# Evaluating the effect of tutoring sessions on second-grade students in Cali-Colombia

Juan Guillermo Salguero Astudillo

**Tutor:** Monica Hernandez Florez

**Reader:** Catherine Guirkingier

Universidad EAFIT
Escuela de Finanzas, Economía y Gobierno
Maestría en Economía Aplicada
Medellín, Colombia
2026

## Abstract

The study estimates the effect of the program Aula Global on second-grade students from 13 public schools in Cali,Colombia during 2024. The intervention consisted of a series of tutoring sessions which aimed to strengthen students' reading and mathematics skills. Using a Difference-in-Differences approach, the study concludes that there is a marginally significant positive effect of the tutoring sessions on students' performance. The results also indicate that individual and schools-level characteristics are associated with learning outcomes, and that the intervention may affect students differently depending on their baseline abilities.

**Key words: Early-grade education; Tutoring interventions; Reading and Mathematics; Difference-in-Differences; Colombia; EGMA; EGRA**

# Contents

## 1. Introduction

There are different reasons why we are interested in improving education in the country. The relationship between the level of education and a society's development is a topic with extensive evidence around the world. From this perspective, the development of strategies to improve education both in terms of coverage and quality is widely recognized as important.

Concern about educational quality has become much more relevant in recent years because of the crisis caused by COVID-19. The new realities caused by the pandemic practically forced countries to reinvent the way in which classes were taught. Once the crisis was over, a need was identified to recover the learning process students had before the pandemic. Moreover, many international organizations have recognized the importance of addressing learning gaps from the earliest possible moment due to the future impact on students' life (World Bank, 2019).

Considering the objective of tackling learning gaps, one intervention that has proved to have positive effects is tutoring sessions with students. This kind of intervention allows students to receive extra time of instruction which may lead to the strengthening of their academic skills. Studies such as Nickow, Oreopoulos, & Quan (2020) show that this type of interventions have positive effects on enhancing students' reading abilities.

Together We Learn is a project resulting from a consortium between the Carvajal Foundation, Parque Explora, and Proantioquía. Under the direction of Partners of the Americas and with USAID resources, the project developed different educational activities between 2022 and 2025 with the objective of improving the quality and educational coverage in some of the cities of Colombia that have been recipients of migration in recent years.

One of the activities developed under this project's umbrella was Aula Global, which consisted of a series of tutoring sessions in language and mathematics for public school students from the second grade of primary school. The strategy addressed learning gaps concerning the basic skills of students in reading and mathematics through different strategies such as games and songs, that enhanced students' motivation to participate while increasing their academic skills (Tulcan Tapia, Valencia Daza, & Hernández Cruz, 2023)

In Colombia, various public and private organizations, such as the Secretaries of Education, the Luker Foundation, and the Carmelita Foundation, among others, have tried to measure the academic progress of students' basic skills using the EGMA (Early Grade Mathematics Assessment) and EGRA (Early Grade Reading Assessment) tests. These tests are relevant because they are a standardized way to evaluate students' reading and mathematics skills and they give the opportunity to monitor students' development.

Given the objectives of the Aula Global program, this study addresses the following research question: What is the impact of the program Aula Global on second-grade students' reading and mathematical skills?

The general objective of this study is to analyze the impact of the program Aula Global on second-grade students from Cali, Colombia during 2024. This will be done through the results obtained from the EGMA and EGRA tests that allow evaluation of the basic skills in language and mathematics of second grade students from Official Educational Institutions in the city. In this sense, the specific objectives of this exercise are the following:

- Develop a Difference-in-Differences (DiD) analysis to isolate the impact of Aula Global tutoring sessions on enhancing reading and mathematics skills on second grade students from Cali Colombia during 2024.
- Determine the impact of other sociodemographic variables like gender, school/campus, frequency of attendance to the institution and having internet on the students' performance.
- Analyze if the program's impact varies depending on the initial students' skill level.

## 2. Theorical Framework

For this type of study, various previous works can be consulted, we might divide the existing literature into three scopes. First, various studies discuss the importance of education for a country's growth. From an economic-society perspective, we can find the work of Becker & Chiswick

(1966), which explores how a person's level of education impacts income distribution in society. Contemporary extensions of this framework emphasize that skill formation is a cumulative process: early abilities facilitate the acquisition of more complex competencies later in life. Cunha & Heckman (2007) explain how early childhood skills follow a dynamic process in which “skills beget skills” meaning that early investments are more productive than later remediation and can substantially reduce long-term inequalities.

Within this context, reading and mathematics skills developed during the first years of primary school are critical, as they become the basis of the student’s future academic performance. When students fail to acquire basic competences at a certain age, their ability to keep with more advanced topics becomes compromised, potentially increasing long term disparities in academic trajectories and labor market outcomes (Heckman & Mosso, 2014).

Recognizing the importance of developing foundational competencies in early childhood is essential for understanding why governments invest in policies that aim to strengthen these skills. Publications such as that of Litsching & Morrison (2013) highlight how increased intergovernmental transfers can improve literacy rates, school attendance and reduce poverty rates. Similarly, works such as that of Manuelli & Seshari (2014) or that of Bils & Klenow (2000) analyze the importance of the relationship between a government's contribution to building human capital and the returns it can have in terms of growth and greater productivity in the future.

When talking about learning gaps its important to present the concept of "learning poverty." According to the World Bank & UNESCO (2019), the term refers to all children who, by the age of 10, cannot read or understand a short paragraph. To analyze the components of this term and the status of countries regarding this issue, we can analyze various sources from the World Bank, UNESCO, UNICEF, USAID, etc., which constantly measure this indicator. This topic turns to be an urgent matter in Colombia and other countries in Latin America where more than 50% of the children of that age might be unable to read a short paragraph (World Bank, 2019).

In this context we must highlight the consequences of the pandemic as a magnifier of the Learning Poverty problematic, especially in Latin America. School closures during 2020–2021 led to

unprecedented interruptions in instructional time and reduced opportunities for guided reading and mathematics practice. International assessments show considerable declines in foundational skills among early-grade learners, with the largest setbacks observed in low-income and vulnerable populations (UNESCO, 2022; UNICEF, 2021). In Colombia, early-grade assessments have documented notable deterioration in basic reading fluency and numeracy immediately after the pandemic, especially in public schools serving disadvantaged communities.

Finally, considering that an analysis will be conducted on the effectiveness of language and math tutoring for elementary school students, a review of studies that conduct similar analyses will be conducted. Studies from Stanford University, Brown University, and Vanderbilt University assess programs like Aula Global. Some studies like the one of Robinson, Pollard, Novicoff, White, & Loeb (2024) show that 2$^{nd}$ grade students that received virtual tutoring increased literacy skills by 0.05-0.08 SD in comparison to control group students. Similarly, Roschelle, Cheng, Hodkowski, Neisler, & Haldar (2020) concluded that students who participated in mathematical tutoring sessions scored significantly more than students who didn't.

The literature reviewed suggests two primary channels through which tutoring interventions such as Aula Global can improve early learning outcomes. First, tutoring increases students' exposure to structured practice in core domains such as phonemic awareness, decoding, reading comprehension, and basic numeracy. Additional guided instructional time has been associated with higher learning gains, particularly among young learners, because it enables repeated practice, individualized feedback, and reinforcement of foundational skills (Robinson et al., 2024; Roschelle et al., 2020).

