Pre-Analysis Plan: Experimental Evaluation of the Effects of Automated Writing Evaluation on Students and Teachers*

Bruno Ferman Lycia Lima Flavio Riva

Last Version: December 9th, 2019. *First Version*: August 28th, 2019. Ferman: Sao Paulo School of Economics - FGV, bruno.ferman@fgv.com, *corresponding author*; Lima: Sao Paulo School of Business Administration - FGV, lycialima@gmail.com.; Riva: Sao Paulo School of Business Administration - FGV, flaviorussoriva@gmail.com. This research underwent ethics review by the Committee on the Use of Humans as Experimental Subjects (COUHES, Protocol #18115953228) at MIT and the ethics' committee at Fundação Getúlio Vargas. This file was uploaded at the American Economic Association Social Science Registry (RCT ID AEARCTR-0003729), after the randomization of schools into the treatment arms and before the researchers had access to data on implementation or administrative data available from the Brazilian Ministry of Education. The collection of primary data and the correction of the written essays will be funded by the Post-Primary Education Initiative from J-PAL (GR-0938). We also thank João Pugliese for detailed comments on first drafts. The authors declare that they have no relevant or material financial interests that relate to the research described in this plan.

1 Introduction

This document describes a clustered randomized experiment with public schools in the state of Espírito Santo (Brazil) designed to study the effects and channels of impact of AWE-based programs. Language teachers and high school senior students in 110 treated schools will participate in one of two alternative programs based on an NLP algorithm embedded in an online platform, and 68 schools will compose the control group — yielding a total of around 300 teachers and 20,000 students. Both programs will span the 2019 academic year and consist of five writing practices elaborated by the implementer to support the development of writing skills and to improve the scores of students in the argumentative essay of the National Secondary Education Exam (ENEM). ENEM is the largest college admission exam in Brazil and the second largest college admission exam in the world, only falling behind the Chinese *gāokǎo*. In 2014, the year when it reached the maximum number of applicants since its creation in 1998, 6.2 million people took the exam.

The *first* treatment arm (standard treatment) uses the algorithm (described by Fonseca et al., 2018) to provide students with instantaneous feedback on syntactic text features — such as spelling mistakes — and with a noisy signal of student achievement, a performance bar with 5 levels. About three days after submitting their essays on the program's platform, students receive a final grading elaborated by human graders hired by the implementer, who correct the essays trying to mimic the real-world exam. This grading includes the final ENEM essay grade on a 1000-points scale, comments on the skills valued in the exam and a general comment on essay quality. In the *second* treatment arm (alternative treatment), the whole experience with the writing task is completed at once, and is based only on interactions with the artificial intelligence: after submitting the essays, students receive the instantaneous feedback on text features and the noisy signal of achievement (as in the first treatment arm), but are also presented to the AWE-predicted grade on a 1000-points scale and to comments selected in the implementers' database among a list of specific comments suited for a skill score. In both treatment arms, the essays and the aggregate and individual grading information generated throughout the year — by the artificial intelligence supervised by human graders in the standard treatment, and only by the artificial intelligence in the alternative treatment — are presented to teachers on a personal dashboard.

The primary goal of the experiment is to document the impacts of the programs on ENEM essay scores. This is a relevant outcome for students, as ENEM is used for admission purposes by many post-secondary institutions in Brazil. Differences in the essay score account for the largest share of the public-private achievement gap in the exam (see details in Section 2). Thus, in principle, both programs could be considered as viable alternatives to make public school students more competitive for admission into (better) post-secondary institutions. We consider the estimation of the main effects of the standard treatment as an efficacy trial of an intervention developed to overcome relevant bottlenecks of effective writing pedagogy. The main idea of the program is that a combination of additional inputs from artificial and human intelligence can help overcome important time constraints teachers may face to support the development of writing skills. Such time constraints are arguably particularly tight for Portuguese public school teachers, because grading and

providing feedback on essays is a time-intensive nonroutine task, and because Portuguese teachers in public schools usually have to divide their time to teach Grammar, Literature, and Writing. Notice, however, that — in contrast to the alternative treatment — scaling up an intervention like the standard program would necessarily entail large marginal costs.[1] Therefore, estimating the effects of the alternative program will inform about the potential of a relatively easily scalable program to improve students' writing skills, which is important from a policy perspective.[2] Finally, the differential effects between the two programs will be informative about the state of art of artificial intelligence — in particular, NLP algorithms —, and its potential to emulate human intelligence without supervision. The ENEM essay is an interesting setting to study this issue, since the essay values both low-level abilities (such as the command over grammar rules) and high-level abilities (such as global coherence). Therefore, a closer look on how the treatments differentially affect different types of abilities will bring valuable information on the current potentials and limitations of artificial intelligence. This differential effect will also be informative about a relevant trade-off between feedback from artificial versus human intelligence, where in the second one students receive a delayed, but arguably better feedback on their essays.

In order to provide a rich understanding of the channels of impact of both interventions and explain potential differences in treatment effects, we will collect primary data on teachers and students at the end of the year. On the *students*' side, we hypothesized that AWE can provide additional opportunities of practice and more individually-customized feedback perceived as high-quality. Additionally, students may be more motivated to study for the essay if they believe they have better prospects in this exam. On the *teachers*' side, we hypothesized that AWE could relieve teachers of some of the burden of responding to students' writing tasks and affect their time allocation across tasks with different complexity. In particular, the interventions may provide more time for teachers to provide individualized counseling taking into account the students' heterogeneity or increase the amount of time preparing classes.

We are also interested in the effects of both programs on the following secondary outcomes: general writing skills and learning in other subjects. While the programs were not designed to directly affect these outcomes, we believe they can have important indirect effects on them. For both groups of outcomes, the effects are *a priori* ambiguous. We start with the discussion of general writing skills. It is possible that training more for the ENEM essay induces spill-overs across writing genres, in particular in skills that are not genre-specific. However, training intensely could potentially reduce performance in writing tasks unrelated to the test. In this sense, observing a reduction in the quality of other writing tasks would indicate that "training for the test" can be detrimental to more general writing skills. We will assess this possibility through an essay in

[1] We are collecting cost data from the implementer using the J-PAL costing template.

