

Shermis, M. D., Burstein, J. C., and Bliss, L. (2008). The impact of automated essay scoring on high stakes writing assessments. In *annual meeting of the National Council on Measurement in Education*.

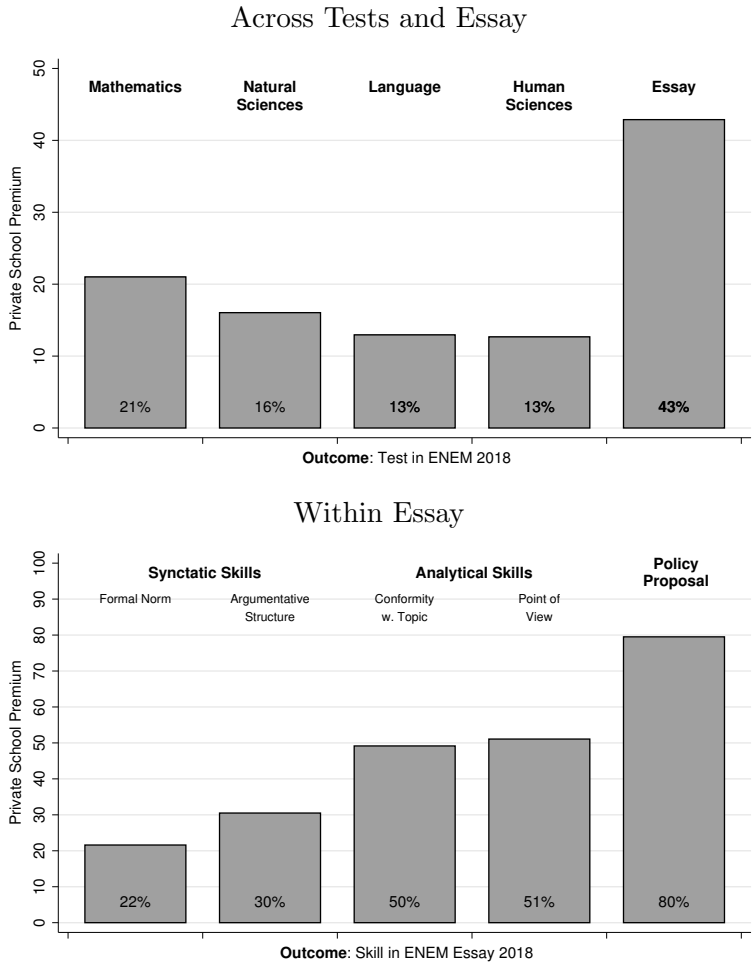
Somers, M.-A., Corrin, W., Sepanik, S., Salinger, T., Levin, J., and Zmach, C. (2010). The enhanced reading opportunities study final report: The impact of supplemental literacy courses for struggling ninth-grade readers. ncee 2010-4021. *National Center for Education Evaluation and Regional Assistance*.

Wilson, J. and Roscoe, R. D. (2019). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, page 0735633119830764.

York, B. N. and Loeb, S. (2014). One step at a time: The effects of an early literacy text messaging program for parents of preschoolers. Technical report, National Bureau of Economic Research.

Young, A. (2018). Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results*. *The Quarterly Journal of Economics*, 134(2):557–598.

Figure 1: Private School Premium in ENEM 2018

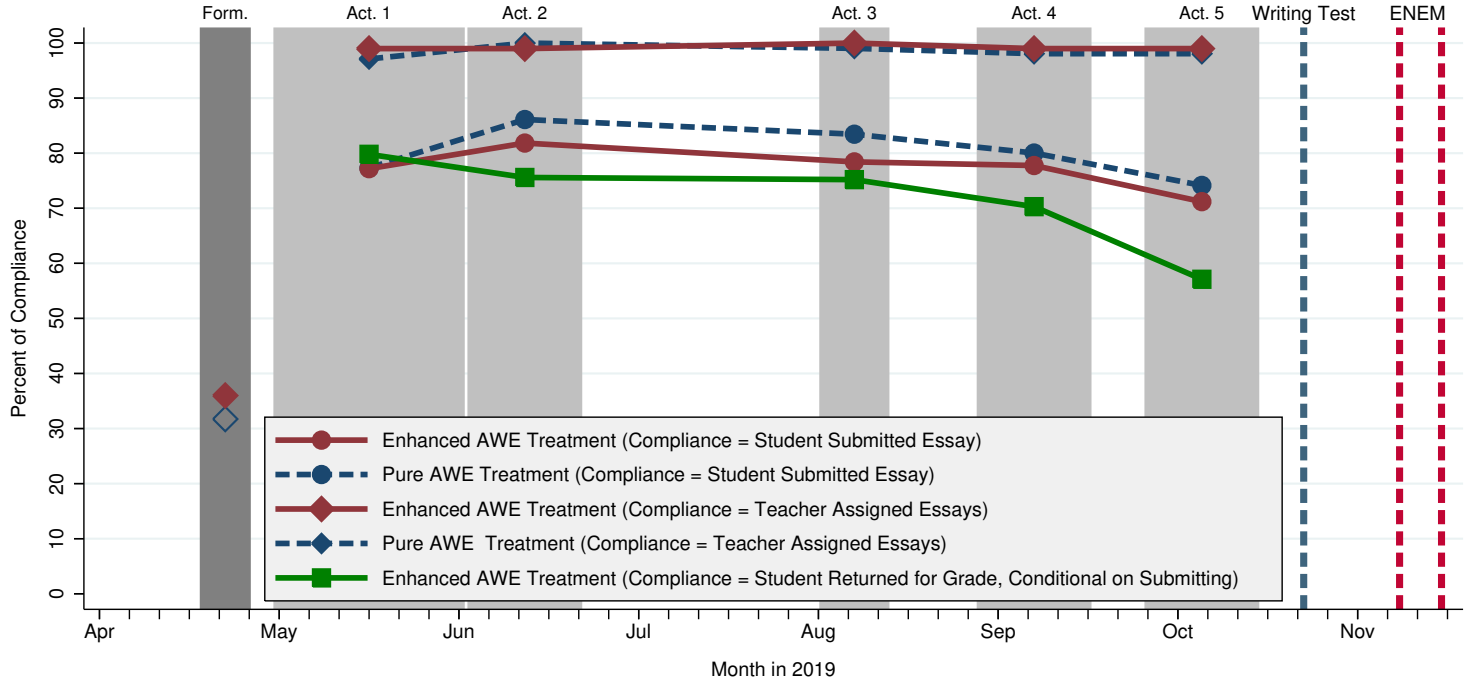


[BACK TO TEXT]

1. These bar graphs illustrate the magnitude of the achievement gaps between public and private schools in the different exams and in the essay of the Brazilian National Secondary Education Exam using data on the universe of high school seniors in Brazil that took each of the tests in 2018. The exam is currently composed of 180 multiple-choice questions, equally divided into four areas of knowledge (Mathematics, Natural Sciences, Language and Codes, Human Sciences), and one written essay. The upper figure relates to the five tests that compose the exam. The lower figure considers each competency in the written essay individually. We excluded from the sample students from schools that are administered at the federal level, which are typically very different from other public schools in Brazil.

2. *Syntactic skills* comprise two competencies: "exhibiting command of the formal written norm of Brazilian Portuguese" and "exhibiting knowledge of the linguistic mechanisms that lead to the construction of the argument"; *Analytic skills* comprise two competencies: "understanding the proposed topic and applying concepts from different areas of knowledge to develop the argument following the structural limits of the dissertative-argumentative prose" and "selecting, relating, organizing and interpreting information, facts, opinions and arguments in defense of a point of view, using pieces of knowledge acquired in the motivating elements and during the schooling"; *Policy proposal* comprises one sub-skill: " elaborating a policy proposal that could contribute to solve the problem in question, respecting basic human rights " (INEP/MEC, 2018).

Figure 2: Timeline and Compliance Among Teachers and Students, by Treatment Arm



[BACK TO TEXT]

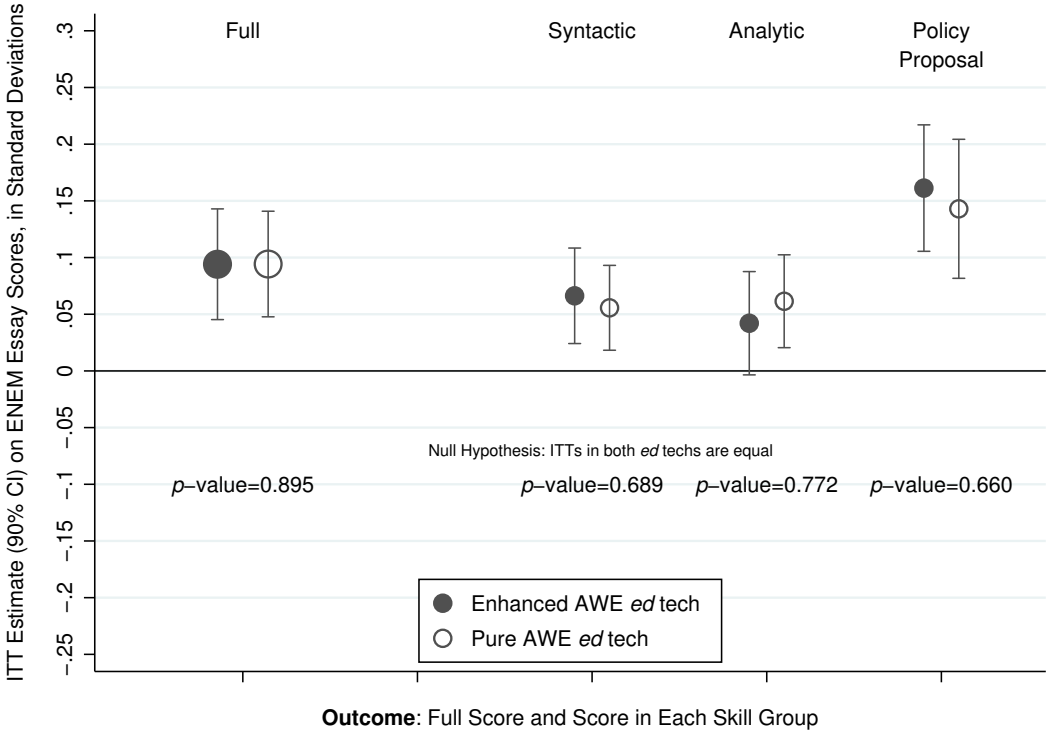
1. This figure presents a timeline of the experiment and indicators of compliance of teachers and students with the interventions in both treatment arms. We always denote compliance events associated with the pure AWE (enhanced) treatment with (non-)dashed lines.