Second, Aula Global incorporates a pedagogical model that emphasizes motivation and student engagement through games, rounds, stories, riddles, and songs. As documented by Tulcan Tapia, Valencia Daza, and Hernández Cruz (2023), this methodology is explicitly designed to create a dynamic and enjoyable learning environment that encourages active participation. This type of interactive instruction can enhance students' willingness to engage with academic tasks, sustain attention, and persist in learning activities—factors that are essential for skill acquisition in the early grades.

Together, increased instructional time and higher student engagement constitute plausible and theoretically grounded mechanisms through which the tutoring intervention may influence performance on EGRA and EGMA assessments.

The rationality of these mechanisms can be supported by existing literature. From an economic perspective, early academic skills follow a cumulative and dynamic process in which current learning increases the productivity of future learning. In this sense, interventions delivered in the early grades can generate disproportionately high returns because foundational skills enhance the efficiency with which later competencies are acquired (Cunha & Heckman, 2007). Within this framework, the amount and quality of instructional time become a key input in learning skills: additional guided practice in reading and mathematics increases exposure to structured learning opportunities, facilitates consolidation of basic skills, helps students that might be lagged in the process and allows teachers to correct misconceptions before they become persistent barriers to academic progression.

The literature also documents that tutoring interventions might impact students differently. The impact variation may be related to the different skill level students have at the beginning of the intervention (Nickow et al., 2020; Elbaum et al., 2000). In this sense, students' characteristics may determine how they can benefit from the interventions they participate in.

In this regard, several studies explain how additional time dedicated to strengthening the basic skills might lead students to acquire the expected academic level, especially in the early grades. Evidence from Lavy (2015), Bellei (2009), and Andersen et al. (2016) find that additional hours of structured instruction in reading and mathematics significantly boost performance by expanding opportunities for guided practice and enabling teachers to address learning gaps more effectively. As a result, early interventions that increase meaningful learning time act directly as a crucial input in the learning process.

Regarding the second mechanism, it is important to consider that in addition to cognitive inputs, the development of early academic skills depends on non-cognitive factors such as motivation,

attention, and engagement. Economic models of skill formation recognize the importance of these socio-emotional attributes as complementary to cognitive investments capable of raising the productivity of instructional time by increasing students' willingness to participate and persist in learning activities (Heckman, Stixrud, & Urzúa, 2006). On this way, pedagogical approaches that stimulate interest and enjoyment can enhance learning outcomes not only directly, through improved comprehension, but also indirectly by strengthening the behavioral skills (like motivation, participation and engagement) that enable students to benefit more fully from instructional opportunities.

## 2.1 Previous studies

It is important to highlight that there have already been publications describing the effects of Aula Global in Cali and the Pacific region of Colombia. Works such as that of Tulcan Tapia, Valencia Daza, & Hernández Cruz (2023) detail the methodology used in tutoring sessions of the activity in Cali and describe some effects identified in previous analyses. In the same way, the work of Barrera Osorio, Gonzalez, Lagos, & Deming (2020) takes the results of EGMA and EGRA to measure the effect of providing information to the family about students' performance in the reading area.

This activity has been evaluated in different ways in the past. First, a pilot study was conducted with the first students who received the intervention in 2017. This report recognized a 0.26 SD increase in language and math skills among students participating in the program compared to the control group (Barrera & Lagos, 2018). This evaluation was conducted with second, third, fourth, and fifth-grade students from different schools in Cali.

In 2024, an internal evaluation done by the Carvajal Foundation was conducted within the project to analyze the effectiveness of a different methodology for developing Aula Global activities. The methodology implemented with these students was interactive audio instructions, meaning that the Carvajal Foundation sent recordings of the instructions of every activity to the schools so the activities could be carried out by their own teachers. Although this methodology was different from the traditional tutoring sessions with trained tutors, the results of the intervention were still positive for the second to fifth grade students.

Although the activity has already been evaluated in the past, the analysis that will be carried out here turns to be relevant because, (1) since the pilot test was carried out, the organizers of the activity have changed the strategy of the tutoring sessions, they have strengthened the selection and training of the teachers who participate in the sessions and the schools that are beneficiaries of the program, school selection varies according to school characteristics (population changes, population dropout rates, school willingness to participate in the programs) and stakeholders decisions (Secretary of Education). (2) The evaluation that will be carried out will be on the face-to-face methodology and not on other methodologies as it has been done internally in recent years. There were some years in which both tests and the tutoring sessions were done remotely due to the COVID-crisis, local security shocks or impossibility of mobility to the institutions. (3) The selection process for treatment and control students will be different, on this occasion, unlike Barrera & Lagos (2018), both groups were in the same school but in different classrooms, this guarantees that the characteristics of the treatment and control students are almost the same, differing only in the classroom to which they belong.

The schools that participated in the program were selected as some of the most needed assistance as beneficiaries of the Aula Global project. Once this selection was made, randomization was conducted in the second-grade classrooms to determine which classroom would be the treatment and which the control group as is explained in the following section.

## 3. Methodology

The methodology consists of a difference-in-difference analysis in which the group of students that were part of the treatment are compared with a control group. The analysis takes as main variable the total score that the students obtained in the EGMA and EGRA tests at the beginning and at the end of 2024 intervention. The tests have a specific focus on phoneme recognition, decoding of simple and invented words, reading fluency, and reading comprehension in language, and recognition of magnitudes, sequences, and basic addition and subtraction operations in mathematics. Every module is considered essential, so all of them will be taken into consideration for the final score.

Regarding the counterfactual analysis, at the classroom level, the control group has been selected randomly among other second grade classrooms from the same schools from the treatment. The final control group, meaning the students from the control classroom that were directly compared to the treatment group, were selected according to the students' score in the pre-test. As one condition for entering the program is to obtain less than 60% correct answers in the diagnostic test, the same rule is used to determine which students will take part of the control group.

Unluckily, the budget for the evaluation allowed the organization only to retrieve information of the students that got less than 60% in the diagnostic test, meaning that a lot of students that got 60% or more in the pretreatment were not evaluated at the end of the intervention.

The analysis also determines the effect of different sociodemographic variables on the performance of the students. There has been collected information on the gender, age, campus, frequency on school attendance (reported by students) and (house) internet connection of students. In addition to this, the analysis might reveal what kind of students are the most beneficiated from the intervention. Since Aula Global is designed to strengthen the most basic skills of the students, it is expected that the students who show the poorest performance in the diagnostic test also show the greater advance in the final one.

## 3.1 Aula Global overview

Since 2017 and together with Harvard University, Vanderbilt University and Luker Foundation, the Carvajal Foundation has been implementing the activity of Aula Global. The activity consists of a series of 24 tutoring sessions to address elementary students' educational gaps of language and mathematics basic skills. The frequency of the tutoring sessions is 2 per week, which means that the activity takes place over a period of between 4 and 5 months. Although in this study we are only considering second-grade students, due to program budget and data availability, the activity is designed for students from second to fifth grade.

The tutoring sessions contain different flexible strategies to strengthen the student's abilities, and on the language session the students work on textual comprehension, writing and discourse elaboration by using varied dynamic activities like poems, riddles, couplets, short stories and songs.