[2] The marginal cost of scaling such programs is low conditional on schools having a computer lab and a working Internet connection. Importantly, one of the advantages of the technology we describe is that it is based on an online platform that works very well with poor internet connections, which are still very common in some schools in Brazil. In 2017, according to the Brazilian Educational Census, 14,891 public schools (out of 18,407) with at least one class of high-school seniors had a computer lab and Internet connection, totaling 1,496,987 (out of 1,777,480) enrolled senior students. Therefore, we consider that the program in the alternative treatment has the potential of being scaled up to more than one million public school students per year at a relatively low marginal cost.

which students will write using a different structure from the ENEM essay. We will also consider the effects of the program on learning in other subjects. On the one hand, the program may crowd out time and effort from other training activities to writing. On the other hand, improvements in writing skills may be complementary to other subjects (like reading). Students' achievement in other subjects may also be positively impacted if students feel more motivated to take the ENEM exam because of the treatment. Finally, for language (non-writing) skills, the program can also positively affect students' scores if it allows Portuguese teachers to better allocate their time for teaching the other subjects as well.

2 Background Information

All educational levels in Brazil can be completed either in private or in public institutions and one of the main features of the Brazilian school market is that public and private schools are very different in terms of quality indicators.[3] Repetition and evasion rates are substantially higher among public school students (Costa, 2013), who also perform considerably worse than private school students in tests that are used for admission at the post-secondary level. Not surprisingly, this large gap is present in the National Secondary Education Exam ("Exame Nacional do Ensino Médio", ENEM), a non-compulsory standardized national high-stakes test in Brazil that is a key determinant of access to higher education in the country. In what follows, we provide background information on the exam and its written essay and describe the public-private achievement gap in the exam.

Since its creation in 1998, the exam is implemented by an autarchy linked to the Brazilian Ministry of Education, the National Institute for Research on Education ("Instituto Nacional de Estudos e Pesquisas Educacionais", INEP). The tests take place each year in October, November or December in two consecutive Sundays and are 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 five grades are used by post-secondary institutions in Brazil and Portugal for admission purposes, using weights chosen independently. The leftmost bars in Figure 1 describe the private school premium among high school seniors in the ENEM 2018, in Brazil (Panel A) and Espírito Santo (Panel B), for each test in the exam and in the written essay. Although the achievement gap in ENEM is a significant feature of all tests, it is remarkably larger in the written essay (at 40-50%) when compared with the multiple-choice tests (at 15-25%). When compared to the multiple-choice Portuguese Language test, which is supposed to measure the other dimension of literacy, the difference between public and private school students in the essay in Espírito Santo is approximately 35 percentage points larger or more than 200%. All in all, this shows that differences in the essay score account for the largest share of the public-private

[3] This partially reflects a segmentation based on student's income. Students with lower family income are predominantly in public schools, which accounted for 86-90% of school enrollment from 2000 to 2015. For instance, in 2014, only 2%, 3% and 4% of high school students in families in the first, second and third income quintile were in private schools, respectively (Almeida et al., 2017).

achievement gap in the exam.

A typical ENEM essay has 25-30 lines and is supposed to conform to the argumentative textual genre. Topics are always introduced by excerpts, graphs, figures or cartoons that frame a social problem using different points of view.[4] Valid essays typically begin with an introductory paragraph, followed by two paragraphs where arguments are developed and a final paragraph featuring an "intervention proposal".[5] The written essays are corrected by graders hired by INEP, who evaluate five specific skills, if an essay is considered valid. Each skill is valued on a 200 points scale with intervals of 40, so that the full grade ranges from 0 to 1000. These skills are described (INEP/MEC, 2018, p. 8) as the ability to:

- adhere to the formal written norm of Portuguese;
- use argumentative linguistic structures;
- conform to the proposed essay topic (prompt), and develop a text using knowledge from different areas;
- select, relate, organize and interpret data and arguments in defense of a point of view;
- elaborate a policy proposal that could contribute to solve the problem in question, respecting the basic human rights' principles.

Table 1 presents the correlations between ENEM grades in all skills for high school seniors present in the day of the ENEM written essay in 2018. The correlation between grades in the first two skills (the use of formal written norm and argumentative linguistic structures) 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. Interestingly, the correlation between the last skill and the other skills — the consistency of the intervention proposal — is not as high (specially for the first two), suggesting that this ability is based on other aspects of the text, and, in particular, on how the policy proposal fits with the text development and the construction of the point of view. The rightmost bars in Figure 1 describe the private school grade premium in each skill. Notably, the gap is increasing in the complexity or sophistication of the skill, reaching 80-100% for the policy proposal.

[4]Since its creation, ENEM has proposed several polemic topics, which attracted great attention from the media: religious intolerance (2016), gender-based violence and its persistence in Brazil (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) child labor in Brazil (2005) and citizenship and social participation (1999).

[5]Even if a test taker manifests interest in having an essay grade (i.e., writes something on the essay sheet), the invalidation (and annulment) of the essay still can occur in some special cases: explicit copy of the motivating prompts or other parts of the Language test, problems with genre and/or subject accordance, insufficient text production for grading purposes (< 7 lines) and the presence of "disconnected" parts in the essay.

3 Treatment Arms

The implementer was created in 2015 and its main goal is to improve writing through the use of artificial intelligence and its applications in linguistics. Its main current product is a pedagogical program based on an online platform that corrects and provides feedback on ENEM written essays using an AWE algorithm (Fonseca et al., 2018) supervised by independently hired human graders. This section describes the standard pedagogical program and an alternative treatment that will help explore the potentials and limitations of relying only on the artificial intelligence as an additional pedagogical resource.