2. The darkly shaded area in April represents the period when teachers and other school professionals were introduced to the *ed*-techs in both treatment arms, either in itinerant presentations or through an online course. Compliance with these presentations is denoted by the hollowed and full diamonds in this darkly shaded region. The p-value from a simple test of proportion equality between treatment arms using a regression of the teachers' indicator of presence in the presentation on strata indicators is 0.469 (standard errors clustered at the strata level).

3. The lightly shaded areas represent the periods when the platform was available for submission of essays for each of the five writing activities of the two interventions, which ran concomitantly. For each activity, the lines present the evolution of compliance throughout the 2019 academic year, for students (connecting circles) and teachers (connecting diamonds). In these two pairs of lines, compliance of students (teachers) is defined as the submission of an essay (the event of having students submit essays through the platform) for all high school senior classes taught. The p-values for comparisons of "parallel trends" are 0.464 for students *ed* tech and 0.245 for teachers. The green line in the bottom of the figure connecting squares depicts the proportion of students in the enhanced treatment arm that submitted essays *and* entered the platform to check the human grader grading for her essay. We can reject a null hypothesis of constant compliance along this margin (p-value¡0.001).

4. The dashed vertical lines in October and November denote, respectively, the writing test we administered with the collaboration of the State's Education Secretary (October 18th) and the ENEM test (two consecutive Sundays, November 3rd and November 10th).

Figure 3: ITT Effects of *ed* techs on ENEM Essay Scores



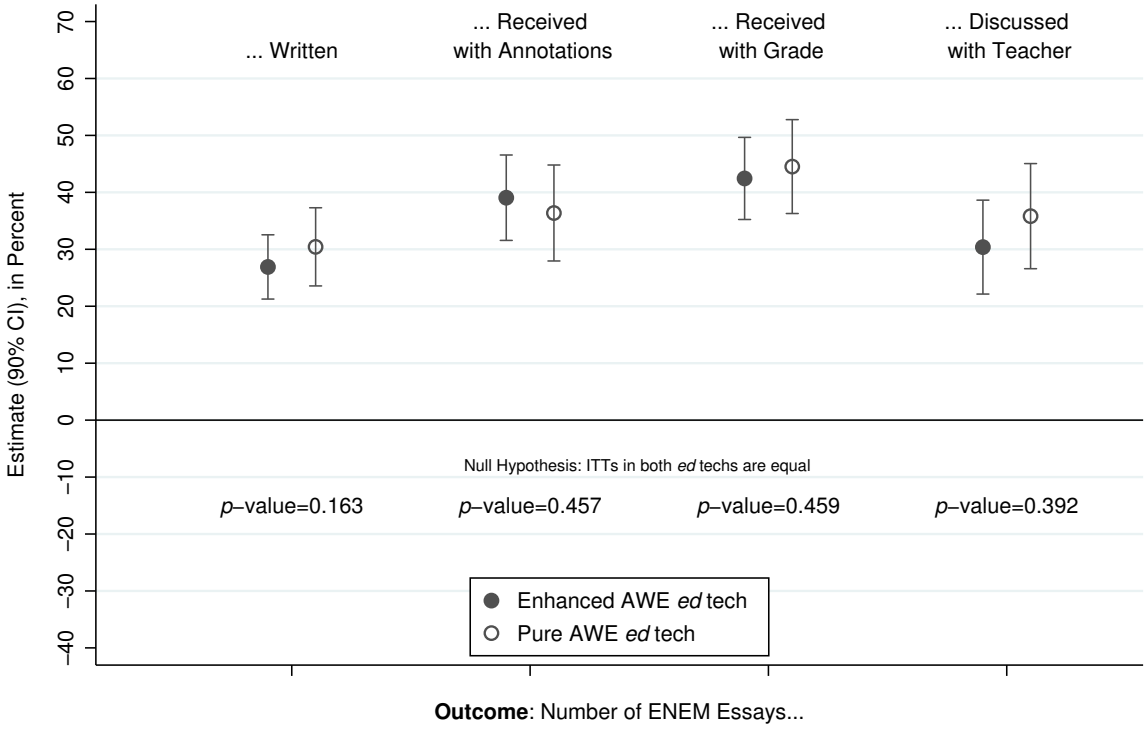
[BACK TO TEXT]

1. This figure plots the average intention-to-treat effects and 90% confidence intervals of the enhanced (full circles) and pure AWE (Automated Writing Evaluation, hollowed circles) *ed*-techs on the Brazilian National Secondary Education Exam essay scores. The pure AWE *ed*-tech is a fully automated system that provides instantaneous scores and feedback to students using natural language processing and machine-learning. The enhanced AWE *ed*-tech uses human graders as an additional resource to enhance grading and feedback quality. Estimates and standard errors used to construct confidence intervals are from specification (1), an ordinary least squares regression with treatment indicators, strata indicators, school-level controls and individual-level controls. Details on controls are in the notes to Table 3.

2. The unit of observation is a score of a student that participated in the official ENEM 2019 written essay or in an independently administered essay with the same structure and grading criteria as the ENEM essay (N=29,359). The topic of the essays were "Democratization of Access to Cinema in Brazil" and "The Construction of a National Brazilian Identity for the Portuguese Language". The description of the competencies comprised in each set of skill can be found in the notes to Figure 1.

3. The p-values are for tests of no difference between effects in both treatment arms are computed using standard errors clustered at the strata level and specification (2), which uses only data from treated schools and additionally controls for the student AI-provided grade on the first writing activity of the year.

Figure 4: ITT Effects of *ed* techs on Training, Feedback and Individualized Pedagogy



[BACK TO TEXT]

1. This figure plots the average intention-to-treat effects and 90% confidence intervals of the enhanced (full circles) and pure AWE (hollowed circles) *ed*-techs on individual training and feedback in its many potential forms —comments on essays, grades and individual discussions with teachers— responded to the incorporation of the *ed* techs. The inputs to the figure are estimated using specification (3), an ordinary least squares regression with treatment indicators, strata indicators, school-level controls and individual-level controls. Details on controls are in the notes to Table 4.

2. The unit of observation is a student that was present in the exam and provided valid answers to the multiple-choice questions. The variables are, in order:

- the number of essays written to train for the ENEM in 2019, top-coded at 10;
- the number of ENEM training essays that received individualized annotations;
- the number of ENEM training that received a grade;
- the number of ENEM training essays graded that were followed by a personal discussion with the teacher.

3. The p-values are for tests of no difference between effects in both treatment arms are computed using standard errors clustered at the strata level.

Table 1: Design Validity — Balance Across Treatment Arms

	Enh. AWE - Control	Pure AWE - Control	Enh. AWE - Pure AWE	*p*-values (clust. strata) (2)=0	(3)=0	(4)=0	Obs.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. ENEM 2018 Cohort							
Essay Full Score	-0.020σ	-0.037σ	0.018σ	0.667	0.167	0.691	17,218
	(0.045)	(0.026)	(0.044)				
Syntactic Skills	-0.013σ	-0.026σ	0.013σ	0.685	0.333	0.743	17,218
	(0.033)	(0.027)	(0.039)				
Analytical Skills	-0.016σ	-0.04σ	0.024σ	0.719	0.102	0.553	17,218
	(0.045	(0.024)	(0.040)				
Policy Proposal	-0.023σ	-0.026σ	0.003σ	0.607	0.432	0.955	17,218
	(0.045)	(0.032)	(0.046)				
Language and Codes	-0.047σ	-0.048σ	0.001σ	0.240	0.184	0.993	17,218
	(0.039)	(0.035)	(0.039)				
Mathematics	-0.042σ	-0.034σ	-0.007σ	0.264	0.335	0.887	16,349
	(0.037)	(0.035)	(0.052)				
Natural Sciences	-0.018σ	-0.043σ	0.025σ	0.631	0.319	0.534	16,349
	(0.038)	(0.043)	(0.039)				
Human Sciences	-0.018σ	-0.041σ	0.023σ	0.622	0.155	0.557	17,218
	(0.037)	(0.028)	(0.039)				
B. Main Sample Cohort							
Baseline Language Score	-0.074σ	0.011σ	-0.080σ	0.146	0.838	0.110	17,739
	(0.049)	(0.053)	(0.051)				
Baseline Math Score	-0.063σ	-0.059σ	0.001σ	0.403	0.325	0.957	17,739
	(0.074)	(0.059)	(0.073)				
Joint test (*p*-value)				0.819	0.802	0.701	

