IDB | Invest

Development through
the Private Sector Series
July 2021

TN
No. 27

Building Socioemotional Skills by Playing:

Experimental Evidence from Brazil

Authors:
Viviane Azevedo
Maria Laura Lanzalot
Ricardo Paes de Barros
Rodolfo Stucchi
Patricia Yañez-Pagans

IDB | Invest

Building Socioemotional Skills by Playing:

Experimental Evidence from Brazil

Cover page design: David Peña Blanco

July 2021

Building Socioemotional Skills by Playing: Experimental Evidence from Brazil *

Viviane Azevedo[†] Maria Laura Lanzalot[‡] Ricardo Paes de Barros[§]
Rodolfo Stucchi[¶] Patricia Yañez-Pagans[‖]

July 13, 2021

Abstract

School programs are increasingly promoting the development of socioemotional skills given widely accepted evidence on the importance of these skills for academic performance and employment. Educational systems have also seen an increase in the use of games to promote learning, but there is still limited rigorous evidence of their causal impacts. We conduct a large-scale randomized control trial looking at the impacts of the MindLab program in Brazil, a well-known methodology designed in Israel and adopted in over 21 countries, that uses games to promote learning. We quantify impacts on reasoning, metacognition, socioemotional and academic outcomes focusing on a sample of fifth-grade students. Results show no significant program impacts at the average level, but highlight large heterogeneities across participating municipalities, with some reporting large and positive effects on reasoning and academic outcomes, while others report lower socioemotional results and academic performance. Multiple robustness checks confirm the results. Administrative program data and interviews with school professionals suggest the importance of program design and teacher engagement with the methodology to produce a positive change in students.

JEL Classification: H52; I20; I24; I28.
Keywords: Education; Socioemotional skills; Gamification

*We would like to thank MindLab for giving us the opportunity to evaluate their innovative education program. We are also grateful to Instituto Ayrton Senna at Insper for their collaboration in the design and implementation of this evaluation in the field. Thank you to LACEA conference participants and Gabriela Aparicio for feedback provided to previous versions of this work. The views expressed in this paper are those of the authors and do not represent those of the Inter-American Development Bank or IDB Invest, its respective Board of Executive Directors or the countries they represent.

[†]Development Effectiveness Division, IDB Invest. vivianea@iadb.org.
[‡]Development Effectiveness Division, IDB Invest. mlanzalot@iadb.org.
[§]Insper. ricardopb@ias.org.br.
[¶]Development Effectiveness Division, IDB Invest. rstucchi@iadb.org
[‖]Development Effectiveness Division, IDB Invest. patriciaya@iadb.org.

1 Introduction

Social and emotional learning (SEL) has become increasingly important in schools. SEL is the process of acquiring the skills to recognize and manage emotions, develop caring and concern for others, make responsible decisions, establish positive relationships, and handle challenging situations effectively (Konishi & Wong, 2018). Multiple studies have shown that SEL can have positive impacts on students' academic achievement, therefore increasing their chances of success both in school and later in life (McCormick *et al.*, 2020; Corcoran *et al.*, 2018; Durlak *et al.*, 2011; Sklad *et al.*, 2012). A study by Bassi *et al.* (2016) shows that besides the academic skills traditionally taught in schools employers are increasingly looking for socioemotional skills, such as critical thinking, responsibility, teamwork, and problem solving, among others. Given these findings, a large number of programs that aim to develop socioemotional skills have been developed over the last decades, and various private sector institutions or providers have emerged to assist educational systems in offering these programs. From a development point of view, as low-income children are at increased risk for socioemotional problems, SEL programs may be key to reducing socioeconomic disparities (Mondi & Reynolds, 2020).