3.1 Standard Program

The program spans the academic year of high school seniors and consists of 5 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. Teachers are also absolutely free to adapt their grading on Portuguese Language to the outputs of the platform. Even though access to the platform contents can be done independently by students outside the school environment, the implementer tries to provide teachers with an instrument to support the development of writing skills *inside* the classroom.

3.1.1 Students' Interface.

After writing and submitting the essay, the online platform interacts instantaneously with the student providing a comprehensive set of text features, presented as descriptive statistics used to compare the essay to "goals" that would bring the student closer to achieve a perfect score.[6] At this point, the student is also presented to a noisy signal of his/her achievement, displayed in a performance bar with 5 levels. This indicator is based on the AWE predicted essay score, but the actual score is not shown to the student. The program withholds the AWE algorithm predicted grade to avoid introducing noise in the evaluation process. The final score is set by human graders independently hired on a task-based contract that pays 3.50 Reais (approximately US$0.85) *per* essay. The graders have access to the essay with all the text features used to predict the grade on the formal written norm skill. For the other skills, the human graders choose a value, ranging from 0 to 200 (in 40 point intervals), without any aid from the algorithm outputs. When the human graders choose a grade 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 can leave a general comment on the essay, the final step before submitting the annotated correction to the student. This process

[6] Some examples are: number of words, number of spelling mistakes and uses of informality tones, intervention elements and social agent markers.

takes, on average three business days. Students receive a text message when their final correction is available.

3.1.2 Teachers' Interface.

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 logged in, started writing and finished the task. Each teacher also has access to a personal dashboard with shortcuts to Excel files containing the aggregate data on enrolled students and their individual grades on the essay and skills for a given activity.

3.2 Alternative Treatment

The user experience in this treatment arm focuses on the instantaneous outputs from the AWE algorithm to explore the possibility that a pedagogical program could be based on information that is only generated by the artificial intelligence. The user interface is very similar to the one in the standard 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.

4 Sample and Randomization Procedure

In March 2019, we received from the State's Education Department ("Secretaria de Estado da Educação", SEDU/ES) a list of public schools in Espírito Santo that were selected to participate in the experiment. The selection of these schools was based on a *survey* on proneness to online technology adaptation conducted by the IT department of SEDU/ES.[7] The final number of treated units in the first arm was chosen based on constraints in the implementer capacity of providing the standard program to more than 55 schools in 2019. The randomization used the following strata: (i) a geographical criterion, given by the 11 regional administrative units in the state; (ii) the average grade in the ENEM 2017 essay;[8] (iii) participation on an implementation pilot in 2018.[9] The whole process led to a total of study sample size of 178 schools divided in 33 strata (of sizes 2 to 8), with 55 units assigned to the standard program, 55 units assigned to the alternative treatment, and 68 schools assigned to the control group.

[7]The sample of experimental schools received 8,000 notebooks distributed by SEDU/ES between February and April, so we do not expect the availability of computers to be a first order concern for implementation purposes.

[8]We used the median or quartiles of the average grade to split schools within a regional administrative unit, according to the original size of the group. We generated an independent stratum for the 6 schools that had no students taking the ENEM test in 2017.

[9]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 artificial intelligence 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.

5 Hypotheses and Data

This section presents our measures of engagement and the hypotheses on (primary and secondary) outcomes and potential mechanisms through which the treatments could impact the behavior of students and teachers. The outcomes and mechanisms are listed in Table 2, where we also present the data that will be used to operationalize our hypotheses and our priors on the final results of the study.

5.1 Implementation and Engagement

A necessary condition for interventions like the ones described in Section 3 to positively affect outcomes is that the software is broadly integrated in instruction. We will present descriptives on:

- the proportion of teachers that use the platform to assign essays for each writing activity proposed by the implementer;
- the proportion of students that submit essays through the platform for each writing activity proposed by the implementer;
- the final number of essays written and submitted by students throughout the year;
- the proportion of students in the standard program treatment arm that come back for the comments of human graders after they arrive.

This information is available from the implementer and the differences in behavior of engagement measures over time may be relevant to understand potentially different effects between the two treatments.

5.2 Primary Outcome: ENEM Test Scores

5.2.1 Main Hypotheses

Our primary outcome is given by ENEM essay scores. This choice was based on two main reasons. First, improving the prospects of students in this essay is the main goal of the implementer. Second, the ENEM score is an important outcome for students, as it is a key mediator of access into a large number of post-secondary education institutions, and the essay is responsible for the greatest public-private gap achievement in this exam. Our alternative hypotheses on the main and differential impacts are:

A.1. *The standard treatment will have an effect different from zero on ENEM essay scores.*

A.2. *The alternative treatment will have an effect different from zero on ENEM essay scores.*

A.3. *The differential effect of the standard treatment relative to the alternative treatment is different from zero.*

Overall, we believe that the two programs will have a positive impact on students' ENEM essay scores if the treatment induces students to write more training ENEM essays, receive more feedback on their writing, receive better quality feedback on their writing, and allows teachers to reallocate their time to other relevant tasks to the students. However, we cannot rule out *a priori* that the treatment may have negative impacts on the students. This may happen if, for example, the training for the ENEM essays that teachers would provide to students in the absence of the program would have been better than the training they receive in the program. We believe this would not be the case, based on conversations with teachers who report strong time constraints for preparing their students for this essay.

We do not have strong priors about the differential effect between the two programs. On the one hand, we believe the standard treatment should provide better quality feedback for students. This should be particularly relevant for the skill of writing a consistent policy proposal that concludes the essay. Such higher quality feedback can also affect the engagement in platform, implying that students in the standard treatment may have not only better quality, but also more feedback. However, such effects should only materialize if students actually go back to the platform to read this more detailed feedback a couple of days after they submitted the essay. In contrast, the feedback from the alternative program is immediate. Therefore, the comparison between the two treatment arms depends crucially on this trade-off of better quality versus timing of the feedback. We present more schematically the possible channels we anticipate are relevant for these programs, and discuss how we are going to test each of those channels in Section 5.3.