Notes: This table investigates balance with respect to student-level variables across experiment arms. Columns 1, 2 and 3 present estimates and standard errors clustered at the strata-level computed using an ordinary least squares regression with indicators for each of the two experiment arms and strata indicators. Columns 4, 5 and 6 present *p*-values testing that the treatment indicators (columns 4 and 5, for the enhanced AWE *ed* tech, and the pure AWE *ed* tech, respectively) and their difference (column 6) are zero. Column 7 presents the number of observations used for inference tests for each variable. We also present, in the bottom rows of the table, *p*-values from a joint test that all covariates are balanced in each comparison. These *p*-values are constructed based on equation (7) from Young (2018), taking into account the randomization protocol. [BACK TO TEXT]

Table 2: Design Validity — Treatment Status and Samples' Attrition

	Control Mean	Only Strata FEs			Strata FEs + Controls			
		Enhanced AWE	Pure AWE	(2)-(3)	Enhanced AWE	Pure AWE	(5)-(6)	Obs.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Attrition								
Nonofficial ENEM Essay Attriter	0.22	0.016	0.012	0.004	0.007	0.002	0.005	19,516
		(0.014)	(0.019)	(0.021)	(0.012)	(0.016)	(0.019)	
Students' Questionnaire Attriter	0.17	0.014	0.018	-0.004	0.007	0.01	-0.003	19,516
		(0.010)	(0.012)	(0.016)	(0.009)	(0.012)	(0.016)	
Official ENEM Essay Attriter	0.27	0.001	-0.025	0.026	-0.001	-0.029	0.028	178
		(0.024)	(0.028)	(0.032)	(0.025)	(0.028)	(0.032)	
Teacher Survey Attriter	0.12	0.076	0.026	0.050	0.072	0.016	0.055	324
		(0.043)	(0.043)	(0.055)	(0.043)	(0.044)	(0.054)	
End-of-year Student Composition								
Not Enrolled in Same School in April 2019	0.01	0.001	0.001	0.001	0.001	0.001	0.001	17,872
		(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	

Notes: This table presents estimates and inference tests for attrition in our main analytical samples. Column 1 presents attrition rates in the control group and columns 2, 3 and 4 present estimates from specification (3), an ordinary least squares regression with indicators for each of the two experiment arms and strata indicators. In columns 5 to 7, we add to this regression the school-level and individual-level controls available in the beginning of the year that we use in our main regressions. In these columns, we always include the 2018 ENEM essay average score as a control, after replacing these observations with the control group school sample mean. Additional controls for the nonofficial ENEM essay and the student's questionnaire are the baseline Language and baseline Mathematics proficiency scores using data from another state's standardized exam that happened right before the treatments were implemented. These controls are also included in the regressions on end-of-year student composition. [BACK TO TEXT]

Table 3: Treatments and ENEM Essay Scores

			Score, By Skill Group				
	Full Score	=	Syntactic Skills	+	Analytic Skills	+	Policy Proposal
	(1)		(2)		(3)		(4)
	Panel A. Main Effects — Specification (1)						
$\widehat{\tau^{\text{Enhanced AWE}}_{\text{ITT}}}$	0.094σ		0.066σ		0.042σ		0.161σ
(*s.e.*, clust. strata)	(0.030)		(0.026)		(0.028)		(0.034)
p-value, clust. strata	0.003		0.015		0.140		<0.001
p-value, rand. inf	0.005		0.014		0.152		0.001
p-value, MHT adj.	0.005		0.056		0.152		0.006
$\widehat{\tau^{\text{Pure AWE}}_{\text{ITT}}}$	0.094σ		0.056σ		0.061σ		0.143σ
(*s.e.*, clust. strata)	(0.028)		(0.023)		(0.025)		(0.037)
p-value, clust. strata	0.002		0.020		0.019		0.001
p-value, rand. inf	0.003		0.032		0.016		0.001
p-value, MHT adj.	0.006		0.064		0.048		0.005
MHT Adjustment	Holm		Holm		Holm		Holm
N_{Scores}	29,359		29,359		29,359		29,359
$N_{Schools}$	178		178		178		178
N_{Strata}	33		33		33		33
	Panel B. Differential Effects — Specification (2)						
$\widehat{\tau^{\Delta}_{\text{ITT}}}$	0.006σ		0.014σ		-0.011σ		0.021σ
(*s.e.*, clust. strata)	(0.042)		(0.034)		(0.039)		(0.047)
p-value, clust. strata	0.895		0.689		0.772		0.660
p-value, rand. inf	0.896		0.676		0.755		0.670
p-value, MHT adj.	—		0.999		0.755		0.999
MHT Adjustment	Holm		Holm		Holm		Holm
N_{Scores}	17,356		17,356		17,356		17,356
$N_{Schools}$	110		110		110		110
N_{Strata}	33		33		33		33

Notes: This table presents estimates, standard errors and inference tests for the average absolute and differential intention-to-treat effects of both *ed* techs on ENEM essay scores. The unit of observation is an essay written in the official 2019 ENEM or in the unofficial ENEM (see Section 4 for details). In Panel A, estimates are from specification (1), an ordinary least squares regression with indicators for each of the two experiment arms, strata indicators and the school average ENEM essay score for the full essay score in column 1, and for the specific group of skills in columns 2, 3 and 4. We also include the following individual-level covariates, as specified in the pre-analysis plan: (i) female indicator; (ii) age dummies ranging from 17 or less to 23 or more; (iii) educational and occupational characteristics of the mother and father of the students; (iv) household income category; (v) baseline Language and baseline Mathematics proficiency scores using data from another state's standardized exam that happened right before the treatments were implemented. These covariates are interacted with the exam indicator to take into account that the set of covariates available for observations from the 2019 ENEM are different from the other exam. We also replace missing school-level and individual-level continuous covariate values with the control group mean and included an indicator for missing in this covariate in the regression. For discrete covariates we created a complementary category for missing variables. In Panel B, estimates are from specification (2), where we only use data from treated schools and control for the student AI-provided grade on the first writing activity of the year. We present standard errors clustered at the strata level in parentheses and three two-sided *p*-values: *p*-values obtained using the standard errors clustered at the strata level; randomization inference *p*-values using the randomization protocol and 1,000 draws of the assignment with replacement; and Holm (1979) adjusted *p*-values using the latter. The multiple hypothesis testing adjustments were made within the cells that have the same shaded background. [BACK TO TEXT]

Table 4: Treatments, Training, Feedback and Individualized Pedagogy

	# ENEM essays ...		Annotations were useful?		# ENEM essays ...		
Dep. Var.:	Written	Comment. Annotat.	Somewhat Useful	Very Useful	Graded	Discus. Ind. with Teacher	Summary Index
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A. Main Effects — Specification (3)						
$\widehat{\tau^{\text{Enhanced AWE}}_{\text{ITT}}}$	1.35	1.37	0.07	0.06	1.64	0.76	1.02σ
(*s.e.*, clust. strata)	(0.17)	(0.16)	(0.01)	(0.02)	(0.17)	(0.13)	(0.13)
p-value, clust. strata	0.001	0.001	0.001	0.007	0.001	0.001	0.001
p-value, rand. inf	0.001	0.001	0.001	0.001	0.001	0.001	0.001
p-value, MHT adj.	0.001	0.001	0.001	0.001	0.001	0.001	0.001
$\widehat{\tau^{\text{Pure AWE}}_{\text{ITT}}}$	1.52	1.28	0.06	0.01	1.72	0.90	0.96σ
(*s.e.*, clust. strata)	(0.21)	(0.18)	(0.01)	(0.02)	(0.19)	(0.14)	(0.15)
p-value, clust. strata	0.001	0.001	0.001	0.585	0.001	0.001	0.001
p-value, rand. inf	0.001	0.001	0.001	0.590	0.001	0.001	0.001
p-value, MHT adj.	0.001	0.001	0.001	0.590	0.001	0.001	0.001
	Panel B. Differential Effects — Specification (3)						
$\widehat{\tau^{\text{Enhanced AWE}}_{\text{ITT}} - \tau^{\text{Pure AWE}}_{\text{ITT}}}$	-0.18	0.09	0.01	0.05	-0.08	-0.14	0.06σ
(*s.e.*, clust. strata)	(0.12)	(0.13)	(0.01)	(0.02)	(0.11)	(0.16)	(0.10)
p-value, clust. strata	0.163	0.457	0.249	0.010	0.459	0.392	0.579
p-value, rand. inf	0.280	0.610	0.220	0.020	0.530	0.400	0.620
p-value, MHT adj.	0.999	0.610	0.999	0.120	0.999	0.999	—
MHT Adjustment	Holm	Holm	Holm	Holm	Holm	Holm	Holm
Control Group Mean	4.89	3.39	0.81	0.44	3.72	2.32	—
Control Group SD	3.17	2.99	0.39	0.50	3.05	2.76	—
$N_{Students}$	14,175	14,180	14,151	14,151	14,162	14,123	13,963
$N_{Schools}$	178	178	178	178	178	178	178
N_{Strata}	33	33	33	33	33	33	33