In recent decades, the educational field has also seen an increase in the use of game-based elements to promote desired behaviors and learning, known as gamification. The interactional nature of games makes them especially suitable for delivering SEL programs (Hromek & Roffey, 2009). Games use strategies such as discussion, role-play, and problem solving to engage players allowing them to learn and practice various socioemotional skills. Despite their advantages, there is still little rigorous causal evidence on the impacts of gamification on learning. A recent review of the literature conducted by Zainuddin *et al.* (2020) shows that gamification has positive impacts on motivation and learning. However, the majority of studies have been conducted on adult learners or higher education students, with little evidence from primary or secondary schools. Moreover, most evidence is connected to high-tech environments but insights are still needed on how these approaches could work in unsophisticated technological conditions, which mostly pertain to those observed in developing countries. The literature has also pointed out that games may bring unintended consequences, such as increased anxiety and lower collaboration (Araya *et al.*, 2019), or lower performance in some assignments (Dominguez *et al.*, 2013). These findings suggest that gamification might not work equally well for all students and that important efforts are required for the design and implementation of these strategies (Dominguez *et al.*, 2013).

We conduct a large-scale randomized control trial looking at the impact of game-based approaches applied to SEL programs and their effects on the development of socioemotional

skills and academic learning. In particular, we evaluate the MindLab program that aims to promote the development of cognitive, social, emotional, and ethical skills in students. The program is based on three pillars: (i) the use of reasoning games as a didactic resource, (ii) the construction of strategies and methods that help to organize thoughts and actions, and (iii) having the teacher as a mediator who conducts the process and intentionally promotes reflections on how to apply learning in everyday situations, thus expanding self-knowledge and reflective and critical thinking. So far, the MindLab method has been adopted by kindergartens and schools in over 21 countries, more than 15,000 teachers have been trained and certified, and over 4 million students have been exposed to the program. Our study is conducted in Brazil where the MindLab program[1] has already been applied in over 1,000 schools reaching more than 350,000 students (MindLab, 2018). In 2017, IDB Invest provided a loan to MindLab Brazil to further increase access to innovative educational methodologies in public and private schools in the country. The company expected to reach more than 160,000 additional students over six years, most of them living in low-income households.

Brazil is a unique context of study. Recent statistics show that 98% of children between the ages of 5 and 14 were enrolled in school in 2016, an enrollment rate similar to the OECD average.[2] Despite the high enrollment rate, the 2015 PISA results showed that the average performance of students in Brazil was significantly below OECD averages.[3] There is also a wide gap between the scores of the best-performing students and the lowest-performing students, reflecting inequality of access to quality education and opportunities (Bourguignon *et al.*, 2007). The challenge of low-quality education is further exacerbated for children who attend public schools. In 2015, 87% of all Brazilian students enrolled in primary education attended public schools, of which over 80% were low-income.[4] The 2015 IDEB (Índice de Desenvolvimento da Educação Básica), which measures the quality of basic education on a scale from 0-10, averaged 4.2 for the last years of public primary schools versus 6.1 for private schools. There are also wide disparities across regions, with the lowest results observed in the less-developed Northeast region of the country (IDEB of 3.5) and the highest ones in the Southeast region (IDEB of 5.1).[5] Among the obstacles identified in public education are: inadequate use of classroom time, lack of adequate study materials, high rates of teacher

[1] The MindLab program is called MenteInovadora in the case of Brazil.

[2] For more information please check Education at a Glance 2018, Brazil: https://www.oecd-ilibrary.org/docserver/eag-2018-en.pdf?expires=1546113290&id=id&accname=ocid194302&checksum=7881DAA0035FFDDB525A50AFF57A0276

[3] For example, in science (401 vs. 493 points), reading (407 vs. 493 points) and mathematics (377 vs. 490 points).

[4] IBGE, Síntese de Indicadores Sociais, Uma Analise das Condições de Vida da População Brasileira, Figure 12. 2018: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101629.pdf. As a proxy for BoP, the first 3 income quintiles were used.

[5] http://ideb.inep.gov.br/resultado/resultado/resultadoBrasil.seam?cid=574317

absenteeism, and ineffective teaching methods (Sandoval, 2012).

Brazil also has the third highest rate of young adults who are not in education, employment or training (NEETs) in Latin America and the Caribbean (after Mexico and El Salvador), meaning that 11.1 million young adults from a total of 48.5 million in the country are in vulnerable situation. According to Novella *et al.* (2018), the combination of high technological and socioemotional skills raises hope regarding how these youth can face workforce challenges in the future.