5.2.2 Data and Outcomes

We will have administrative data on the 2019 ENEM test scores (out of 1000) for the students in the experimental sample of schools. In addition to that, we also partnered with the SEDU/ES to administer an essay that follows the same textual genre and grading criteria of ENEM. Such essay will be an additional part of the standardized state exam that students take in Espírito Santo, and will be presented to teachers and students as an initiative of the SEDU/ES, not related to the programs we are evaluating. These essays will be graded by one of Brazil's leading education testing firms, "Centro de Políticas Públicas e Avaliação da Educação" (CAEd/UFJF).

The decision to collect these data was based on the following reasons. First, due to recent changes in the autarchy that is in charge of the ENEM exam, we believe that there is a small, but positive, probability that the microdata from the exam will not be available for research purposes. Second, we think it is important to have control over the theme of the essays, to guarantee that (by chance) students in the treatment group are not benefited by writing an essay on a topic that they had just trained in one of their program-related writing practices. Third, we can include better individual-level controls in regressions using our collected measures as outcomes, because in this case we can match students' outcomes in our writing tests with information on their proficiency before the treatments. A final potential concern is that the treatments could affect the enrollment rates in the ENEM exam. We don't believe this will be a crucial problem because all high school

seniors must take ENEM to graduate in the state we are implementing the program. Nevertheless, we discuss the potential problems of differential attrition and the adjustments we will use if this is a problem in Section 6.4.1.

In addition to the total essay score, we will also have information on the scores for each writing skill (see details on the writing skills considered in the ENEM exam in Section 2). We will aggregate the specific writing skills considered in the ENEM exam in three writing skill groups:

I. *syntactic features*, comprising the use of formal written norm and the use of argumentative linguistic structures;

II. *lower-level semantic features*, comprising the conformity to the argumentative genre and prompts and the use of data to defend a point of view;

III. *higher-level semantic features*, comprising the global consistency, creativity quality of the intervention proposal.

This grouping is based on the fact that the skills therein show increased level of sophistication and we anticipate that the artificial intelligence will be more able to provide valuable inputs to the syntactic features — which involve more routine correction procedures — when compared to the semantic ones, either lower- or higher-level.

5.3 Mechanisms

5.3.1 Hypotheses

As explained in Section 5.2, we anticipate that some mechanisms can help us understand the results from testing hypotheses **A**.1, **A**.2 and **A**.3 above. Our alternative hypotheses are that the treatments will:

M.1. *increase the number of essays written/assigned to train for the real ENEM essay, inside and outside the classroom;*

M.2. *increase the amount of feedback a student receives after writing essays;*

M.3. *improve the quality of the feedback a student receives after writing essays;*

M.4. *increase students' aspirations to enter in a post-secondary institution;*

M.5. *increase the Portuguese Language teacher expectations about his/her students' future educational attainment;*

M.6. *change the knowledge of the teachers about their students' strengths and weaknesses*;

M.7. *change the time allocation of the Portuguese Language teacher, reducing time allocated to correcting classwork, and increasing time allocated to preparing lectures, and giving individual support to students. It can also reduce the number of extra-hours worked.*

The effect described in **M**.6. is, a priori, ambiguous. On the one hand, the platform aggregates information from the students, making it easier for the teacher to digest. On the other hand, it may reduce such knowledge if teachers read and grade fewer essays from the students.

5.3.2 Data

The mechanisms listed above will be assessed through the collection of primary data. We have partnered with SEDU/ES to include multiple-choice questions in the state's standardized exam ("Programa de Avaliação da Educação Básica do Espírito Santo", PAEBES) questionnaire and will also collect these data independently through student and teacher surveys. From the student side, the variables collected will be:

- number of essays written to train for the ENEM in the last year (mechanism **M**.1);
- number of essays written to train for the ENEM in the last year, which received individualized correction and/or comments (mechanism **M**.2);
- number of essays written to train for the ENEM in the last year, which received a grade (mechanism **M**.2);
- number of essays written to train for the ENEM in the last year, which were followed by a personal discussion with the teacher after receiving a grade (mechanism **M**.2);
- perception on the usefulness of the correction and/or comments for improving specific skills for the ENEM test (mechanism **M**.3), which we will interpret as a sign of feedback quality;
- plans for 2020 (work, college, or both), which we will use to understand whether the programs shift students' aspirations towards attaining post-secondary education (mechanism **M**.4).

From the teacher side, the variables collected will be:

- number of essays assigned to train for the ENEM in the last year (mechanism **M**.1);
- number of essays assigned to train for the ENEM in the last year inside the classroom (mechanism **M**.1);
- number of essays assigned to train for the ENEM in the last year, which were individually graded (mechanism **M**.2);
- number of essays assigned to train for the ENEM in the last year, which were followed by a discussion about common mistakes (mechanism **M**.2);
- number of essays assigned to train for the ENEM in the last year, which were followed by a discussion about good essays (mechanism **M**.2);

- teachers' perceptions about the proportion of their students that will be admitted in post-secondary institutions (mechanism **M**.5);
- 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 (mechanism **M**.6);
- teachers' predicted average grade of their students in public schools in the written essay of ENEM 2019, which we will compare to the actual average grade in the exam's essay, at the school level (mechanism **M**.6).

We are also interested in documenting how the interventions affected the time allocation of teachers in general, and their allocation to nonroutine *vs.* routine tasks (mechanism **M**.7), in particular. To this end, we will ask collect the following data:

- number of hours working inside and outside the school dedicated in a typical week in 2019 (considering classes taught to senior public high school students) to:
  - teaching;
  - preparing lectures and materials for activities, including home assignments;
  - correcting essays;
  - correcting classwork and homework related to Grammar or Literature;
  - giving individual support to students (one-on-one tutoring), guiding and counseling those with academic problems or special interests;
  - meeting with the school's staff.

  We will also collect information on:
- number of hours dedicated to work outside the school environment in a typical week in 2019;
- perceptions about the availability of time during the year to improve the knowledge of students about writing, Grammar and Literature, in a scale of 1 to 5.