In the mathematics sessions the students focus on counting skills, the base 10 number system and basic operations (Tulcan Tapia, Valencia Daza, & Hernández Cruz, 2023)

The tutoring sessions consist of one-hour sessions and are guided by a trained tutor who visits the school two times a week to conduct the activities, normally the activities replace the student's regular classes.

## 4. Data

The standardized tests used for this analysis were the EGMA (Early Grade Mathematics and EGRA (Early Grade Reading Assessment) tests. These tests were developed by the company RTI, based on a request from the World Bank and USAID. The Carvajal Foundation has always used these tests to measure the diagnosis of students at the beginning of the activity and the final state once they finish the intervention. Despite this, on few occasions it considers collecting information from a control group for comparison, mainly due to budget issues regarding the activities.

The test contains different tasks that the student must complete at a certain time. The tasks are divided into the following categories:

- Reading abilities: Evaluates the students reading skills through different activities like simple words reading, pseudoword reading, short text reading, reading comprehension and oral comprehension.
- Mathematics abilities: Evaluates the students mathematics skills through different activities like number comparisons, sequences and identification of missing numbers, additions and subtractions.

The following table shows the number of items in each activity, the final score of the students is taken as the summatory of all the items correctly answered. Every item is a task that the students must complete, could be a word the student must read (e.g.: Reading of a passage has 100 items, meaning that the student is asked to read a 100-word text in a certain time), or a question he must answer (like a mathematical problem or a question about the text).

| Test | Task | Number of Items |
|---|---|---|
| EGRA | Simple words reading | 50 |
| | Pseudowords reading | 50 |
| | Reading of a passage | 100 |
| | Reading comprehension | 6 |
| | Oral comprehension | 4 |
| EGMA | Number comparisons | 10 |
| | Sequences | 10 |
| | Additions | 25 |
| | Subtractions | 27 |

*Table 1: Description of EGMA and EGRA tasks*

The EGMA and EGRA tests have already been used for other studies in Colombia, for example the study by Barrera-Osorio, Gonzalez, Lagos & Deming (2020) used data from this assessments to evaluate the improvement in students' skills when their parents were constantly informed of their progress, here the authors recognized an initial positive transitory impact on students whose parents were informed about their advances.

The sample for the analysis comprises 458 second-grade students from 13 official educational institutions in Cali. Regarding that every institution has its own treatment and control classroom, there were 13 treatment classrooms and 13 control classrooms, Within the sample, there are 204 students who belonged to the program and received tutoring sessions in reading and mathematics during 6 months of their scholar year in 2024. The remaining 254 individuals were students from the same institutions that belonged to the control classroom. Figure 1 shows the geographical position of the schools in Cali, Colombia. As can be seen, most of the schools are in the east part of the city, where the lower-income households can be found.

The treatment sample contains only students who attended a minimum of 17 of the 24 tutoring sessions planned throughout the program's implementation. This is done in accordance with a recommendation made by the implementation technical team, which specifies that this is the minimum number of sessions that a student must participate in to advance their basic skills. The

rest of the students that did not complete the minimum number of sessions were not part of the analysis.

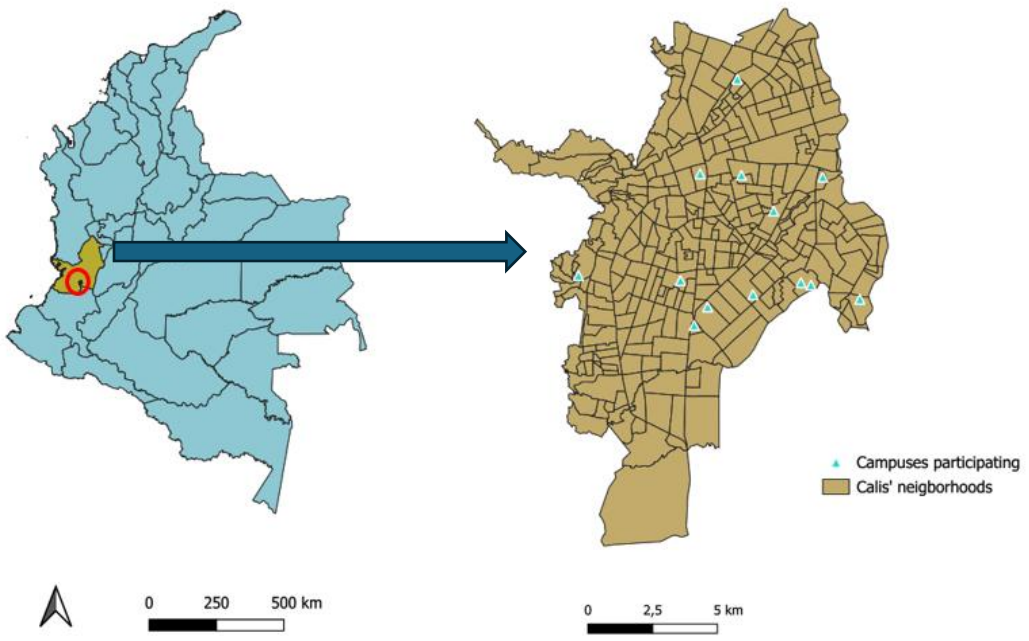



*Figure 2: Geographical location of the selected schools*

The data contains sociodemographic information regarding the school, age, group (classroom), and children's gender. It also contains the children's scores on standardized reading and mathematics tests. The following table summarizes each of the variables mentioned:

| Variable name | Type | Description |
|---|---|---|
| score_pre | Numeric | Reflects the number of items that the student correctly responded to in the post-treatment test. |
| score_post | Numeric | Reflects the number of items that the student correctly responded in the diagnostic test. |
| group | Dummy | Takes the value of 1 if student belongs to the treatment group and 0 if it belongs to the control group. |
| gender | Categorical | Describe the gender of the student (male/female) |

| | | |
|---|---|---|
| schedule | Categorical | Describes whether the student attends classes in the morning, in the afternoon, or follows a unique schedule combining both. |
| freq | Categorical | Describes the frequency of school attendance, distinguishing between more than three times per week, two to three times per week, or fully virtual attendance. |
| internet | Categorical | Describes if students have internet connection in their house. |
| school | Categorical | Reflects the school the student attends |
| campus | Categorical | Reflects the campus/site the student attends. A school may have 2 campuses. |

*Table 2: Description of the variables of the study*

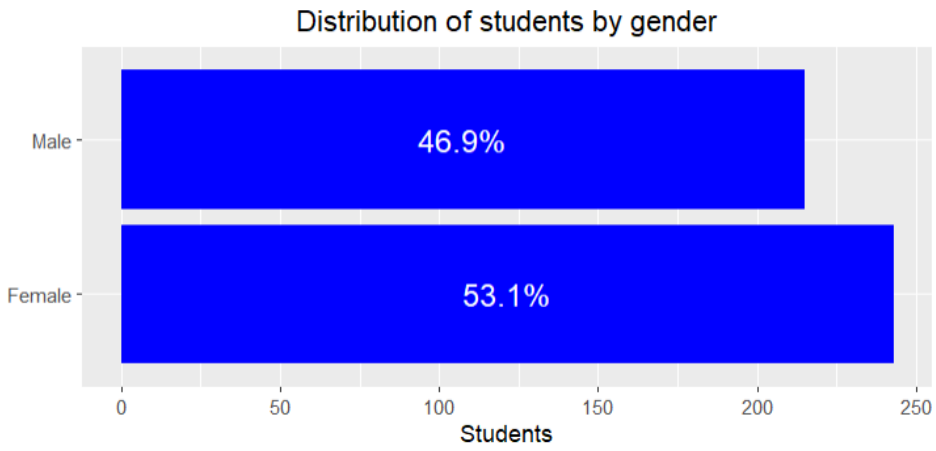
The diagnostic test was conducted in both treatment and control classrooms. In the treatment classrooms the diagnostic results determined which kids would receive the tutoring. To be part of the program, the kid would have to have a grade of less than 60% correct answers in the diagnostic test. The same test was made for the control group to select kids with the same characteristics in terms of grade to be the final control group. Then the only difference between the groups is that some kids received the treatment, and others did not.