Notes: This table presents estimates, standard errors and inference tests for the average absolute and differential intent-to-treat effects of both *ed* techs on training behavior and on the quantity and quality of feedback. All outcomes in columns (1) to (6) were collected in the students' questionnaire in the state's standardized exam. In Panel A and B, estimates are from specification (3), an ordinary least squares regression with indicators for each of the two experiment arms, using as controls strata indicators and the school average ENEM essay score for the full essay score in column 1, and for the specific group of skills in columns 2, 3 and 4. We also include the following individual-level covariates, as specified in the pre-analysis plan: (i) female indicator; (ii) age dummies ranging from 17 or less to 23 or more; (iii) educational and occupational characteristics of the mother and father of the students; (iv) household income category; (v) baseline Language and baseline Mathematics proficiency scores using data from another state's standardized exam that happened right before the treatments were implemented. We present standard errors clustered at the strata level in parentheses and three two-sided *p*-values: *p*-values obtained using the standard errors clustered at the strata level; randomization inference *p*-values using the randomization protocol and 1,000 draws of the assignment with replacement; and Holm (1979) adjusted *p*-values using the latter. The multiple hypothesis testing adjustments were made within the cells that have the same shaded background. The summary index in column (7) is computed based on the procedure suggested by Anderson (2008) and relies on constructing a positively weighted mean of the standardized outcomes of the observations with non-missing outcomes in columns (1)-(6). [BACK TO TEXT]

Table 5: Treatments and Teachers' Perception on Time Constraints

	Time available to improve students' knowledge in ... (1-5 scale)			
Dep. Var.:	Writing	Grammar	Literature	Summary Index
	(1)	(2)	(3)	(4)
	Panel A. Main Effects — Specification (3)			
$\widehat{\tau^{\text{Enhanced AWE}}_{\text{ITT}}}$	0.30	0.36	0.32	0.26σ
(*s.e.*, clust. strata)	(0.17)	(0.18)	(0.20)	(0.15)
p-value, clust. strata	0.045	0.026	0.061	0.044
p-value, rand. inf	0.049	0.009	0.025	0.043
p-value, MHT adj.	0.196	0.054	0.125	0.086
$\widehat{\tau^{\text{Pure AWE}}_{\text{ITT}}}$	0.02	-0.01	0.01	0.02σ
(*s.e.*, clust. strata)	(0.20)	(0.18)	(0.18)	(0.17)
p-value, clust. strata	0.456	0.521	0.506	0.454
p-value, rand. inf	0.413	0.505	0.533	0.412
p-value, MHT adj.	0.999	0.999	0.533	0.412
	Panel B. Differential Effects — Specification (3)			
$\widehat{\tau^{\text{Enhanced AWE}}_{\text{ITT}} - \tau^{\text{Pure AWE}}_{\text{ITT}}}$	0.28	0.36	0.33	0.24σ
(*s.e.*, clust. strata)	(0.20)	(0.15)	(0.23)	(0.17)
p-value, clust. strata	0.173	0.025	0.170	0.154
p-value, rand. inf	0.119	0.040	0.088	0.102
{p-value, MHT adj.}	0.119	0.120	0.176	—
MHT Adjustment	Holm	Holm	Holm	Holm
Control Group Mean	2.67	3.04	2.84	—
Control Group SD	1.2	1.1	1.2	—
$N_{Teachers}$	280	279	279	279
$N_{Schools}$	173	173	173	173
N_{Strata}	33	33	33	33

Notes: This table presents estimates and inference tests for the average absolute and differential treatment effects on Language teachers' perception on time constraints to improve their students' abilities. The unit of observation is a teacher that participated in our endline survey and provided an answer for the question in each column. Estimates in both panels are from specification (3), an ordinary least squares regression with indicators for each of the two experiment arms, strata dummies, the average 2018 ENEM essay score and dummies for schools for which we don't observe this average. We present standard errors clustered at the strata level in parentheses and three upper one-sided p-values: p-values obtained using the standard errors clustered at the strata level; randomization inference p-values using the randomization protocol and 1,000 draws of the assignment with replacement; and Holm (1979) adjusted p-values using the latter. The multiple hypothesis testing adjustments were made within the cells that have the same shaded background. The summary index in column (4) is computed based on the procedure suggested by Anderson (2008) for observations with non-missing outcomes in columns (1)-(3). [BACK TO TEXT]

Table 6: Treatments, Labor Supply and Task Time Allocation

Dep. Var.:	Working outside school	Correcting written essays	Correcting other homework	Preparing classes	Providing individual support	Share Non-Rout. Tasks
	Average hours worked weekly *per* group of task...					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A. Main Effects — Specification (3)					
$\widehat{\tau^{\text{Enhanced AWE}}_{\text{ITT}}}$	-1.19	-0.47	-0.02	-0.71	-0.71	2.44
(*s.e.*, clust. strata)	(0.98)	(0.75)	(0.52)	(1.21)	(0.72)	(2.18)
p-value, clust. strata	0.116	0.539	0.975	0.564	0.328	0.272
p-value, rand. inf	0.117	0.568	0.973	0.489	0.302	0.196
{*p*-value, MHT adj.}	0.234	0.999	0.973	0.999	0.999	0.392
		p-value (diff. in allocation, χ^2) = 0.469				
$\widehat{\tau^{\text{Pure AWE}}_{\text{ITT}}}$	-0.15	-0.27	-0.02	0.07	-0.35	1.84
(*s.e.*, clust. strata)	(1.16)	(0.82)	(0.60)	(1.16)	(0.60)	(1.83)
p-value, clust. strata	0.895	0.744	0.978	0.954	0.566	0.323
p-value, rand. inf	0.873	0.756	0.979	0.946	0.537	0.278
{*p*-value, MHT adj.}	0.873	0.999	0.999	0.999	0.999	0.278
		p-value (diff. in allocation, χ^2) = 0.982				
	Panel B. Differential Effects — Specification (3)					
$\widehat{\tau^{\text{Enhanced AWE}}_{\text{ITT}} - \tau^{\text{Pure AWE}}_{\text{ITT}}}$	-1.03	-0.20	0.01	-0.77	-0.37	0.60
(*s.e.*, clust. strata)	(0.90)	(0.84)	(0.67)	(1.19)	(0.63)	(2.19)
p-value, clust. strata	0.258	0.815	0.999	0.519	0.568	0.786
p-value, rand. inf	0.189	0.779	0.996	0.458	0.462	0.758
p-value, MHT adj.	—	0.999	0.996	0.999	0.999	
		p-value (diff. in allocation, χ^2) = 0.628				
MHT Adjustment	Holm	Holm	Holm	Holm	Holm	Holm
Control Group Mean	6.14	6.05	3.82	10.19	3.14	37.2
Control Group SD	5.88	5.41	3.29	6.97	4.50	10.6
$N_{Teachers}$	270	264	270	273	265	262
$N_{Schools}$	173	173	173	173	173	173
N_{Strata}	33	33	33	33	33	33

Notes: This table presents estimates and inference tests for the average absolute and differential treatment effects on Language teachers' labor supply (in total and outside schools) and time allocation across different types of tasks. All outcomes were initially elicited in hours and then winsorized at the top 1%. The unit of observation is a teacher that participated in our endline survey and provided an answer for the question in each column. Estimates in both panels are from specification (3), an ordinary least squares regression with indicators for each of the two experiment arms, strata dummies, the average 2018 ENEM essay score and dummies for schools for which we don't observe this average. We present standard errors clustered at the strata level in parentheses and three two-sided (upper one-sided, in column 1) *p*-values: *p*-values obtained using the standard errors clustered at the strata level; randomization inference *p*-values using the randomization protocol and 1,000 draws of the assignment with replacement; and Holm (1979) adjusted *p*-values using the latter. The multiple hypothesis testing adjustments were made within the cells that have the same shaded background. [BACK TO TEXT]

Table 7: Treatments and Future Education

Dep. Var.:	Share admitted PSE 2020	Plans for 2020 include PSE	Summary Index
	(1)	(2)	(3)
	Panel A. Main Effects — Specification (3)		
$\widehat{\tau^{\text{Enhanced AWE}}_{\text{ITT}}}$	4.95	0.01	0.41σ
(*s.e.*, clust. strata)	(4.22)	(0.01)	(0.47)
p-value, clust. strata	0.250	0.652	0.390
p-value, rand. inf	0.201	0.602	0.344
{p-value, MHT adj.}	0.804	0.999	0.688
$\widehat{\tau^{\text{Pure AWE}}_{\text{ITT}}}$	1.86	0.01	0.32σ
(*s.e.*, clust. strata)	(4.20)	(0.01)	(0.43)
p-value, clust. strata	0.661	0.687	0.468
p-value, rand. inf	0.619	0.609	0.379
{p-value, MHT adj.}	0.619	0.999	0.379
	Panel B. Differential Effects — Specification (3)		
$\widehat{\tau^{\text{Enhanced AWE}}_{\text{ITT}} - \tau^{\text{Pure AWE}}_{\text{ITT}}}$	3.09	0.01	0.10σ
(*s.e.*, clust. strata)	(4.70)	(0.02)	(0.60)
p-value, clust. strata	0.516	0.957	0.874
p-value, rand. inf	0.455	0.953	0.843
{p-value, MHT adj.}	0.999	0.953	—
MHT Adjustment	Holm	Holm	Holm
Control Group Mean	41.06	0.73	—
Control Group SD	22.73	0.44	—
Regression Level	Teacher	Student	School
N	272	14,152	163