Our experimental evaluation is restricted to a sample of 56 schools (28 treated and 28 controls) in four municipal networks and focuses on a sample of 2,532 fifth-grade students. Given the fairly small number of clusters and to reduce efficiency loss, a matched-pair cluster randomized experiment was designed. Since the number of schools in each network was more than twice the number of places available for the MindLab program, schools were paired in each network based on their similarity in baseline educational characteristics. Then a lottery was drawn for each pair, randomly choosing one school to participate in the program and the other to be in the control group. The study was conducted with local experts in education to identify and develop a series of tests aimed at measuring reasoning, executive functions, metacognition, and socioemotional abilities. In addition, we explore impacts on academic achievement by looking at Brazilian standardized tests. Surveys from parents and teachers were also collected to better understand their perceptions about the changes brought by the program. We use these surveys to check the robustness of our results and to explore some of the mechanisms driving the observed effects.

Results show, on average, no significant program impacts on reasoning, metacognition, or math and Portuguese scores. We also observe negative effects, although small, on children's socioemotional outcomes. The treatment group's test scores on this dimension are 15% of a standard deviation lower than those of the control group. Results in previous SEL evaluations have been diverse. Durlak *et al.* (2011) summarize the results of 213 SEL program evaluations (with almost half of them being experimental studies) finding, on average, 57% of a standard deviation on socioemotional skills (and between 20% to 30% on positive attitudes, positive social behaviors, fewer conduct problems, and lower levels of emotional distress) in programs with no implementation problems.[6] More recently, however, Corcoran *et al.* (2018) conclude that for higher quality randomized control trials with larger samples effect sizes range from -14% to 73% for reading, from -22% to 81% for math, and -2% for science. The highest impacts

[6] It is important to note that more than half of the programs evaluated in this meta-analysis present similar characteristics to the MindLab program: they are applied by the same teacher that already works with the classroom (53%); the program duration is less than a year (77%); and they are offered to students in the first phase of elementary school (56%). However, the authors did not find any difference regardless of who implemented the treatment (i.e., the class teacher or another professional) or the level of the application (i.e., programs that involve the entire school community or just one classroom).

were observed in very comprehensive programs, including a school-wide climate development program, counselor's program, parent program, and community program, among others.[7]

Interestingly, our study finds important heterogeneity across municipalities, with one municipality showing robust positive results for reasoning tests (49% of a standard deviation) and academic tests (over 50% of a standard deviation), while other municipalities present negative and significant effects in socioemotional outcomes (around 27% of a standard deviation) and academic tests (around 42% of a standard deviation). We explore some of the mechanisms that may be driving these results and find that some school, teacher, and student' characteristics seem not to play a role in explaining the differential effects. However, administrative program data and surveys with school staff reveal the importance of adequate program implementation and of school teachers and staff having positive views and feeling engaged with these new methods. In successful cases, teachers were clearly involved and had positive perceptions of the program and its methods. In cases where the program did not reach the proposed objectives, opposite views were evident. This goes in line with other studies in the literature such as Durlak *et al.* (2011) that show that implementation problems moderate program outcomes.

Multiple robustness and placebo tests confirm our results and the validity of the identification strategy. Overall, results highlight the potential positive effects of introducing gamification in SEL programs and their impacts on building socioemotional skills and improving academic learning. However, the results also highlight the importance of program design and implementation, with careful consideration of the heterogeneity across settings. Qualitative evidence suggests that there are important implications on how the methodology is being explained and transferred to teachers during program development.