Considering the quality of the data, we can predict that most students in both groups had low test scores. This can be determined from the intervention design, since, being focused on students who performed poorly on the tests, it is normal that most scores are below 50%.

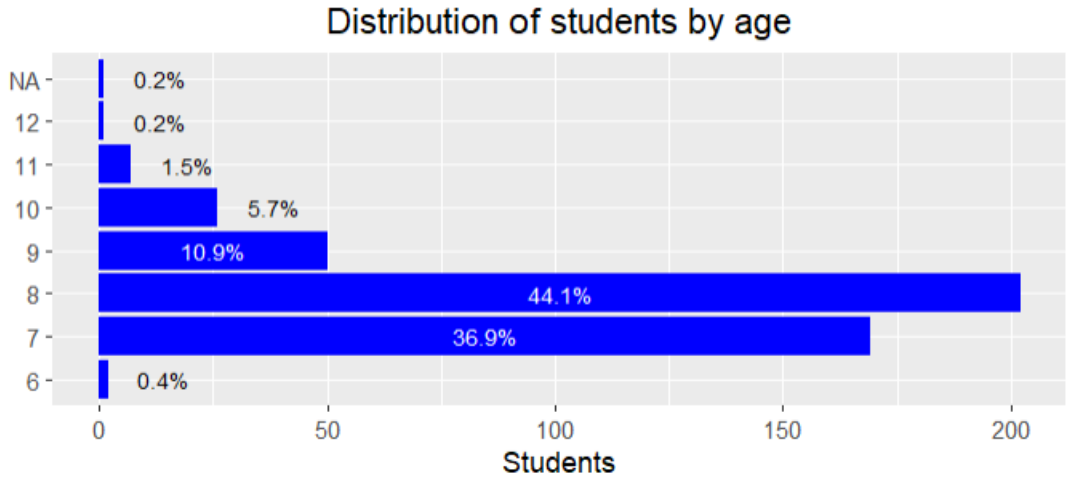
Another interesting point in this regard is the entry rule for the program, which requires a score of less than 60%. During the program, some participating students had to withdraw from school for various reasons (mostly a change of residence), so at a certain point the program had to replace them with students who had obtained a score higher than 60% on the initial test. However, these cases are minimal, and these students are included in the analysis of the project's effect.

## 5. Exploratory analysis

To begin the analysis, it's important to present basic statistics of the variables. The following graphics show the main distribution of the variables presented:

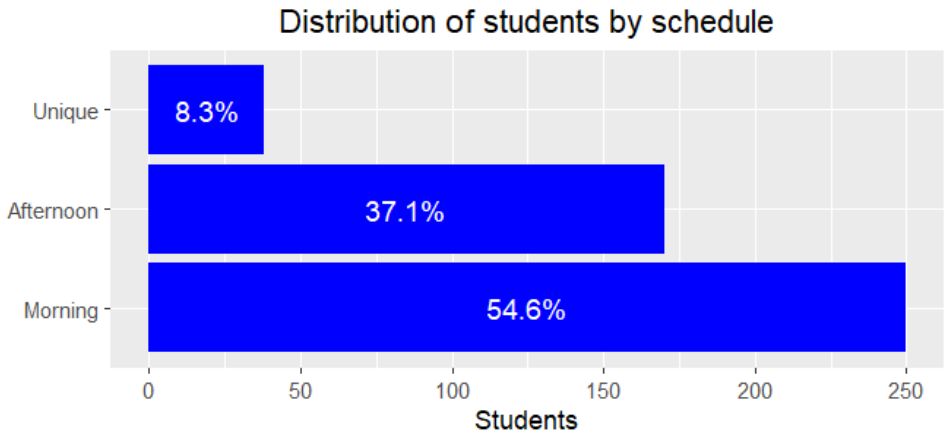



*Graphic 3: Distribution of students by gender*

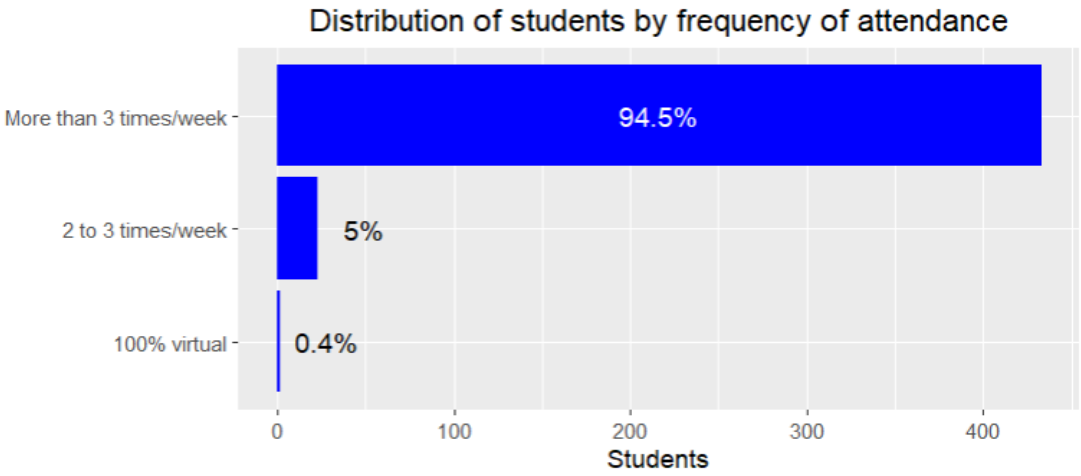



*Graphic 2: Distribution of students by age*

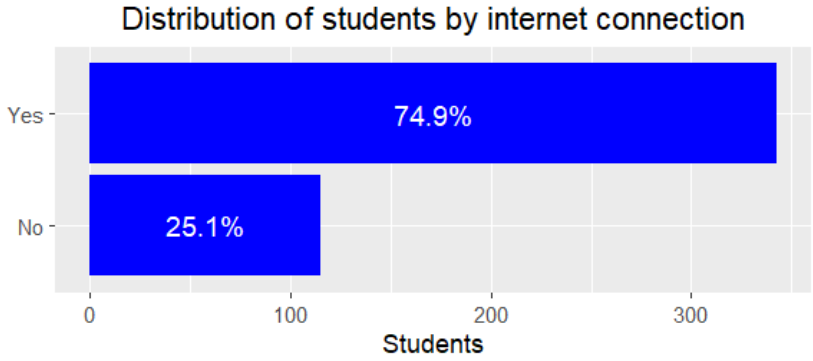
As can be seen in the distribution graphics shown above, there is almost no difference in the distribution of the students in the variable gender and most of the students are 7 and 8 years old (average second grade). However, there is presence of over-age students in the sample. The following graphics summarizes the distribution of students by other characteristics like, the schedule they attend school, the fact that they have or do not have an internet connection in their houses and the frequency of school attending (self-reported).