5.4 Secondary Outcomes

5.4.1 Hypotheses

In addition to estimating the effects on the primary outcome described in Section 5.2, we will also consider the effects of the program on a series of secondary outcomes that can be indirectly affected by the program. The alternative hypotheses on these outcomes are:

B. *The treatments will have an effect different from zero on general writing skills.*

On the one hand, when training for the ENEM essay, students will practice more writing and receive more feedback on their writing, which can improve writing skills in general. On the other

hand, treatments may hinder the development of general writing skills if the feedback it provides is too specific for the ENEM essay. While our prior is that treatments will positively affect general writing skills, we will estimate the effects on general writing skills to test whether "training to the test" is a relevant concern.

C. *The treatments will have an effect different from zero on learning in non-Portuguese topics.*

On the one hand, treatments may crowd out effort from other subjects, as students now spend more time training for the ENEM essay. On the other hand, there might be complementarities in learning. Moreover, the treatments might affect students' aspirations to enter in a post-secondary institution, which may increase effort for learning other subjects.

D. *The treatments will have an effect different from zero on other Portuguese non-writing topics.*

Since writing and non-writing Portuguese classes are taught by the same teachers, we expect all of the mechanisms considered in hypothesis **C**, plus an additional one that the treatments may affect the time allocation of the Portuguese teacher.

5.4.2 Data

With respect to hypothesis **B**, students will write another essay in the same day when the ENEM-like essay will be collected. This essay will have a different structure from the ENEM essay, and was developed to evaluate students' general writing skills. We engaged one of Brazil's leading education testing firms, "Centro de Políticas Públicas e Avaliação da Educação" (CAEd/UFJF), to design the prompts and grade the essays under our supervision. Such essay will be an additional part of the standardized state exam that students take in Espírito Santo, and will be presented to teachers and students as an initiative of SEDU/ES, not related to the programs we are evaluating. With respect to hypothesis **C**, we will combine information from the ENEM Mathematics, Natural Sciences, Human Sciences tests, and the mathematics standardized exam administered by the SEDU/ES. In order to test hypothesis **D**, we will combine information from the ENEM Language and Codes test and the Portuguese standardized exam administered by the SEDU/ES.[10]

6 Research Plan

6.1 Identification and Estimation

Restricting the analysis to schools that were included in the experimental sample, the causal impact of being offered a chance to participate in the programs is identified and can be studied by comparing the outcomes of schools randomly selected for treatment conditions and the outcomes

[10]Microdata on the ENEM exam is available from INEP, while standardized exam (called PAEBES) will be provided from the Secretary of Education.

of schools randomly selected to form the control group. The intention to treat (ITT) effects of the two treatment arms will be estimated based on the regression

$$Y_{ise} = \tau^{\text{Standard}} W_s^{\text{Standard}} + \tau^{\text{AWE}} W_s^{\text{AWE}} + \mathbf{X}'_{ise}\mathbf{\Lambda} + \epsilon_{ise}, \tag{1}$$

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 ENEM-like essay that will be included in the state standardized exams. We will append in this regression information on these two scores to maximize the power of our experiment. The variable W_s^{Standard} (W_s^{AWE}) is an indicator that takes value 1 if school s was assigned to the version of the program with(out) human graders. The vector $\mathbf{X}_{ise}$ includes strata fixed effects, and individual- and school-level covariates. We also include an indicator variable for the exam. The covariates are also interacted with this indicator, to take into account that the set of covariates available for observations from the 2019 ENEM are different from the other exam.[11] We will use clustered standard errors at the school level to take into account not only that the error term ϵ_{ise} may be correlated for different students in the same school, but also that we have information on up to two test scores for each student.

The differential ITT effect between the two treatment arms will be estimated based on the regression

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

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 $\mathbf{X}_{ise}$.[12] 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 be graded by humans), which will potentially improve the power in this comparison. We will run similar regressions for the mechanisms and secondary outcomes.[13] For the outcomes collected at the teachers' survey, since we had open questions on number of essays and on the time allocation, we will winsorize the data at the top 1% to avoid problems with outliers.

We will also estimate the distributional effects of the program in order to understand whether

[11] For both tests, we include as covariates: gender, age dummies ranging from 17 or less to 23 or more, educational and occupational characteristics of the mother and father of the students, household income category and the school average in the 2018 ENEM essay score. For the test we will administer, we will also include as covariates individual-level baseline Portuguese and Math test scores (from the standardized exams administrated by the SEDU/ES). We will replace missing covariate values with the control group mean and include a dummy for missing in the regression.

[12] We will 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 exam

[13] For the ENEM skill groups, we will control for the specific skill group of the regression instead of the final ENEM essay score. Likewise, for the regressions using other outcomes we will control for the pre-treatment school average of the respective outcome when available.

the program had differential impacts on different points of the distribution.

6.2 Inference and Power

We will present standard errors clustered at the school level, and we will also present p-values based on randomization inference using the randomization protocol. For the primary outcome, we simulated with data from the 2018 ENEM exam to assess the minimum detectable effect (MDE) of our study. This provides a good approximation for the MDE if we only had data from the 2019 ENEM exam. In this case, we would have an MDE of around 0.1σ for the ITT of each treatment arm, if we consider testing each treatment arm independently. These numbers are based on a significance level of 5% and a power of 80%. An effect size of 0.1σ falls within the range of positive outcomes found in many studies of educational interventions.

Since we are testing the main effects of two different treatments, and we are interested on whether either one of the interventions has an effect, we will also present p-values correcting for multiple hypothesis testing (MHT), based on the procedure proposed by Romano and Wolf (2005). Assuming that both treatment arms have an effect of 0.1σ, we would be able to reject the null for at least one of these two hypotheses with probability greater than 80% (considering a significance level of 5%). We consider the test on the differential effect between the two arms as a separate hypothesis, so we will not correct for MHT in this case.