This study makes three important contributions to the literature. First, it contributes to the still very limited rigorous causal evidence looking at the impacts of gamification on learning. As shown by Zainuddin *et al.* (2020), most studies have focused on adult learners or higher education students and high-technology settings. Of those focused on primary or secondary school children, the majority have used mixed methods with reduced sample sizes or qualitative approaches. A recent experimental paper by Araya *et al.* (2019) shows that gamification increased math learning for primary-level students in the ConectaIdeas program in Chile, but at the same time it increased math anxiety and reduced preferences to collaborate in teams. Second, to the best of our knowledge, this is the first rigorous causal study measuring the effects of gamification in SEL programs. In addition, this study provides evidence on the impacts of these programs on low-income students enrolled in primary education and in a low-technology environment, thus results may be relevant to guide

[7] This was the case of the evaluation of the Positive Action program in the United States.

policy-making in developing countries. Finally, this is the first experimental evaluation of the MindLab program, which has been broadly adopted across the world. This program has had numerous quasi-experimental evaluations in the past showing positive results (Garcia & Abed, 2009; Garcia *et al.*, 2011; Garcia & Abed, 2013; Azevedo *et al.*, 2019), but as Corcoran *et al.* (2018) indicate non-randomized studies of SEL interventions overstate effect sizes, therefore the experimental design presents a more accurate view of the potential impacts of this program.

The rest of the paper is structured as follows. The next section presents a detailed description of the MindLab program and discusses the results from previous evaluations of this program. Section 3 explains the empirical strategy and describes the outcomes and data sources used in the analysis. Section 4 presents the results and Section 5 explores the different mechanisms that could be driving the results. Section 6 presents the robustness and placebo tests and Section 7 presents the main conclusions and discusses the implications of our findings.

2 The MindLab Program

2.1 Program description

The MindLab Group was founded in 1994 in Israel and has since expanded around the world. The method has been adopted by thousands of kindergartens and schools in over 21 countries, including the United States, Portugal, England, Spain, Italy, Hungary, Turkey, Hong Kong, China, and Japan. MindLab Group estimates that more than 15,000 teachers have been trained and certified and over 4 million students have been exposed to the program. The program is based on strategy games that can be used as an educational tool to improve socioemotional and cognitive skills. The games aim at creating an awareness of the "thinking processes", promoting a stronger emphasis on the development of skills such as critical thinking, problem solving, teamwork, and effective communication.

MindLab's methodology is based on a tripod composed of thinking games, metacognitive methods, and the teacher as the mediator. It proposes a curriculum organized into activities designed to develop cognitive, emotional, ethical, and social skills from kindergarten to high school. Reasoning games act as pedagogical resources that simulate concrete everyday situations in a playful context. Metacognitive methods are internal tools that organize thoughts and actions to play (and act in everyday situations) more consciously, autonomously, and responsibly. Finally, the mediator teacher plays an important role in the learning process and in the development of students' socioemotional, cognitive, and executive functions, promoting

reflections that help students build self-knowledge and self-control. Teachers are trained in the theoretical aspects of the methodology and reflect on practical experience with specialists throughout implementation of the program.

In 2006 a Brazilian company, "MindLab Brasil" (MindLab, henceforward), developed a program called "MenteInovadora" that adapts the MindLab Group's method to the Brazilian context. After over ten years of operation in Brazil, the MenteInovadora program has been applied in both public and private schools from pre-school to high school, covering all phases of basic education. Starting with 33 schools in 2010, the company has expanded to over 1,000 schools reaching more than 350,000 students across the country (MindLab, 2018).

MindLab has developed or adapted about 70 reasoning games to the Brazilian context. The curriculum is structured progressively by semester to develop specific skills in students according to their age group and development needs for students ages 4 to 18. Every semester, the teachers receive manuals which explain and structure every single lesson in detail in order to support the development of the set of skills of the curriculum for the respective semester. MindLab offers an initial in-person teacher training that lasts 16 hours, followed by a 14-hour distance training, and eight monthly sessions that last two hours each.

The typical class begins with the teacher explaining the strategy game used during the session and the goals for the class. The students then engage in playing the games in small groups, while the teacher acts as a facilitator and helps them apply different "thinking models", such as identifying the key problems and planning strategies before applying them, or recognize the need for cooperation. At the end of the session, the facilitator and the students discuss the different strategies applied, how they may apply them in real-life situations, and how the students could act differently in these situations based on the strategies they practiced. The program lasts 50 minutes per week in a 34-week curriculum, representing less than 5% of the traditional curriculum of 800 classroom hours per school year.