Notes: This table presents estimates and inference tests for the average absolute and differential treatment effects Language teachers' expectations with respect to admission into PSE and and on students' aspirations with respect to post-secondary education (PSE). The unit of observation is a a teacher that participated in our endline survey and provided an answer for the question in column (1) and a student in experimental schools who participated in the state's standardized test in 2019 and provided a valid answer in column (2). Estimates in both panels are from specification (3), an ordinary least squares regression with indicators for each of the two experiment arms, strata dummies, the average 2018 ENEM essay score and dummies for schools for which we don't observe this average and other individual-level controls we are able to link to students in our data. We present standard errors clustered at the strata level in parentheses and three two-sided p-values: p-values obtained using the standard errors clustered at the strata level; randomization inference p-values using the randomization protocol and 1,000 draws of the assignment with replacement; and Holm (1979) adjusted p-values using the latter. The multiple hypothesis testing adjustments were made within the cells that have the same shaded background. The summary index in column 3 is computed based on the procedure suggested by Anderson (2008) and relies on constructing a positively weighted mean of the standardized outcomes of the observations with non-missing outcomes in columns 1 and 2. Since we cannot link students' answers in the questionnaire with the teacher data, we collapse both answers at the school level in order to compute estimates on the summary index. [BACK TO TEXT]

Table 8: Treatments and Knowledge About Students

Dep. Var.:	How much feels knows strengths and weaknesses of students (1-10 scale)... Writing	Grammar	Literature	Pred. - Real 2019 Essay	Summary Index
	(1)	(2)	(3)	(4)	(5)
	Panel A. Main Effects — Specification (3)				
$\widehat{\tau^{\text{Enhanced AWE}}_{\text{ITT}}}$	0.05	-0.17	0.08	2.2	0.54σ
(*s.e.*, clust. strata)	(0.18)	(0.21)	(0.19)	(24.0)	(0.95)
p-value, clust. strata	0.782	0.410	0.682	0.928	0.573
p-value, rand. inf.	0.834	0.442	0.702	0.914	0.552
{p-value, MHT adj.}	0.999	0.999	0.999	0.999	0.552
$\widehat{\tau^{\text{Pure AWE}}_{\text{ITT}}}$	0.01	-0.24	0.04	-30.3	-0.57σ
(*s.e.*, clust. strata)	(0.17)	(0.21)	(0.25)	(21.9)	(0.69)
p-value, clust. strata	0.978	0.257	0.873	0.176	0.415
p-value, rand. inf.	0.983	0.289	0.884	0.132	0.498
{p-value, MHT adj.}	0.983	0.999	0.999	0.999	0.999
	Panel B. Differential Effects — Specification (3)				
$\widehat{\tau^{\text{Enhanced AWE}}_{\text{ITT}} - \tau^{\text{Pure AWE}}_{\text{ITT}}}$	0.05	0.07	0.04	32.5	1.11σ
(*s.e.*, clust. strata)	(0.23)	(0.29)	(0.32)	(20.4)	(0.90)
p-value, clust. strata	0.813	0.812	0.906	0.121	0.227
p-value, rand. inf.	0.828	0.786	0.891	0.071	0.229
{p-value, MHT adj.}	0.999	0.999	0.891	0.284	—
MHT Adjustment	Holm	Holm	Holm	Holm	Holm
Control Group Mean	8.19	8.34	7.98	85.6	—
Control Group SD	1.32	1.19	1.38	125.7	—
Regression Level	Teacher	Teacher	Teacher	School	School
N	279	278	278	164	164

Notes: This table presents estimates and inference tests for the average absolute and differential treatment effects on teachers' perceptions on how much they know about the strengths and weaknesses of their students in writing essays, and on Grammar and Literature, in a scale of 1 to 10 (columns 1 to 3) and on the absolute difference between teachers' predicted average grade of their students in public schools in the written essay of ENEM 2019 and the actual average grade in the exam's essay, at the school level (column 4). The unit of observation in columns 1 to 3 is a teacher that participated in our endline survey and provided an answer for the question in each column. Estimates in both panels are from specification (3), an ordinary least squares regression with indicators for each of the two experiment arms, strata dummies, the average 2018 ENEM essay score and dummies for schools for which we don't observe this average. We present standard errors clustered at the strata level in parentheses and three two-sided p-values: p-values obtained using the standard errors clustered at the strata level; randomization inference p-values using the randomization protocol and 1,000 draws of the assignment with replacement; and Holm (1979) adjusted p-values using the latter. The multiple hypothesis testing adjustments were made within the cells that have the same shaded background. The summary index in column 5 is computed based on the procedure suggested by Anderson (2008) and relies on constructing a positively weighted mean of the standardized outcomes of the observations with non-missing outcomes in columns 1 to 4, after collapsing the data at the school level using the number of student's of each teachers as weights for variables in columns 1 to 3. [BACK TO TEXT]

Table 9: Treatments and Secondary Outcomes

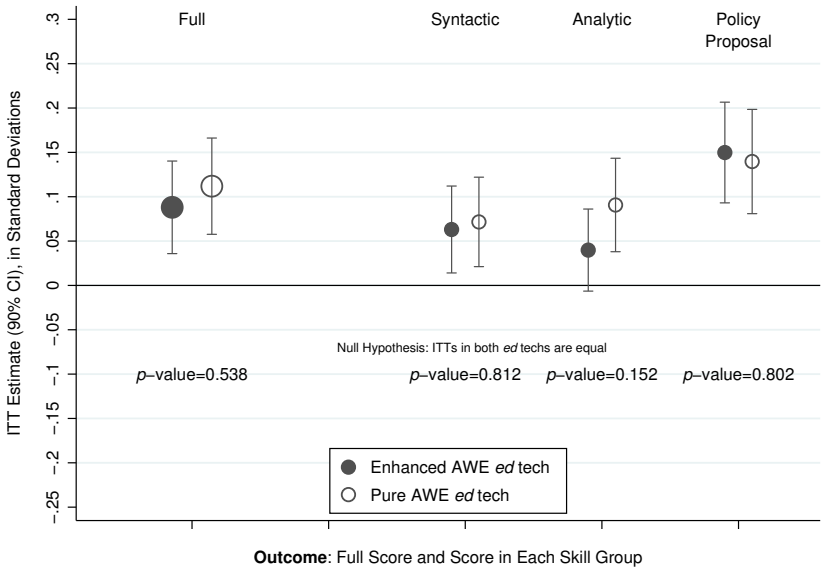
Dep. Var.:	Scores in Writing (Narrative Biography)	Scores in Language Related Tests	Scores in Non-Language Related Tests
	(1)	(2)	(3)
	Panel A. Main Effects — Specification (3)		
$\widehat{\tau^{\text{Enhanced AWE}}_{\text{ITT}}}$	-0.001σ	-0.007σ	0.001σ
(*s.e.*, clust. strata)	(0.035)	(0.021)	(0.021)
p-value, clust. strata	0.974	0.756	0.950
p-value, rand. inf	0.973	0.730	0.951
p-value, MHT adj.	0.973	0.999	0.999
$\widehat{\tau^{\text{Pure AWE}}_{\text{ITT}}}$	0.041σ	-0.001σ	0.005σ
(*s.e.*, clust. strata)	(0.044)	(0.023)	(0.019)
p-value, clust. strata	0.361	0.960	0.807
p-value, rand. inf	0.352	0.969	0.862
p-value, MHT adj.	0.999	0.999	0.999
	Panel B. Differential Effects — Specification (3)		
$\widehat{\tau^{\text{Enhanced AWE}}_{\text{ITT}} - \tau^{\text{Pure AWE}}_{\text{ITT}}}$	-0.042σ	-0.006σ	-0.003σ
(*s.e.*, clust. strata)	(0.037)	(0.029)	(0.024)
p-value, clust. strata	0.267	0.850	0.892
p-value, rand. inf	0.355	0.812	0.891
p-value, MHT adj.	0.999	0.999	0.908
MHT Adjustment	Holm	Holm	Holm
N_{Scores}	15,032	30,608	90,198

Notes: This table presents estimates and inference tests for the average absolute and differential treatment effects on secondary outcomes. In column 1, the outcome is the standardized grade in a narrative essay administered at the same day as the unofficial ENEM. In column 2, we pool standardized scores in the ENEM 2019 Language and Codes test and the PAEBES 2019 Language (reading) exam administered by SEDU/ES. In column 3, we pool standardized scores in the ENEM 2019 Mathematics, Natural Sciences, Human Sciences tests, and the PAEBES 2019 Mathematics, Physics and Chemistry standardized exams. Estimates in both panels are from specification (1), an ordinary least squares regression with indicators for each of the two experiment arms, strata indicators and the school average ENEM 2018 scores related to each family of outcomes. We also include the following individual-level covariates, as specified in the pre-analysis plan: (i) female indicator; (ii) age dummies ranging from 17 or less to 23 or more; (iii) educational and occupational characteristics of the mother and father of the students; (iv) household income category; (v) baseline Language and baseline Mathematics proficiency scores using data from another state's standardized exam that happened right before the treatments were implemented. These covariates are interacted with the exam indicator to take into account that the set of covariates available for observations from the 2019 ENEM are different from the other exam. We also replace missing school-level and individual-level continuous covariate values with the control group mean and included an indicator for missing in this covariate in the regression. For discrete covariates we created a complementary category for missing variables. We present standard errors clustered at the strata level in parentheses and three two-sided *p*-values: *p*-values obtained using the standard errors clustered at the strata level; randomization inference *p*-values using the randomization protocol and 1,000 draws of the assignment with replacement; and Holm (1979) adjusted *p*-values using the latter. The multiple hypothesis testing adjustments were made within the cells that have the same shaded background. [BACK TO TEXT]