Importantly, as explained in Sections 5.2.2 and 6.1, for our primary outcome we will combine information from the 2019 ENEM essay scores with information on the ENEM-like essay administered by the SEDU/ES. Since we will have more information than we used in the simulations for the power calculation, we see the numbers above as extremely conservative. Moreover, since we will have more individual-level covariates for this other ENEM essay, we expect these numbers to be conservative even if microdata from the 2019 ENEM do not become available (as we explain in Section 5.2.2, we see that as an unlikely, but possible, event).

When we consider the effects of the two treatment arms for the three ENEM groups of skills separately, we will present marginal p-values and p-values corrected for MHT in two dimensions (multiple treatments and multiple outcomes).[14] For the mechanisms, we will consider four different families of variables (see details in Table 2). First, we will construct an index on the amount of training and amount/quality of feedback, as proposed by Anderson (2008), including all the information on mechanisms **M**.1 to **M**.3. In addition to uncorrected p-values, we will also present p-values corrected for MHT to take into account that we have two treatment arms. We will then present effects on each component of this family, also presenting both uncorrected p-values and p-values correcting for both multiple treatments and multiple outcomes. The second family of mechanisms is related to aspirations towards post-secondary education. We will construct an index comprising information on the mechanisms **M**.4 and **M**.5. A third family of mechanisms is about teacher knowledge about students (mechanism **M**.6), which will be treated similarly. A

[14] When we consider the main effects of the treatments, we will correct for the fact that we are testing 6 hypotheses (3 outcomes $\times$ 2 treatments). When we consider the differential effects, we will correct for the fact that we are testing 3 hypotheses (3 outcomes).

final family of mechanisms is related to teachers' time allocation. We will look on effects on the share of hours allocated to the sum of (i) "preparing lectures and materials for activities, including home assignments", (ii) "giving individual support to students (one-on-one tutoring), guiding and counseling those with academic problems or special interests" (mechanism **M**.7). We will also estimate the effect on the number of extra hours and on indicators of teachers' perception on time availability. We will present p-values taking into account that we have multiple outcomes and two treatments in this family. We will also present a p-value for the null that the treatment had no effect for all the variables used to describe mechanism **M**.7, excluding the ones that describe perceptions on time available. This will be informative about whether the treatments changed the time allocation of teachers. Finally, for the secondary outcomes, we will present effects on indices of general writing skills, achievement on Portuguese (non-writing) exams, and achievement on non-Portuguese exams. We will present naive p-values (not taking into account MHT), and p-values taking into account that we have multiple secondary outcomes and two treatments.

6.3 Heterogeneity in Observables

To understand how treatment effects differ in terms of observable characteristics, we will expand the regression models (1) and (2) including interactions with specific variables.[15] We will consider heterogeneity with respect to the following variables:

1. Gender;
2. Race (white and Asian *vs.* others);
3. Socio-economic status, using data on household income (below versus above the median);
4. Quartiles of achievement in the distribution of a baseline Portuguese Language standardized test administered by SEDU/ES;
5. An indicator of whether the school is full or part time;
6. Number of classrooms that the teacher teaches (below versus above the median);
7. An indicator of whether the student participates in a pre-ENEM training program from the SEDU/ES.[16]

If we find heterogeneous effects on the primary outcome in one of these dimensions, then we will also investigate whether there are heterogeneous effects in the mechanisms we described in Section 5.3. This can potentially provide further insights on the relevance of each mechanism.

[15] We will interact each treatment dummy with each value of the specific variable we are considering for heterogeneous effect, and also include this variable in level. We will run separate extended regression models (1) and (2) for each specific variable we are considering for heterogeneous effects.

[16] SEDU/ES offers this program to around 2000 students, based on their interest and on their school achievement in the previous year. This program is available for all students in the state, not only the ones in our experiment. The definition of which students participated in this program was finalized before our randomization.

6.4 Potential Challenges

6.4.1 Attrition

While, as argued in Section 5.2, we do not believe differential attrition will be a relevant problem for our primary outcome, after having access to the essay scores in both sets of data we will be able to test whether either one of them presents evidence of differential attrition. If we find evidence of differential attrition for at least one dataset, then we will consider two alternative strategies. The first strategy is to present upper and lower bounds on the effects based on Lee bounds (Lee, 2009).[17] We will also consider discarding the dataset with higher differential attrition. Note that the first strategy has the advantage of using all of the data, which implies lower standard errors, while the second one does not use all data, but potentially identifies a tighter set. We will focus our analysis on the strategy that gives the tighter confidence intervals, and present the alternative one in the appendix. See Ferman and Ponczek (2017) for a discussion on this strategy.

6.4.2 Partial Compliance

The interventions we are analyzing can only be effective if students and teachers actually use the platform. This could be a challenge if teachers and students are not motivated to participate in the program, or if there are relevant infra-structure problems (e.g., lack of computers or unstable internet connection). Even though these interventions do not require a high quality internet connection to work, infra-structure could still represent a relevant bottleneck in the implementation of these programs.

Importantly, the implementer has information on usage, so we will have information on the proportion of students that actually used the platform, and the intensity of usage. In recent interactions with the implementer we were informed that all treated schools are actively participating in the programs, and in the first two activities roughly 85% of the students submitted their essays.

Partial compliance from students in the control schools participating in the program would not be possible, as the implementer has control over which schools receive the program.

6.4.3 Student and Teacher Mobility

A potential threat to the validity of our experiment would be teachers and/or students switching to different schools because of the treatment. This could happen if, for example, more motivated teachers and/or students move to treated schools to participate in the program. We believe such movements are extremely unlikely, because the randomization and disclosure of the treated schools were made in the beginning of May 2019, a couple of months after the school year began. Nevertheless, we will get administrative data from the SEDU/ES on teachers and students initial allocation and transfers to check if this is a relevant concern.

[17] In this case, the type of exam (whether the ENEM essay or the ENEM-like essay) would be a covariates, so that we allow for the differential attrition to differ depending on the exam.

7 Follow-up Papers

In addition to the short-run outcomes described in Section 5, we will also work on subsequent papers focusing on medium- and long-run outcomes.