*Graphic 3: Distribution of students by schedule*




*Graphic 4: Distribution of students by frequency of attendance*




*Graphic 5: Distribution of students by internet connection*

As can be seen in the graphics shown above, most of the students attend school in the morning although there is an important number of children who go in the afternoon and most of the kids in the sample attend school more than three times per week. Regarding the last variable, we can conclude that there is an important size of sample that doesn't have internet connection in their houses. Although the graphics shown are a useful way to check differences on the distribution of the variables, another section will present balance tests made to analyze the same distribution within the treatment and control groups.

## 6. Method of analysis: Differences in Differences

A DiD technique was applied to estimate the effect of program. The method focused on the final score of the standardized test as the outcome variable, which measures the number of correct answers.

DiD is widely used in applied micro econometrics to estimate causal effects in this kind of context; the method compares changes over time between treatment and control groups. The main assumption we care about when performing a DiD analysis is the parallel trends assumption which states that in the absence of tutoring sessions (treatment), the children from the treatment group would have performed just as the students in the control group.

The method is a standard strategy in policy evaluation specially when pre-intervention information is available (Bertrand, Duflo, & Mullainathan, 2004; Wing, Simon, & Bello-Gómez, 2018). In literature there can be found many examples of the use of this method to analyze the effects of interventions related to education. Cortes, Goodman, and Nomi (2015) for example estimated the impact of a mathematics intensive program for low performing students using this method. Lavy (2015) also used DiD to analyze the effects of having different instruction times in different schools in Chicago.

The DiD[1] analysis turns out to be a good method to study interventions related to education, especially when there is information on the pretreatment period of both treatment and control groups. In this design, although the selection of the classrooms is randomized, the final decision on the eligible students in the tutoring sessions depends on the results of the diagnostic test, therefore, the selection of participants is not fully random.

Another option for the analysis could be the Regression Discontinuity Design (RDD) which is used when the individuals are split around a specific threshold (like the test score) that determines whether the individuals are part of the intervention or not. In our case this method turns out to be unfeasible due to the lack of information of students. As said before, due to budget limits, the organization only retrieved information from the students that got less than 60% in the diagnostic test. The only reason why there are some students whose scores are higher than the limit is because they are replacing students who were participating in the program (score with less than 60%) and abandoned at some moment. Due to the lack of individuals above the threshold and the examples of students with scores above it, using RRD method is not recommended.

## 6.1 Parallel trends and Balance tests

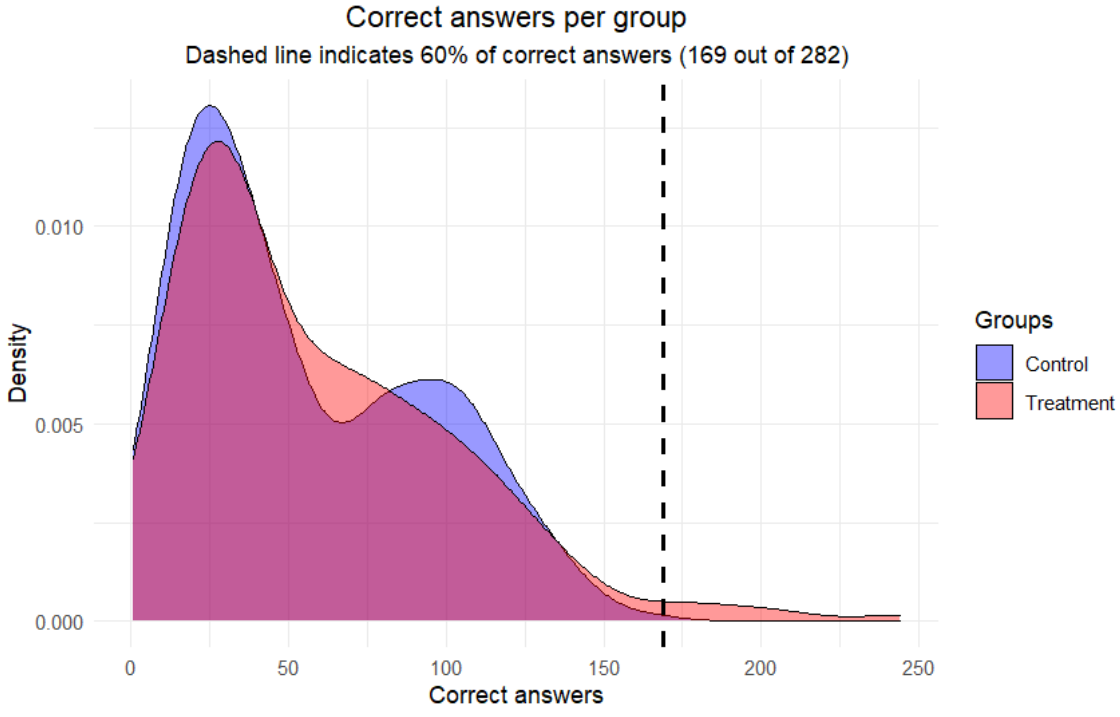
As mentioned before, when working with the DiD method the most important assumption refers to the parallel trends of the population. The assumption states that, in the absence of treatment, both the treatment and control groups would have followed the same trend over time (Angrist & Pischke, 2009). Due to the lack of information several moments before the treatment period we have difficulties in estimating the trends of both groups before the diagnostic tests. Although we cannot directly check pre-treatment trends, we can show that both groups are similar in their observable characteristics. This allows us to infer that the differences observed between the groups at the end of the analyzed period correspond to the effect of the intervention (Imbens & Wooldridge, 2009).

[1] Although the analysis is implemented using a Difference-in-Differences framework, the assignment of students to treatment and control groups was conducted prior to the intervention following a quasi-random process at the institutional level. Therefore, the study design can be considered closely related to a randomized controlled trial, with DiD used to improve estimation efficiency and account for baseline differences.

The first thing to consider regarding similarities between the treatment and control groups is the design of the quasi-experiment. The assignment of treatment and control groups was conducted within the same 13 schools, where in each school two second-grade classrooms were randomly selected as the treatment and control groups, some of these schools had more than two classrooms.

If we consider that the treatment and control groups share the same schools and campuses (in some cases even the same teachers) it's reasonable to assume that both would be affected by the same external shocks and the same institutional environment, in the same order, they might tend to have the same demographic characteristics as most of the students tend to live near the schools they attend. We also saw on the map that most of the schools are located within the same part of the city, which can be an indicator of similar characteristics among the students in the sample.