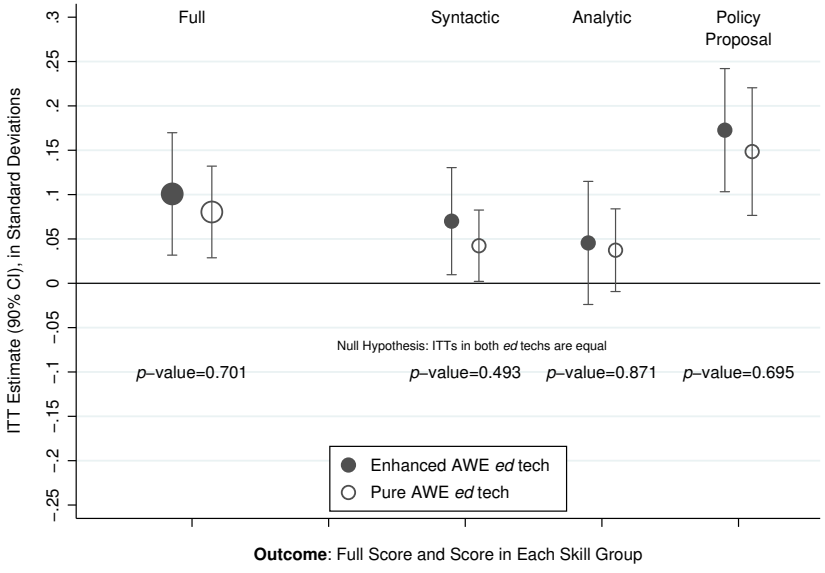
Appendix A Supplementary Material

Figure A.1: Treatment Effects on Scores (Adm. and Primary Data)

Official ENEM 2019 Essay



Nonofficial ENEM 2019 Essay



Notes: This figure replicates the main results of the paper by using each set of data on scores separately. The specification and controls included are the same as in Table 3 and Figure 3 [BACK TO TEXT]

Table A.1: Descriptive Statistics, Sample Selection and Balance Tests

	Sample Selection			*Balance*						
	Espírito Santo	Exper. Sample	t-test p-value	Control Schools	AWE Schools	Stand. Schools	(4)=(5) p-value	(4)=(6) p-value	(5)=(6) p-value	All eq. p-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
ENEM 2018										
Written Essay	457.0	476.0	**0.066**	480.2	470.9	476.5	0.260	0.596	0.677	0.530
	(180.4)	(176.3)		(175.1)	(177.2)	(176.7)				
Formal written	113.6	117.2	**0.023**	117.9	116.4	117.2	0.428	0.538	0.924	0.692
norm	(33.9)	(32.2)		(31.9)	(32.5)	(32.4)				
Argumentative	108.6	111.8	**0.082**	112.6	111.1	111.8	0.377	0.717	0.646	0.675
structure	(35.5)	(34.2)		(33.7)	(34.7)	(34.2)				
Conformity	88.5	92.8	**0.075**	93.6	91.7	92.9	0.242	0.677	0.566	0.499
	(47.4)	(47.4)		(47.5)	(47.4)	(47.1)				
POV construction	83.6	87.9	**0.074**	88.9	86.6	88.2	0.136	0.645	0.448	0.319
	(44.6)	(44.3)		(44.3)	(44.3)	(44.3)				
Policy Proposal	62.7	66.3	0.154	67.3	65.0	66.4	0.461	0.551	0.972	0.731
	(47.7)	(48.0)		(47.9)	(47.9)	(48.4)				

(*cont.*)

Language and Codes	500.7	509.8	**0.022**	512.4	507.3	509.4	0.200	0.260	0.993	0.398
	(66.6)	(66.1)		(66.7)	(65.7)	(65.6)				
Mathematics	511.8	518.2	0.119	521.7	514.8	517.6	0.389	0.342	0.828	0.576
	(86.1)	(85.9)		(87.8)	(84.1)	(85.3)				
Natural Sciences	474.5	478.3	0.255	480.5	475.0	479.2	0.282	0.676	0.496	0.537
	(60.6)	(60.7)		(61.7)	(60.1)	(60.0)				
Human Sciences	545.6	553.8	**0.048**	555.9	550.5	555.0	0.256	0.682	0.519	0.502
	(76.7)	(74.4)		(74.3)	(74.7)	(74.1)				
Other covariates										
Number of employees	46.7	73.3	0.001	73.8	75.7	70.3	0.510	0.538	0.233	0.489
	(24.0)	(26.1)		(21.8)	(30.6)	(26.3)				
Number of classrooms	9.4	14.0	**<0.001**	13.7	13.8	14.5	0.830	0.443	0.591	0.743
	(4.1)	(4.9)		(3.8)	(4.3)	(6.5)				
Broadband Internet	0.80	0.93	**0.004**	0.92	0.91	0.96	0.794	0.544	0.447	0.712
	(0.40)	(0.25)		(0.27)	(0.29)	(0.19)				
Average age (Language teachers)	39.6	41.6	**0.001**	41.7	41.3	41.7	0.731	0.984	0.741	0.925
	(4.8)	(4.4)		(4.7)	(4.2)	(4.1)				
Share of teachers	0.02	0.04	**0.049**	0.04	0.04	0.05	0.998	0.814	0.809	0.965
holding a Masters' degree	(0.08)	(0.10)		(0.10)	(0.08)	(0.12)				
Share of blacks, "pardos" or	0.58	0.62	**0.095**	0.61	0.64	0.62	0.437	0.784	0.563	0.730
indigenous	(0.22)	(0.17)		(0.19)	(0.17)	(0.16)				
Share women	0.50	0.52	**<0.001**	0.52	0.52	0.52	0.697	0.528	0.324	0.608
	(0.04)	(0.04)		(0.04)	(0.04)	(0.04)				
Latitude	-19.9	-20.0	0.124	-20.0	-20.0	-20.0	**0.098**	0.342	0.312	0.254
	(0.8)	(0.8)		(0.8)	(0.7)	(0.7)				
Longitude	-40.7	-40.6	0.148	-40.6	-40.6	-40.6	0.240	0.517	0.526	0.500
	(0.5)	(0.5)		(0.5)	(0.4)	(0.4)				
Rural	0.19	0.01	**<0.001**	0	0.02	0	0.246	0.708	0.255	0.497
	(0.39)	(0.08)		-	(0.14)	-				
Joint test (p-value)							0.905	0.825	0.568	
Number of Observations	276	178		68	55	55				

Notes: Mean and standard deviation (in parentheses) of school-level characteristics in the rows are presented in columns (1), (2), (4), (5) and (6). Statistics are for all state schools in Espírito Santo with at least one high school classroom in column (1), for the experimental sample of schools in column (2), for the control group of schools in column (3), for the standard program in column (4) and for the alternative treatment in column (5). Column (3) presents p-values for t-tests comparing the groups in columns (1) and (2). In columns (7)-(10), we present $p-$values for $t-$ tests comparing the groups of experimental schools, indicated in the header of each column. The $p-$values are from regressions with strata fixed-effects, using standard errors clustered at the school level in the first two groups of variables and using robust standard errors in the third group of variables. The number of schools in the experimental sample is 178 but there were 6 (2) schools in the experimental sample that did not have students applying for ENEM in 2017 (2018), and 4 schools that opened in 2019 (the number of observations thus varies across groups of covariates). The written essay in 2017 has a superscript s because this variable was used for stratification. We did not use the 2018 data for stratification purposes because the public microdata was still not available when we performed the randomization. We also present p-values from a joint test that all covariates are balanced in each comparison. These p-values are constructed based on equation (7) from Young (2018), taking into account the randomization protocol. All p-values lower than 0.10 are in bold.

Table A.2: Design Validity — Balance Across Treatment Arms (Conditional on Non-missing in the Nonofficial ENEM essay)

	Control	Enhanc. AWE	Pure AWE	Enhan. AWE -	*p*-values (clust. strata)			
	Mean	- Control	- Control	Pure AWE	(2)=0	(3)=0	(4)=0	Obs.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Baseline Language Score	0.062	-0.064 (0.05)	0.002 (0.049)	-0.065 (0.05)	0.212	0.975	0.201	14,400
Baseline Mathematics Score	0.056	-0.043 (0.081)	-0.064 (0.059)	0.021 (0.077)	0.601	0.286	0.783	14,400
Male	0.439	-0.001 (0.01)	-0.016 (0.01)	0.015 (0.01)	0.917	0.118	0.149	14,257
Age Range								
Age up to 17	0.53	-0.008 (0.022)	0.009 (0.017)	-0.018 (0.017)	0.718	0.576	0.301	14,123
Aged 18	0.37	-0.01 (0.021)	-0.015 (0.018)	0.005 (0.014)	0.635	0.412	0.742	14,123
Aged 19	0.081	0.016 (0.005)	0.002 (0.007)	0.014 (0.009)	0.007	0.797	0.12	14,123
Aged 20 +	0.019	0.003 (0.004)	0.004 (0.004)	-0.001 (0.005)	0.433	0.340	0.857	14,123
Mothers' Education								
Mother has incomplete primary	0.075	0.013 (0.008)	0.015 (0.008)	-0.002 (0.01)	0.129	0.076	0.853	14,230
Mother has complete primary	0.196	0.011 (0.011)	0.012 (0.016)	0 (0.015)	0.318	0.481	0.977	14,230
Mother HS dropout	0.15	0.005 (0.01)	0.003 (0.009)	0.002 (0.007)	0.613	0.733	0.765	14,230
Mother completed HS	0.357	-0.022 (0.016)	-0.031 (0.017)	0.009 (0.02)	0.188	0.084	0.649	14,230
Mother completed PSE	0.092	-0.01 (0.008)	0.008 (0.011)	-0.018 (0.008)	0.257	0.451	0.041	14,230
Mother has more than PSE	0.056	-0.005 (0.006)	-0.005 (0.005)	0 (0.007)	0.406	0.311	0.986	14,230