2.2 Previous evaluations of the program

MindLab Group programs have been repeatedly evaluated over more than a decade globally. The first two assessments were carried out by Green & Gendelman (2003, 2004) in Israel. Green & Gendelman (2003) evaluate the impact of this methodology on the development of strategic thinking, resource management and planning. Using a sample of 195 students from eight classrooms in five Israeli schools - most of whom (70%) were in their fifth year of elementary school- they tested the students' pre-treatment abilities and randomly assigned the students to the treatment and control status balancing those characteristics. The results show that the methodology had a strong positive impact on students' strategic thinking (20% to 60% of a standard deviation) and, on resource management and planning.

In their second evaluation, Green & Gendelman (2004) estimate the impact of 4 classroom hours of the methodology on strategic reasoning and mathematics and language performance. With a sample of 35 students from two classrooms (19 students in the treated classroom and 16 in the control classroom) in their third year of elementary school in one Israeli school, they found positive impacts on strategic reasoning and language (more than 50% of a standard deviation); and a positive but not statistically significant effect on math. A third international evaluation was carried out by Ajello *et al.* (2012) in Italy, with 300 students in the second and forth year of elementary school. The authors contrasted the problem-solving and decision-making process of students who participated in a program that used the MindLab methodology and another group that did not have access to this program.[8] They found that those who participated in the program more often adopt a more reflective attitude, while students who did not participate in the program quit more often or chose more impulsive solutions.

The Brazilian version of the program has been evaluated annually from 2009 to 2013. These evaluations were conducted by renowned Brazilian educational research institutes in partnership with company staff, but none were randomized control trials. The first national study was conducted by the Institute for Educational Evaluation and Development (INADE) in partnership with MindLab (Garcia & Abed, 2009). This study was carried out with students in their fifth year of elementary education from 10 public and private schools, and evaluated the impact of three months (12 classes) of the program on student performance in mathematics and Portuguese before and after the treatment. They estimated that completing one quarter of the program has an impact of 10% of a standard deviation in mathematics and of 2% in Portuguese. In 2010, they expanded the universe of assessment, using a sample of 3,000 students in 50 public and private schools, finding much more favorable results. The magnitudes obtained indicate that completing one quarter of the program has an impact of 17% of a standard deviation in mathematics, 40% in Portuguese, and 14% in natural sciences (Garcia & Abed, 2010). In 2011, with a sample of 9,000 students and almost 150 schools, the results were between those obtained in the two previous studies (Garcia *et al.*, 2011).[9] Although the timeframe of the evaluation is short, given that results are based on before and after comparisons they do not isolate any changes that may have occurred over time aside from the implementation of the MindLab program.

In 2012, with about 12,000 fifth and ninth grade students (5,000 and 7,000 students, respectively) from nearly 200 schools, they perform the same type of assessment (comparing the performance before and after the treatment) for each grade separately. For fifth grade

[8] It is not clear from the paper what methodology was followed to find comparable units.

[9] Completion of one quarter of the program had an impact of 12% of a standard deviation in mathematics, 12% in Portuguese, and 14% in natural sciences.

students they added a control group using a propensity score approach (Garcia *et al.*, 2012). The magnitude of the results obtained in this fourth study are lower than those obtained in the three previous studies, which did not have a comparison group. The progress among students in the schools that had been participating in the program for more than a year was 13% of a standard deviation higher in mathematics and 6% higher in Portuguese. No impact was found for students in schools that had recently joined the program. This fourth national study also evaluated the program's impact on five socioemotional skills: ethics, empathy, autonomy, self-efficacy and impulsivity.[10] Findings show that the program had a substantial impact on student self-efficacy and autonomy, some impact on impulsiveness and ethical sense; and no significant effect on empathy.