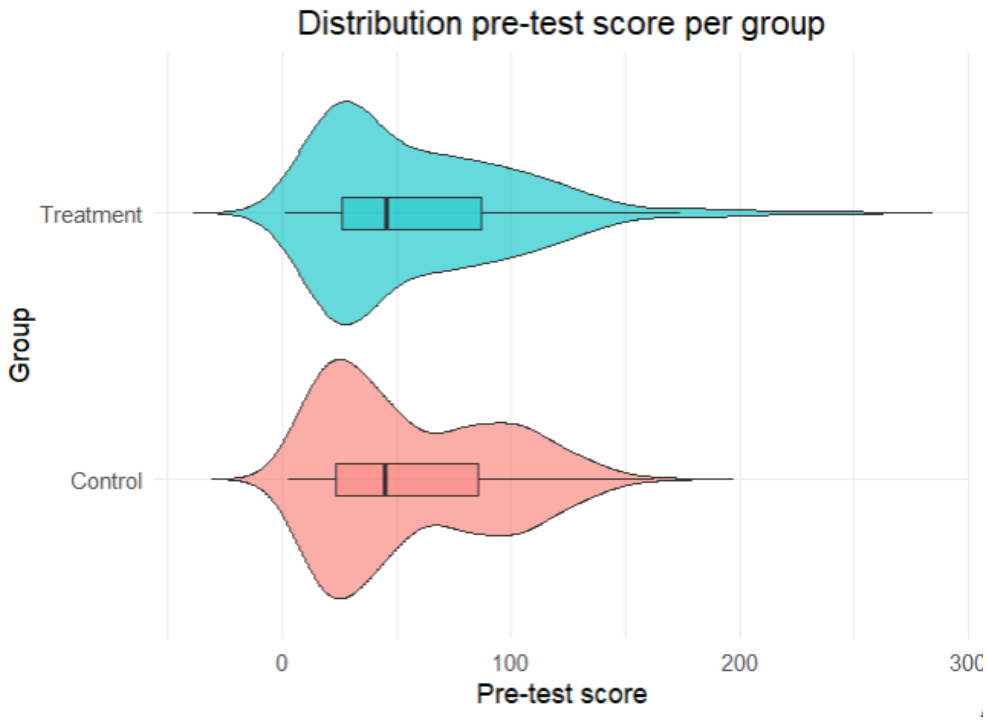
Following the objective of demonstrating similarities between the treatment and the control group we can conduct different balance tests to see graphically and statistically the different characteristics of the groups. The first aspect we test is the difference of both groups around the diagnostic test scores. The following graphic shows the distribution of the diagnostic test scores of the students of both groups.



*Graphic 6: Distribution of diagnostic scores per group, density graphic*

The graphic shows no major differences between the distribution of the scores of both groups during the pre-treatment period. As noted, before, both groups show a concentration of students with the lowest scores. This is due to the design of the selection process in which only the students with scores below 60% of correct answers were selected as the final treated group. Simultaneously, the students in the control group were selected with the same criteria so both groups ended up having almost the same number of correct answers.

To complement the previous graphic, the following violin graphic compares the dispersion of the pre-scores of the treatment and control groups. The graphic shows that the variability of both groups is highly comparable, and the boxplots show that both groups share similar median and interquartile range, meaning there are no differences in the central tendency or dispersion scores before the intervention, in other words, both groups have a similar baseline in terms of academic performance.




*Graphic 7: Distribution of diagnostic scores per group, violin graphic*

The following balance test checks for potential differences between both groups regarding the sociodemographic characteristics. The table below summarizes the results of the balance test. The table shows no important differences between students in the control and treatment groups. Pre-treatment scores, age, gender composition, and school attendance frequency are comparable across groups.

| Variable | Control | Treatment | Difference (T-C) |
|---|---|---|---|
| Pre-test score | 55.9<br>(37.9) | 59.5<br>(43.1) | 3.6<br>(3.8) |
| Age | 7.9<br>(1.0) | 7.9<br>(0.9) | 0.0<br>(0.1) |
| Female (%) | 54.7 | 51.0 | -3.7<br>(4.7) |
| Internet access (%) | 77.6 | 71.6 | -6.0<br>(4.1) |
| Attendance >3 times/week (%) | 94.9 | 94.1 | -0.8<br>(2.2) |
| Morning schedule (%) | 50.4 | 59.8 | 9.4**<br>(4.7) |
| Afternoon schedule (%) | 40.6 | 32.8 | -7.7*<br>(4.5) |

Notes: Continuous variables report mean with SD in parentheses below. Binary variables report percentages. Difference column shows Treatment - Control with standard error in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

*Table 3: Balance table*

The balance test does not include the variables of school or campus due to the design of the model. As the treatment and control groups were randomly selected and considering that during the interventions some of the students from both groups withdrew from their respective institutions, the variables of school and campus were left aside. Nonetheless, the fixed effects of the institution

are considered in the analysis. There were significant differences observed in schedule, for which the variable is controlled in the DiD analysis.

Considering that most of the variables are categorical, the analysis will not include correlation table in this document, the presentation of the variables in relation to the scores of the student has already been analyzed in the balance tables. A correlation matrix would not provide additional meaningful insights.

## 7. Model and results

The estimated model used for this analysis was a DiD specification. This model allows us to estimate the causal effect of the intervention by analyzing the differences between both treatment and control groups over time. The first estimated model includes only three variables: the variables of time (difference between pre and post scores), group (difference between treatment and control groups), and the interaction between them:

$$(1)\ score_{it} = \beta_0 + \beta_1 Time_t + \beta_2 Group_i + \beta_3 (Time_t * Group_i) + \varepsilon_{it}$$

The purpose of this first model is to get a first insight into the effect of the time, group and the DiD before adding any control variables. The results of this estimation can be seen in the first column of *Table 4*. The results of the first estimation confirm three main insights. First, there is evidence that both groups improved over time. Second, at the beginning of the intervention there were not statistically significance differences between the groups. Third, there is evidence of a positive effect associated with the treatment over time, which, although it is not significant, indicates the expected direction of the effects that becomes stronger once controls are added.

The second estimated model covers the same three variables as the first one plus the controls covering the sociodemographic variables. This second model controls students' characteristics such as age, gender, frequency of school attendance, having internet in their houses, and schedule:

$$(2)\ score_{it} = \beta_0 + \beta_1 Time_t + \beta_2 Group_i + \beta_3 (Time_t * Group_i) + \beta_4 Gender_i + \beta_5 Age_i + \beta_6 Internet_i + \beta_7 Schedule_i + \beta_8 Frequency_i + \varepsilon_{it}$$

The results of the second estimated model can be seen in the second column of *Table 4*. The results confirm the first 3 insights of the previous model plus more information about the incidence of demographic variables of the students. According to the results, the gender and frequency of attendance have no significant effect on the difference of scores across time. On the other hand, been a year older is related to scoring 9 points less in the test and being in the morning or afternoon schedule is associated with scoring 16 points higher in the test. Similarly, having internet connection is associated with a rise in 11 points in the test scores.

Thanks to the second estimation we can confirm that students' characteristics may have a certain influence on the difference of the scores in time, but there is one special distinction that is yet to be made, and it's the effects of the school/campus on the performance of the students. The school may have important characteristics regarding infrastructure, teachers' performance, school environment that may impact the effect of the tutoring sessions on students.

Considering this the third estimated model includes the same covariates as model 2 but incorporates campus fixed effects ($\gamma_{it}$). It is important to mention that the fixed effects (FE) of the school are not included in the model due to the correlation between the schools and the campuses. A school may have 2 campuses but they covariate almost at the same time, so including both fixed effects in the model would cause almost perfect multicollinearity.

$$(3)\ score_{it} = \beta_0 + \beta_1 Time_t + \beta_2 Group_i + \beta_3 (Time_t * Group_i) + \beta_4 Gender_i + \beta_5 Age_i + \beta_6 Internet_i + \beta_7 Schedule_i + \beta_8 Frequency_i + \gamma_{it} + \varepsilon_{it}$$

As can be seen in column 3 of *Table 4*, the use of the campus fixed effects indeed improves model and captures differences across institutions. First conclusion taken from the estimation is that the intervention becomes marginally significant, showing moderate positive results, indicating that students in treatment showed a higher improvement (of 9 points on average) in comparison to the

control group over time. This suggests that the program's impact persists even after controlling demographic and institutional characteristics.