7.1 Medium-run Outcomes: Post-secondary Education

E. *In the medium-run, the treatments might have direct effects on the likelihood of enrollment, college degree chosen, the quality of and progression through post-secondary educations institution for individuals who enroll, taking into account that ENEM mediates the entry into post-secondary institutions and that general writing skills might affect achievement.*

7.1.1 Data

We will test the hypotheses above using the following outcomes:

- course chosen in a post-secondary institution in the years following the interventions;
- enrollment in a post-secondary institution in the years following the interventions;
- progression in a post-secondary institution in the years following the interventions;
- quality of the post-secondary institution in the years following the interventions.

Administrative microdata from post-secondary institutions is also available from INEP in the Post-Secondary Education Brazilian Census ("Censo da Educação Superior"). These data can be linked to our data using student individual identifiers.

7.2 Long-run Outcomes: Labor Markets

F. *In the long-run, treatments may have impacts on labor market outcomes, either because admission into post-secondary education changes the opportunities in the labor market or because the program ended up fostering writing skills that affect productivity.*

7.2.1 Data

We will consider this hypothesis by looking at the following indicators of labor market attachment, job quality and productivity (conditional on *formal* employment):

- an indicator of whether the individual could be matched to some record with a positive wage in the same year, which we will interpret as the event of at least one month of employment in the formal labor market;
- establishment or firm size;

- occupation;
- yearly mean work earnings in the formal labor market;
- yearly mean work earnings in the formal labor market in logarithm, accounting for selection.

Outcomes. Follow-up data on labor market outcomes will use the Labor's Annual Social Information Report ("Relação Anual de Informações Sociais", RAIS), a confidential longitudinal data set of compulsory administrative records reported by every employer in the formal market. It omits workers without signed work cards ("carteira assinada"), including interns, the self-employed, elected officials, domestic workers and other smaller categories. For workers with multiple employment spells in a given year, we will keep the observation with the highest observed wage.

References

Almeida, A. M. F., Giovine, M. A., Alves, M. T. G., and Ziegler, S. (2017). Private education in argentina and brazil. *Educação e Pesquisa*, 43(4):939–956.

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.

Costa, G. L. M. (2013). O ensino médio no brasil: desafios à matrícula e ao trabalho docente. *Revista brasileira de estudos pedagógicos*, 94(236).

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.

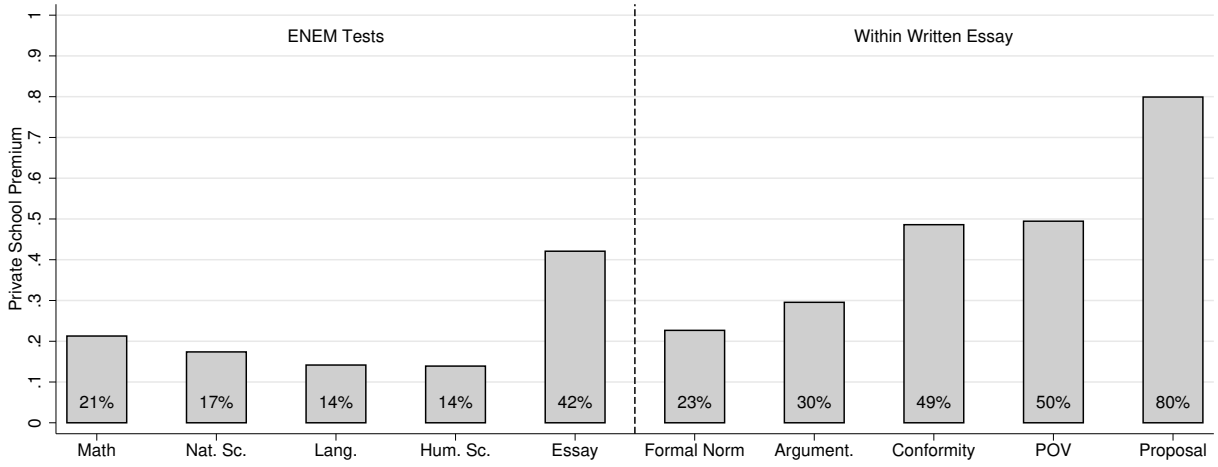
INEP/MEC (2018). Redação do enem 2018. *Cartilha do Participante*.

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

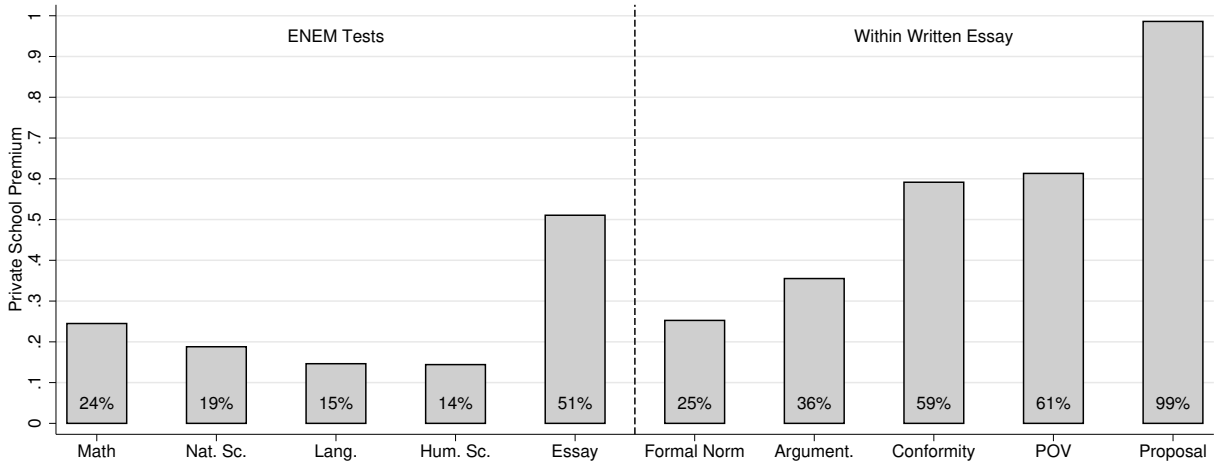
Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