The final annual assessment returns to the before and after approach and was conducted only for sixth grade students from 45 private schools (Garcia & Abed, 2013). The results are consistent with those obtained in previous studies, but with smaller magnitudes.[11]

Azevedo *et al.* (2019) is the first evaluation of the MindLab program performed by an independent team. The study is a non-experimental evaluation that exploits the expansion of the program since 2010, the annual monitoring of grade completion rates by the School Census, and the biannual measurement of student proficiency at the end of fifth and ninth grades by the standardized Prova Brasil test to assess the impact of the MindLab program on grade completion rates, and math and Portuguese proficiency. The study finds heterogeneous effects depending on the comparison groups used, the period considered in the analysis, and the grade evaluated (fifth or ninth). In terms of grade completion rates they find a positive effect in the fifth grade (2 percentage points), but no effect in the ninth grade. Nevertheless, when they assess the effect on math and Portuguese proficiency, they find positive evidence only in the ninth grade (close to 15% of a standard deviation).

The high variability in the results observed could be due to the different methodologies and sample sizes used. Results could also be sensitive to the way the program has been adopted and implemented by schools and teachers. Our study provides additional evidence on the MindLab program based on a large-scale experimental evaluation. As such, our results are subject to less restricted identification assumptions but cannot be generalized out of sample. The paper also focuses on measuring multiple outcomes related to soft skills, such as reasoning, metacognition, socioemotional and executive functions. These outcomes were measured using tests specifically designed by local experts in education.

[10] Measures for each of these skills were obtained from students' self-perceptions (self-reports).

[11] Completion of one quarter of the program had an impact of 10% of a standard deviation in math and 3% in Portuguese.

3 Methodology & Data

3.1 Randomized Design

In 2017, MindLab sold their method to 450 public schools in 13 municipal networks in Brazil. Of those, eight networks (comprising 429 schools) were invited to participate in a randomized control trial (RCT) to test the effectiveness of the program (the five networks left behind were already exposed to the treatment in the past or had a small number of schools). For a variety of reasons,[12] most of the available spots for the program were already allocated to schools prior to the RCT design. Only 72 (17%) out of the 429 schools could be allocated randomly across all eight educational systems that joined the RCT. In those municipalities the majority of schools were going to implement the program for fifth grade students. Therefore, the evaluation was restricted to fifth grade students and we ended up working with five municipalities. However, one of these five municipalities was left out of the evaluation because the school year was delayed due to a general strike. Therefore, the experimental study was performed with a sample of 56 schools (28 treated and 28 controls) in four municipal networks. In addition, as the amount of fifth grade groups (classes) per school varied significantly in our 56 schools, and given time and financial constraints, we limited the survey to only two fifth grade groups per school. In the schools with more than 2 groups in fifth grade, the evaluated groups were randomly selected from all available groups.

Given the fairly small number of clusters and to reduce efficiency loss, a matched-pair cluster randomized experiment was designed (Imai *et al.*, 2009; Bruhn & McKenzie, 2009).[13] The number of schools in each network was twice the number of spots available for the MindLab program, thus the schools were paired in each network using the degree of similarity between them and then a lottery was drawn for each pair randomly choosing one school to participate in the program and another to be its counterfactual. Since the number of vacancies was defined by each municipal network and varied across networks, a different number of draws were made in each network based on the spots available. Table 1 shows the number of treated and control schools in each network.[14]

[12] Some restrictions encountered were: previous exposure to the MindLab program and discretionary reasons.

[13] It is important to mention that Imai *et al.* (2009)'s matched pair design may well have more statistical power than the unmatched cluster randomization design even if one has only three matched pairs.

[14] To protect the privacy of the schools that participated in the evaluation, we do not provide the name of the municipalities that were treated by MindLab as part of this study.

Table 1: Schools Evaluated - By Municipality

Status	Muni. #1	Muni. #2	Muni. #3	Muni. #4	Total
Treated	5	13	5	5	28
Control	5	13	5	5	28
Total	10	26	10	10	56

Source: MindLab

To reduce the variability of the estimated impacts and increase statistical efficiency, the information used to form the pairs should include the best possible predictors of the outcomes of interest in the absence of the treatment. The outcomes of interest for the program, which we explain later in more detail, consist of measures of reasoning, executive functions, metacognition, and socioemotional abilities, all of which were not observable at the time of the draw. However, historical information on outcomes of indirect interest and that may be correlated, such as math and Portuguese performance and grade completion rates, are available for all public schools in the country for the fifth and ninth year of elementary school and for the third year of high school every two years for the period 1995-2015. Therefore, the matching process was based on these indirect results.