**Difference-in-Differences Models Summary**

| | *Dependent variable:* | | |
|---|---|---|---|
| | Student Score | | |
| | Simple DiD | DiD + Controls | DiD + Controls + Campus FE |
| | (1) | (2) | (3) |
| Post-treatment (time) | 39.909*** | 39.933*** | 39.933*** |
| | (4.214) | (4.114) | (3.816) |
| Treatment group (group) | 3.579 | 3.759 | 0.327 |
| | (4.465) | (4.374) | (4.096) |
| Diff-in-Diff (did) | 9.556 | 9.533 | 9.533* |
| | (6.314) | (6.158) | (5.712) |
| Male | | 4.871 | 6.594** |
| | | (3.106) | (2.915) |
| Age | | -9.247*** | -8.635*** |
| | | (1.641) | (1.584) |
| Internet access | | 11.228*** | 10.137*** |
| | | (3.584) | (3.368) |
| Morning schedule | | 16.796*** | -11.875* |
| | | (5.773) | (6.837) |
| Afternoon schedule | | 16.365*** | -14.062* |
| | | (5.918) | (7.983) |
| Low frequency | | -2.641 | -1.011 |
| | | (6.779) | (6.368) |
| Constant | 55.882*** | 103.082*** | 141.817*** |
| | (2.980) | (14.354) | (14.892) |
| Observations | 916 | 914 | 914 |
| $R^2$ | 0.185 | 0.233 | 0.349 |
| Adjusted $R^2$ | 0.183 | 0.225 | 0.333 |
| Residual Std. Error | 47.489 (df = 912) | 46.272 (df = 904) | 42.923 (df = 892) |
| F Statistic | 69.125*** (df = 3; 912) | 30.507*** (df = 9; 904) | 22.746*** (df = 21; 892) |

Standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, *p<0.1

*Table 4: Estimation results*

Regarding the other variables, the gender of the student turned out to be slightly significant related to the test score. The age continues being negatively correlated with the scores and surprisingly the schedule changed the direction of the effect they had on the score once the fixed effects were applied. The change in the schedule variable might be related to the fact that this characteristic is not actually related to the personal characteristics of the student, though it is more related to the institutions where they study, therefore, when applying the campus FE, the remaining effect the schedule had on the score changed.

The last important insight left from the last model is the effect internet connection may have in the scores across time, in both second and third model the estimated incidence of having internet connection was positive related with the scores, associated to 10-11 additional positive answers when the student had internet connection in their house.

## 7.1 Heterogeneity analysis

As an additional step in the analysis, it was interesting to see whether some types of students were more benefited by the program than others. Considering that the program aimed at students whose basic reading and mathematics skills were low, the students with lower scores at the beginning of the program should be the ones that were most benefited at the end of the intervention.

To test this hypothesis, a fourth model was estimated including the triple interaction between time, group and an indicator for students scoring below 30% (84 pts) in the pre-test. The 30% threshold was selected as the main heterogeneity specification and is related to the design of the program, which established that only students that scored below 60% would be considered for participation in the tutoring sessions. Additionally, additional cut-offs were explored as complementary robustness checks to assess the sensitivity of the results to alternative baseline performance (25%, 35% and 40% thresholds were tested). The fourth model was estimated as follows:

$$(4)\ score_{it} = \beta_0 + \beta_1 Time_t + \beta_2 Group_i + \beta_3 (Time_t * Group_i) + \beta_4 Low84_i + \beta_5 (Time_t * Low84_i) + \beta_6 (Group_i * Low84_i) + \beta_7 (Time_t * Group_i * Low84_i) + \beta_8 X_i + \gamma_{it} + \varepsilon_{it}$$

In addition, a fifth model was estimated considering the threshold of 60 pts (21,3%) as an exploratory robustness exercise. Visual analysis of the score distribution in the density and violin plots suggested a possible threshold at this point, motivating the additional specification to evaluate whether the effects differed under this limit. The fifth model was estimated as follows:

$$(5)\ score_{it} = \beta_0 + \beta_1 Time_t + \beta_2 Group_i + \beta_3(Time_t * Group_i) + \beta_4 Low60_i + \beta_5(Time_t * Low60_i) + \beta_6(Group_i * Low60_i) + \beta_7(Time_t * Group_i * Low60_i) + \beta_8 X_i + \gamma_{it} + \varepsilon_{it}$$

In both models the term of interest is the triple interaction term $\beta_7(Time_t * Group_i * LowL_i)$ which measures the heterogeneous treatment effect for students identified as low performers according to each cutoff point. A significant negative effect means that the intervention had smaller relative gains among low performing in comparison to higher-performing students, on the other hand, a positive significant coefficient indicated that low performing students benefited more from the intervention. The coefficient $\beta_3(Time_t * Group_i)$ captures the average treatment effect of the intervention just as the previous three models. Likewise, the terms $\beta_8 X_i + \gamma_{it}$ stand for controlling school and student's characteristics.

*Table 5* shows the results of the estimation. The results indicate that the intervention had a stronger differentiated effect when using a threshold of 84 pts in the pre-treatment score. The triple interaction term was negative and marginally significant, meaning that the students below the cut-off experienced smaller relative gains from the intervention in comparison to students above the threshold. When analyzing results under the 60-pts cutoff the estimation remained negative but lost significance, meaning that the improvement of students regarding that threshold was less differentiated.

These results suggest that extremely low performing students may need additional support or a longer exposition to the treatment to catch their peers' level and be equally benefited from the intervention. Moreover, the program appears to be beneficiating more students who are struggling with their reading and mathematics abilities but that have some previous basic skills and are not at

the bottom of the performance distribution. The controls estimations were omitted from the table as they do not vary between both models and are not the main interest of the heterogeneity analysis.

**Heterogeneity Analysis: Comparison of Thresholds (<84 vs <60)**

| | *Dependent variable:* | |
|---|---|---|
| | Student Score | |
| | Low Performers <84 | Low Performers <60 |
| | (1) | (2) |
| Post-treatment (Time) | 35.362*** | 40.041*** |
| | (5.410) | (4.368) |
| Treatment group (Group) | 5.087 | -0.510 |
| | (5.884) | (4.567) |
| Low Performers | -66.665*** | -65.731*** |
| | (4.564) | (4.064) |
| Diff-in-Diff (Time × Group) | 21.601*** | 14.695** |
| | (8.164) | (6.370) |
| Time × Low Performers | 6.284 | -0.176 |
| | (6.343) | (5.581) |
| Group × Low Performers | -3.833 | -3.984*** |
| | (6.867) | (1.140) |
| Triple Interaction (Time × Group × Low Performers) | -16.481* | -9.012 |
| | (9.544) | (8.284) |
| Observations | 914 | 914 |
| $R^2$ | 0.645 | 0.671 |

Robust standard errors in parentheses. Both models include campus fixed effects. *** p<0.01, ** p<0.05, * p<0.1

*Table 5: Heterogeneity analysis*

## 8. Discussion and comparison with other studies

The estimates obtained in this analysis are aligned with several studies that show moderate positive effects of tutoring on early grade reading outcomes. For example, Nickow, Oreopoulos, and Quan (2020) found an average impact of 0.37 standard deviations (SD) in achievement across K-12 tutoring interventions. Another meta-analysis that serves as example is the one from Elbaum, Vaughn, Hughes, & Moody, (2000) which reported effects around 0.41 SD in their meta-analysis of reading interventions. Considering that the standard deviation of the pre-test was 40, the estimated effect calculated of Aula Global corresponded to 0.24 SD, which can be considered as consistent and moderated according to the cited literature. Although the magnitude of the effect is not as large as other studies, the results are consistent in the addition of controls and point in the same direction as the studies mentioned.