Figure 1: Public *vs.* Private Gap in ENEM's Tests and Written Essay

Panel A: Brazil (N=1,149,023)



Panel B: Espírito Santo (N=25,553)



Notes: This figure illustrates the magnitude of learning gaps in the National Secondary Education Exam ("Exame Nacional do Ensino Médio", ENEM) in public *vs.* private schools. The figure in the top (bottom) panel is based on data on the universe of high school seniors in Brazil (Espírito Santo) that were present for each of the tests in 2018 (we excluded students from federal schools, which are typically very different from other public schools in Brazil). The leftmost bars relate to the five tests that compose the exam: Mathematics (Math); Natural Sciences (Nat. Sc.); Language and Codes (Lang.); Human Sciences (Hum. Sc.); and the written essay (Essay), respectively. The rightmost bars consider each skill in the written essay individually: adhere to the formal written norm of Portuguese (Formal Norm); use argumentative linguistic structures (Argument.); conform to the proposed essay topic (prompt), and develop a text using knowledge from different areas (Conformity); select, relate, organize and interpret data and arguments in defense of a point of view (POV); elaborate a policy proposal that could contribute to solve the problem in question, respecting the basic human rights' principles (Proposal).

Table 1: Correlations — ENEM 2018 Written Essay

Formal written norm	1.00				
Argumentative Structure	**0.82**	1.00			
Conformity and Text Develop.	0.60	0.64	1.00		
POV construction	0.64	0.68	**0.94**	1.00	
Proposal	0.56	0.65	0.74	0.75	1.00

Notes: Correlation between skills in the ENEM written essay data on the universe of high school seniors in Brazil present in 2018.

Table 2: Hypotheses and Families of Outcomes and Mechanisms

Groups of variables	Data	Hypotheses	Prior
Primary Outcome			
ENEM Scores	ENEM 2019 written essay + ENEM-like essay (pooled and separated)	A.1-A.3	AWE> 0, Standard> 0 $\Delta \lesseqgtr 0$
ENEM Scores (*per* Skill)			
Syntax			
Formal norm + Argumentative struct.	ENEM 2019 written essay + ENEM-like essay (pooled and separated)	A.1-A.3	AWE> 0, Standard> 0 $\Delta = 0$
Lower-level Semantics			
Conformity and Text Dev. + POV construction	ENEM 2019 written essay + ENEM-like essay (pooled and separated)	A.1-A.3	AWE> 0, Standard> 0 $\Delta \lesseqgtr 0$
Higher-level Semantics			
Policy Proposal	ENEM 2019 written essay + ENEM-like essay (pooled and separated)	A.1-A.3	AWE> 0, Standard> 0 $\Delta \lesseqgtr 0$
Mechanisms			
Training and Feedback			
Number of ENEM training essays	teacher and student surveys	M.1	AWE> 0, Standard> 0 $\Delta = 0$
Number of ENEM training essays, in class	teacher survey	M.1	AWE> 0, Standard> 0 $\Delta = 0$
Number of essays followed by class discussion about common mistakes	teacher survey	M.2	AWE> 0, Standard> 0 $\Delta \lesseqgtr 0$
Number of essays followed by class discussion about good practices	teacher survey	M.2	AWE> 0, Standard> 0 $\Delta \lesseqgtr 0$
Number of essays commented or annotated	student survey	M.2	AWE> 0, Standard> 0 $\Delta \lesseqgtr 0$
Number of ENEM essays individually graded	teacher and student surveys	M.2	AWE> 0, Standard> 0 $\Delta \lesseqgtr 0$
Number essays followed by discussion with teacher	student survey	M.2	AWE> 0, Standard> 0 $\Delta \lesseqgtr 0$
Perceived usefulness of comments or annotations	student survey	M.3	AWE> 0, Standard> 0 $\Delta \lesseqgtr 0$
Aspirations and Expectations — Post-secondary Education			
Plans for 2020 include PSE	student survey	M.4	AWE> 0, Standard> 0 $\Delta \lesseqgtr 0$
Expectations % students in PSE 2020	teacher survey	M.5	AWE> 0, Standard> 0 $\Delta \lesseqgtr 0$

(*cont.*)

(*cont.*)

Groups of variables	Data	Hypotheses	Prior
Teacher Knowledge About Students			
Perceptions about how much knows about weaknesses and strengths	teacher survey	M.6	AWE$\lesseqgtr$ 0, Standard$\lesseqgtr$ 0 $\Delta \lesseqgtr 0$
Difference between predicted *vs.* actual ENEM essay grade	teacher survey	M.6	AWE$\lesseqgtr$ 0, Standard$\lesseqgtr$ 0 $\Delta \lesseqgtr 0$
Teachers' Time Allocation			
Share of hours working (*per* group of tasks)	teacher survey	M.7	AWE$\lesseqgtr$ 0, Standard$\lesseqgtr$ 0 $\Delta \lesseqgtr 0$
Number of extra-hours	teacher survey	M.7	AWE$<$ 0, Standard$<$ 0 $\Delta \lesseqgtr 0$
Perception of time available, *per* subject	teacher survey	M.7	AWE$<$ 0, Standard$<$ 0 $\Delta \lesseqgtr 0$
Secondary Outcomes			
General Writing Skills	written essay	B	AWE$\lesseqgtr$ 0, Standard$\lesseqgtr$ 0 $\Delta \lesseqgtr 0$
Non-Language-related subjects	ENEM 2019 PAEBES 2019	C	AWE$\lesseqgtr$ 0, Standard$\lesseqgtr$ 0 $\Delta \lesseqgtr 0$
Language-related subjects (non-writing)	ENEM 2019 PAEBES 2019	D	AWE$\lesseqgtr$ 0, Standard$\lesseqgtr$ 0 $\Delta \lesseqgtr 0$

Notes: This tables lists the main hypotheses on primary outcomes, secondary outcomes and mechanisms.