(*cont.*)

	Control Mean	Enhanc. AWE - Control	Pure AWE - Control	Enhan. AWE - Pure AWE	*p*-values (clust. strata) (2)=0	(3)=0	(4)=0	Obs.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Fathers' Education								
Father has incomplete primary	0.092	0.006	0.006	0	0.435	0.495	0.971	14,217
		(0.007)	(0.009)	(0.01)				
Father has complete primary	0.203	0.014	0.015	-0.001	0.339	0.347	0.964	14,217
		(0.014)	(0.015)	(0.016)				
Father HS dropout	0.145	-0.005	-0.001	-0.004	0.455	0.9	0.575	14,217
		(0.007)	(0.008)	(0.008)				
Father completed HS	0.294	-0.005	-0.012	0.008	0.7	0.46	0.673	14,217
		(0.012)	(0.016)	(0.018)				
Father completed PSE	0.063	-0.009	-0.003	-0.006	0.266	0.747	0.371	14,217
		(0.008)	(0.008)	(0.007)				
Father has more than PSE	0.023	-0.005	-0.001	-0.005	0.12	0.796	0.139	14,217
		(0.003)	(0.003)	(0.003)				
p-value of joint test					0.746	0.861	0.470	

Notes: This table investigates balance with and student-level variables across experiment arms for students that participated in the nonofficial ENEM essay. Column 1 presents means and standard deviations (in brackets) of variables listed in rows. Columns 2 and 3 present estimates and standard errors clustered at the strata-level computed using specification (3), an ordinary least squares regression with indicators for each of the two experiment arms and strata indicators. We use these estimates to test for differences between the enhanced AWE and the pure AWE treatments and report the point estimate and standard error in column 4. Column 5 presents the number of observations used for inference tests for each variable in columns 5 to 8 [BACK TO TEXT].

Table A.3: Design Validity — Balance Across Treatment Arms (Conditional on Complying with Official ENEM essay)

	Control Mean	Enhanc. AWE - Control	Pure AWE - Control	Enhan. AWE - Pure AWE	*p*-values (clust. strata) (2)=0	(3)=0	(4)=0	Obs.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Male	0.423	-0.007	-0.022	0.015	0.444	0.078	0.239	14,268
		(0.009)	(0.012)	(0.013)				
Age Range								
Age up to 17	0.374	-0.011	0.01	-0.021	0.641	0.653	0.242	14,268
		(0.024)	(0.021)	(0.017)				
Aged 18	0.45	-0.009	-0.028	0.019	0.658	0.119	0.232	14,268
		(0.019)	(0.017)	(0.016)				
Aged 19	0.127	0.008	-0.001	0.009	0.401	0.91	0.325	14,268
		(0.009)	(0.011)	(0.009)				
Aged 20 +	0.049	0.012	0.02	-0.008	0.18	0.064	0.458	14,268
		(0.009)	(0.01)	(0.01)				
Mothers' Education								
Mother has incomplete primary	0.195	0.011	0.008	0.003	0.431	0.59	0.865	14,268
		(0.013)	(0.014)	(0.017)				
Mother has complete primary	0.177	0.007	0.005	0.002	0.548	0.673	0.886	14,268
		(0.011)	(0.012)	(0.013)				
Mother HS dropout	0.151	-0.003	-0.008	0.005	0.675	0.394	0.523	14,268
		(0.007)	(0.009)	(0.008)				
Mother completed HS	0.309	-0.011	-0.012	0	0.306	0.445	0.993	14,268
		(0.011)	(0.015)	(0.015)				
Mother completed PSE	0.044	-0.006	0.001	-0.008	0.246	0.818	0.065	14,268
		(0.005)	(0.006)	(0.004)				
Mother has more than PSE	0.025	-0.005	-0.001	-0.004	0.045	0.657	0.262	14,268
		(0.002)	(0.003)	(0.003)				

Notes: (*cont.*)

	Control Mean	Enhanc. AWE - Control	Pure AWE - Control	Enhan. AWE - Pure AWE	p-values (clust. strata) (2)=0	(3)=0	(4)=0	Obs.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Fathers' Education								
Father has incomplete primary	0.14	0.024	0.007	0.016	0.054	0.588	0.266	14,268
		(0.012)	(0.014)	(0.014)				
Father has complete primary	0.155	0.014	0.016	-0.002	0.11	0.166	0.843	14,268
		(0.008)	(0.011)	(0.012)				
Father HS dropout	0.17	-0.002	-0.003	0.001	0.775	0.739	0.888	14,268
		(0.008)	(0.01)	(0.006)				
Father completed HS	0.363	-0.024	-0.009	-0.015	0.088	0.62	0.429	14,268
		(0.014)	(0.018)	(0.019)				
Father completed PSE	0.066	-0.007	-0.01	0.003	0.322	0.16	0.639	14,268
		(0.007)	(0.007)	(0.006)				
Father has more than PSE	0.071	-0.004	0.003	-0.007	0.57	0.638	0.332	14,268
		(0.006)	(0.007)	(0.007)				
HH Income Category								
Category 1	0.239	0.005	-0.009	0.014	0.767	0.493	0.463	14268
		(0.018)	(0.013)	(0.019)				
Category 2	0.31	0.016	0.015	0.001	0.151	0.252	0.971	14268
		(0.011)	(0.013)	(0.015)				
Category 3	0.239	-0.004	0.01	-0.014	0.686	0.400	0.301	14268
		(0.009)	(0.012)	(0.013)				
Category 4	0.212	-0.017	-0.016	-0.001	0.274	0.257	0.952	14268
		(0.016)	(0.014)	(0.014)				

Notes: (*cont.*)

	Control Mean	Enhanc. AWE - Control	Pure AWE - Control	Enhan. AWE - Pure AWE	*p*-values (clust. strata) (2)=0	(3)=0	(4)=0	Obs.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Mothers' Occupation								
Low Skill	0.695	0.025	0.011	0.014	0.111	0.453	0.391	14,268
		(0.015)	(0.015)	(0.016)				
Medium Skill 15	0.158	-0.024	0.002	-0.026	0.023	0.910	0.047	14,268
		(0.01)	(0.014)	(0.013)				
High Skill	0.019	-0.004	0.001	-0.004	0.145	0.829	0.092	14,268
		(0.003)	(0.002)	(0.003)				
Fathers' Occupation								
Low Skill	0.712	0.01	0.008	0.002	0.316	0.57	0.864	14,268
		(0.01)	(0.013)	(0.014)				
Medium Skill 15	0.186	-0.01	-0.002	-0.008	0.301	0.891	0.495	14,268
		(0.009)	(0.012)	(0.012)				
High Skill	0.011	-0.004	-0.002	-0.002	0.067	0.267	0.452	14,268
		(0.002)	(0.002)	(0.002)				
p-value of joint test					0.910	0.161	0.197	

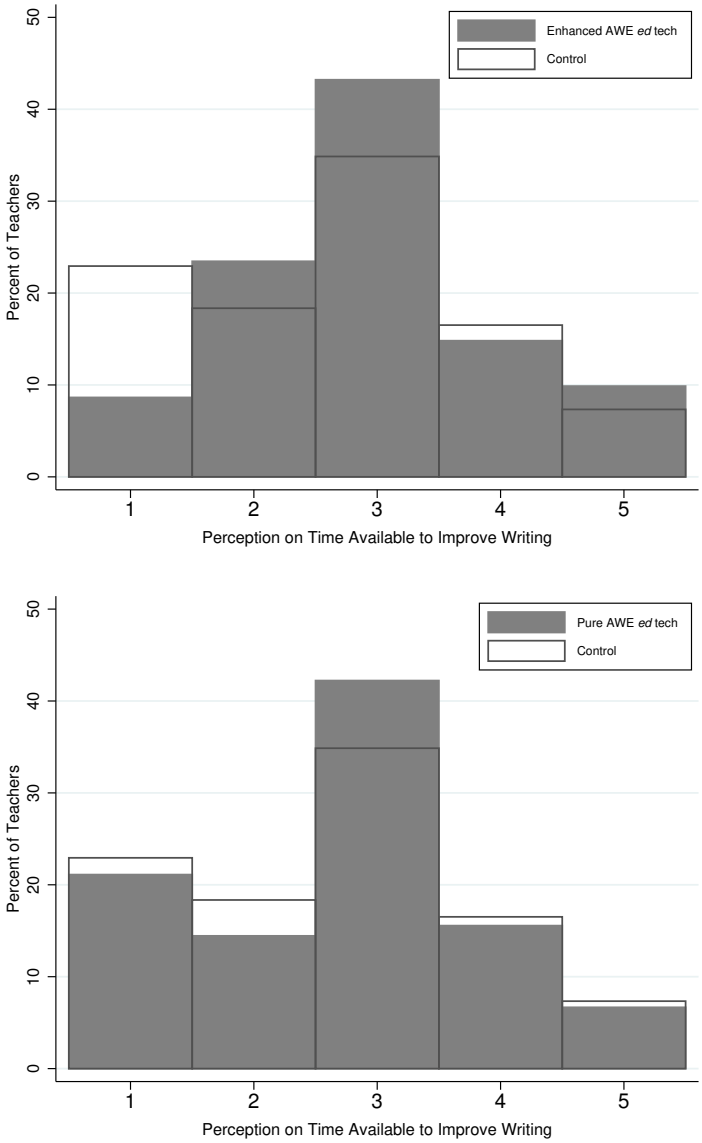
Notes: This table investigates balance with and student-level variables across experiment arms for students that participated in the official 2019 ENEM essay. Column 1 presents means and standard deviations (in brackets) of variables listed in rows. Columns 2 and 3 present estimates and standard errors clustered at the strata-level computed using specification (3), an ordinary least squares regression with indicators for each of the two experiment arms and strata indicators. We use these estimates to test for differences between the enhanced AWE and the pure AWE treatments and report the point estimate and standard error in column 4. Column 8 presents the number of observations used for inference tests for each variable in columns 5 to 8. [BACK TO TEXT]