It is important to highlight the similarity between our results and previous studies of the same program. The results of 0.24 SD are almost the same as the 0.26 SD identified by Barrera and Lagos (2018) when they evaluated the program 7 years ago, regarding the methodological changes made to the program during the years we can conclude that the intervention shows consistent effects and is still relevant.

The difference between the effects reported in the other studies and the one of Aula Global may be related to the contextual characteristics of the studies and could be related to the intensity of the intervention or differences in the measurement of the effects. Nonetheless, the direction and the magnitude of the estimated effects are the similar to the other studies and validate the conclusions obtained in this analysis.

## 9. Limitations and next steps

It is important to notice that no matter how strong the design of this study was there were some limitations in the data presented. The fact that there was no information available about previous development of the students didn't allow to check the parallel trends assumption. Moreover, the available students' information left out of the analysis some important characteristics like parents' education or household income which has proven to have a strong incidence on the learning outcomes of students. In addition to this, it is important to recognize that maybe with a bigger sample size there would have been easier to have more significant results.

If this work continues it would be interesting to follow the students across the years to check if the ones in the treatment groups remained with better grades, showed more persistence in school or have access to better universities or scholarships in the future. Future analyses could benefit from incorporating longitudinal data and richer background information on student's households.

## 10. Conclusions

The analysis developed provides empirical support for the program, as it indicates that tutoring sessions had a moderate positive effect on the reading and mathematics skills among second grade students. The effect was consistent with similar interventions, and with previous studies conducted on the same program.

The final effect of the program was relatively more beneficial for students who were struggling with their reading and mathematics skills but that had, at the same time, some basic previous abilities developed. The students who showed poor skills at the beginning of the intervention tended to benefit less from the program than their peers. This suggests heterogeneous effects depending on baseline performance within the observed sample.

The results highlight the role of the school environment and the students' characteristics in the learning process of children. The analysis showed that there were significant differences among

the campuses included in the program and the demographic characteristics also proved to be significant in analyzing the results.

The results also provide insights into the design of similar programs. First, a single intervention cannot be expected to have extremely high and significant results in such a short period of time, as these results identified effects within a significant but moderate range.

Second, as mentioned, student performance depends largely on the school environment and the socioeconomic conditions in which they live. Therefore, it is essential that future programs intended for this population include lines of action that address the students and their families from different perspectives. It is crucial that interventions analyze the school context where the student learns and consider that the heterogeneity of the population can determine the degree to which a student can benefit from a policy.

The results highlight the need to continue developing strategies that measure the effectiveness of interventions and that collect information to allow for more comprehensive and robust evaluations. In this case, the limited information on student performance in several periods before the test made it difficult to visualize student skills trends over time. Similarly, the available individual information did not include key aspects such as parental income level or educational attainment; these variables have proven to be highly significant covariates of students' learning abilities. Therefore, it is recommended that they be considered and included in the baseline data for future similar studies.

Developing assessments that allow for the evaluation of learning trends is especially important in the context of this study. As stated in the introduction, various global organizations have warned about the high levels of Learning Poverty in the region. This problem was exacerbated by the global COVID-19 pandemic, and today, more than ever, all learning recovery strategies developed in Latin American countries are relevant (World Bank, 2019).

## 11. Bibliography


Andersen, S., Humlum, M., & Nandrup, A. (2016). Increased instruction time and student learning: Evidence from a natural experiment. *The Scandinavian Hournal of Economics*, 28(5), 3-22.

Angrist, J., & Pischke, J. (2009). Mostly harmless econometrics: An empiricist's companion . *Princeton University Press*.

Barrera, F., & Lagos, F. (2018). *Tutoring, Professional Development, and Educational Improvement: Evidence from Cali, Colombia.* Harvad University.

Becker, G. S., & Chiswick, B. R. (1966). Education and the Distribution of Earnings. *American Economic Review, 56(1/2)*, 358-369.

Bellei, C. (2009). Does lengthening the school day increase students academic achivement? Results from a natural experiment in Chile . *Economics of Education Review*, 28(5), 629-640.

Cortes, K., Goodman, J., & Nomi, T. (2015). Instensive math instruction and educational attainment: Long-run impacts of double-dose algebra. *Journal of Human Resources*, 50(1), 108-158.

Cunha, F., & Heckman, J. (2007). The technology of skill formation. *American Economic Review*, 92(2), 31-47.

Elbaum, B., Vaughn, S., Hughes, M., & Moody, S. (2000). How effective one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research. *Journal of Educational Psychology* , 92(4), 605-619.

Heckman, J., & Mosso, S. (2014). The economics of human development and social mobility. *Annual Review of Economics*, 6, 689-733.

Imbens, G., & Wooldridge, J. (2009). Recent developmnets in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1), 5-86.

Lavy, V. (2015). Do differences in school's instruction time explain international achivement gaps? Evidence from developed and developing countries . *The Economic Journal*, 125(588), F397-F424.

Litsching, S., & Morrison, K. (2013). The impact of Intergovernmental Transfers on Education Outcomes and Poverty Reduction. *American Economic Journal: Applied Economics, 5(14)*, 206-240.

Manuelli, R., & Seshari, A. (2014). Human Capital and the Wealth of Nations. *American Economic Review, 104(9)*, 2736-2762.

Nickow, A., Oreopoulos, P., & Quan, V. (2020). The Impressive Effects of Tutoring on PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence. *National Bureau of Economic Research*.

Robinson, C., Pollard, C., Novicoff, S., White, S., & Loeb, S. (2024). *The Effects of Virtual tutoring on Young Readers: Results from a Randomized Controlled Trial. EdWorkingPaper.* Annenberg Institute for School Reform at Brown University.

Roschelle, J., Cheng, B., Hodkowski, N., Neisler, J., & Haldar, L. (2020). *Evaluation of an online tutoring program in elementary mathemathics.* San Mateo, CA.

Tulcan Tapia, M., Valencia Daza, L., & Hernández Cruz, A. (2023). *Aula Global, a Groundbreaking Initiative for Learning Recovery. Childhood Education, 99 (4)*, 6-13.

UNESCO, World Bank, Unicef and OECD. (2022). *From learning recovery to education transformation. Insights and Reflections from the 4th Survey on National Education Responses to COVID-19 School Closures.* UNESCO Publishing.

UNICEF. (2021). *COVID-19 and learning losses: Rebuilding education systems*. UNICEF.
World Bank. (2019). *Ending Learning Poverty: What Will It Take?* Retrieved from World Bank.