Table A.4: Treatments, Collective Training and Feedback (Teacher Survey)

	Teachers — Assignments, Grades and Collective Feedback					
	# ENEM Essays...					
Dep. Var.:	Assign. in class	Assign.	Graded good	Discuss. bad	Discuss. Index	Summary
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A. Main Effects — Specification (3)					
$\widehat{\tau^{\text{Enhanced AWE}}_{\text{ITT}}}$	-0.13	0.19	0.51	0.76	0.41	0.16σ
p-value, clust. strata	0.526	0.442	0.351	0.322	0.372	0.282
[p-value, rand. inf.]	0.493	0.419	0.307	0.267	0.312	0.208
{p-value, MHT adj.}	0.999	0.999	0.999	0.999	0.999	{0.208}
$\widehat{\tau^{\text{Pure AWE}}_{\text{ITT}}}$	-0.44	1.13	0.30	0.45	0.58	0.25σ
p-value, clust. strata	0.607	0.213	0.425	0.380	0.325	0.191
p-value, rand. inf.	0.630	0.213	0.448	0.384	0.350	0.192
p-value, MHT adj.	0.630	0.999	0.999	0.999	0.999	0.3844
	Panel B. Differential Effects — Specification (3)					
$\widehat{\tau^{\text{Enhanced AWE}}_{\text{ITT}} - \tau^{\text{Pure AWE}}_{\text{ITT}}}$	0.31	−0.94	0.21	0.31	−0.16	-0.09σ
(p-value, clust. strata)	0.870	0.502	0.889	0.832	0.897	0.768
[p-value, rand. inf.]	0.797	0.352	0.863	0.782	0.869	0.715
{p-value, MHT adj.}	0.999	0.999	0.999	0.999	0.869	—
MHT Adjustment	Holm	Holm	Holm	Holm	Holm	Holm
Control Group Mean	16.5	10.3	13.2	9.9	11.5	—
Control Group SD	12.2	9.4	10.8	7.6	9.7	
						—
$N_{Teachers}$	271	270	269	260	267	259

Notes: This table presents estimates and inference tests for the average absolute and differential average treatment effects on the: amount of writing that teachers assigned to students to train for the ENEM essay during the year and on the amount of collective feedback that they gave back to students. All outcomes in columns (1) to (5) were initially elicited as an open-ended question on the number of essays and then winsorized at the top 1%, as specified in the pre-analysis plan. The unit of observation in columns (1) to (6) is a teacher that participated in our end-line survey and provided an answer for the question in each column. Estimates in both panels are from specification (3), an ordinary least squares regression with indicators for each of the two experiment arms and strata dummies. We present three upper one-sided p-values below each coefficient in Panel A and three two-sided p-values below each coefficient in Panel B: p-values obtained using the standard errors clustered at the at the strata level, in parentheses; randomization inference p-values using the randomization protocol and 1,000 draws of the assignment with replacement, in brackets; and Holm (1979) adjusted p-values using the latter, in curly brackets. These adjustments were made within the cells that have the same shaded background. The summary index in columns (6) is computed based on the procedure suggested by Anderson (2008) and relies on constructing a positively weighted mean of the standardized outcomes of the observations with non-missing outcomes in columns (1)-(5). [BACK TO TEXT]

Figure A.2: Treatments and the Distribution of Perceptions on Time



[BACK TO TEXT]

1. This figure plots the distribution of the index on the subjective perceptions of teachers on how constrained they feel to improve their students' abilities on writing using the time available inside and outside the classroom. The upper figure compares teachers using the enhanced AWE *ed* tech and the control group, whereas the bottom figure compares teachers using the pure AWE *ed* tech and the control group.

2. The possible answers in this question followed a 5-point Likert scale and ranged from "Time very insufficient" to "Time more than sufficient".

Table A.5: Lee Bounds on Teachers' Significant Treatment Effects

	ITT	Lee Bounds	
	Coeff.	Lower	Upper
	(1)	(2)	(3)
Time Available For... (Table 5)			
Writing	0.30	0.10	0.52
(*s.e*, clust. strata)	(0.17)	(0.23)	(0.19)
Grammar	0.36	0.05	0.49
(*s.e*, clust. strata)	(0.18)	(0.20)	(0.20)
Literature	0.32	0.08	0.50
(*s.e*, clust. strata)	(0.20)	(0.20)	(0.20)
Summary Index	0.26σ	-0.17σ	0.44σ
(*s.e*, clust. strata)	(0.15)	(0.14)	(0.15)
Average Hours Worked Weekly... (Table 6 column 1)			
Outside School	-1.19	-1.33	1.03
(*s.e*, clust. strata)	(0.98)	(0.89)	(0.86)

Notes: This table depicts the original coefficients and standard errors from Tables 5 and 6 (column 1), lower (column 2) and upper (column 3) Lee (2009) bounds on significant coefficients arising from the analysis of the absolute effects enhanced AWE *ed* tech using equation (1). The dependent variable is listed in the rows of the table. The specification used to compute the bounds does not include strata fixed effects nor controls. Bootstrapped standard errors using 500 replications are in parentheses for columns 2 and 3. [BACK TO TEXT]

Table A.6: Heterogeneity in Treatment Effects

Heterog. Margin:	Gender		Socio-economic (HH Income) Status		Race	
	Boys	Girls	Below Median	Above Median	White or Asian	Non-White nor Asian
	(1)	(2)	(3)	(4)	(5)	(6)
$\widehat{\tau^{\text{Enhanced AWE}}_{\text{ITT}}}$	0.073σ	0.114σ	0.098σ	0.087σ	0.145σ	0.096σ
(*s.e.*, clust. strata)	(0.033)	(0.038)	(0.035)	(0.037)	(0.035)	(0.042)
p-value diff., clust. strata	0.387		0.568		0.245	
p-value diff., MHT adj.	0.999		0.999		0.999	
$\widehat{\tau^{\text{Pure AWE}}_{\text{ITT}}}$	0.079σ	0.108σ	0.143σ	0.097σ	0.089σ	0.087σ
(*s.e.*, clust. strata)	(0.034)	(0.035)	(0.035)	(0.037)	(0.035)	(0.042)
p-value diff., clust. strata	0.631		0.324		0.865	
p-value diff., MHT adj.	0.999		0.999		0.999	

Heterog. Margin:	Shift		Quartiles of Baseline Language Achievement			
	Full Shift	Non-Full Shift	Quartile 1	Quartile 2	Quartile 3	Quartile 4
	(7)	(8)	(9)	(10)	(11)	(12)
$\widehat{\tau^{\text{Enhanced AWE}}_{\text{ITT}}}$	0.298σ	0.086σ	0.044σ	0.154σ	0.119σ	0.145σ
(*s.e.*, clust. strata)	(0.171)	(0.046)	(0.052)	(0.060)	(0.056)	(0.066)
p-value diff., clust. strata	0.175		0.158			
p-value diff., MHT adj.	0.999		0.999			
$\widehat{\tau^{\text{Pure AWE}}_{\text{ITT}}}$	0.300σ	0.080σ	0.031σ	0.120σ	0.123σ	0.101σ
(*s.e.*, clust. strata)	(0.089)	(0.037)	(0.041)	(0.049)	(0.048)	(0.053)
p-value diff., clust. strata	0.999		0.201			
p-value diff., MHT adj.	0.999		0.999			

Heterog. Margin:	Number of Classes Taught (Teacher)	
	Above Median	Below Median
	(13)	(14)
$\widehat{\tau^{\text{Enhanced AWE}}_{\text{ITT}}}$	0.088σ	0.116σ
(*s.e.*, clust. strata)	(0.058)	(0.058)
p-value diff., clust. strata	0.715	
p-value diff., MHT adj.	0.999	
$\widehat{\tau^{\text{Pure AWE}}_{\text{ITT}}}$	0.182σ	0.060σ
(*s.e.*, clust. strata)	(0.057)	(0.042)
p-value diff., clust. strata	0.099	
p-value diff., MHT adj.	0.999	

Notes: This table presents estimates and inference tests for the average absolute and differential treatment effects on sub-samples singled out in the pre-analysis plan. Estimates are from specification (1), an ordinary least squares regression with indicators for each of the two experiment arms, strata dummies, ans the controls listed in the footnotes to table 3. We present standard errors clustered at the strata level in parentheses and two-sided *p*-values comparing whether the effects are equal in the sub-samples: *p*-values obtained using the standard errors clustered at the strata level; and Holm (1979) adjusted *p*-values using the latter. The multiple hypothesis testing adjustments were made within the cells that have the same shaded background. [BACK TO TEXT]