More specifically, pairs were matched on a predictor of the progress in the IDEB for years prior to the intervention (i.e., between 2015 and 2017). The IDEB is a synthetic indicator that incorporates information on student proficiency and grade completion rates. This predictor was estimated as a linear function of the IDEB in earlier years and of a variable capturing the socioeconomic level of students in the school, called the INSE.[15] Therefore, in a first step, we estimate the following equation:

$$\Delta IDEB_{s,(2015,2013)} = \beta_0 + \beta_1 INSE_{s,2013} + \beta_2 IDEB_{s,2013} + \beta_3 IDEB_{s,2011} + \mu_s \qquad (1)$$

Where $\Delta IDEB_{s,(2015,2013)}$ is the change in the $IDEB$ index for school s between years 2013 and 2015. $INSE_{s,t}$ and $IDEB_{s,t}$ are the specific year t values for INSE and IDEB in school s. In a second step we use the estimated coefficients from this regression to predict the change in IDEB between 2015 and 2017, using as covariates the updated values for the

[15] INSE is an index of the socioeconomic level of Brazilian schools created in 2014 by the National Institute for Educational Studies and Research "Anísio Teixeira" (INEP, for its acronym in English). INSE's objective is to contextualize school performance in the national exams, taking into consideration external factors that can affect academic performance, such as students' household income and parents' schooling level. For more information please see: http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/indicador-de-nivel-socioeconomico-das-escolas-de-educacao-basica-inse-2015-e-publicado-pelo-inep/21206

IDEB in 2013 and 2015.

After forming the 28 pairs of schools by similarity (i.e., closer $\Delta\widehat{IDEB}_{i,(2017,2015)}$), we randomly select in each pair a school to participate in the MindLab program. To ensure the transparency and integrity of the draw, the treatment was assigned in a public raffle, in which one school representative for each pair of schools removes a ball from an opaque urn containing two balls (yellow and orange), and receives the envelope with the same color as the ball selected in the raffle. Whether or not the school will participate in the program depends on the color of the selected ball and the card inside the envelope that says "Treatment" or "Control". The card's text inside the envelope was also randomly assigned. Therefore, the public raffle does not fully determine the chances of participating in the program, but it guarantees the transparency of the randomization.

3.2 Baseline Data

Brazil has rich educational data, which allows us to obtain information on the quality of the education system, socioeconomic characteristics of the students, characteristics of the schools and the teachers. In particular, we use four sources of information, all from the National Institute for Educational Studies and Research "Anísio Teixeira" (commonly known as INEP) for the baseline year 2015:

(i) Prova Brasil is a national education assessment administered every two years. It consists of standardized tests and socioeconomic questionnaires applied to kids in the fifth and ninth grade of elementary school and the third year of high school. The tests cover Portuguese, with a focus on reading, and math, with a focus on problem solving. In the socioeconomic questionnaire, students provide contextual information about factors associated with their development, such as family background, parents' education levels, and family income. Teachers and school principals also complete questionnaires that gather demographic data, professional profiles, and working conditions. It is important to clarify that we do not have a baseline of the same cohort of students who were treated in 2017, but rather we have a baseline of their schools' fifth grades performances.

(ii) The School Census collects data on a yearly basis regarding school infrastructure characteristics, school enrollment levels, grade completion rates, grade repetition, and dropout rates.

(iii) The School Socioeconomic Index (INSE) is an indicator constructed by the INEP to contextualize school performance in national exams, taking into consideration external factors that can affect academic performance, such as students' household income and

parents' schooling levels. The students are grouped in eight ordinal levels and INSE is calculated as a simple arithmetic mean of students' socioeconomic level assessment.[16]

(iv) The Basic Education Development Index (IDEB) is a synthetic indicator that incorporates information on student proficiency in math and Portuguese and grade completion rates. It was created in 2005 to monitor student achievement and progression flows in primary and lower secondary education. IDEB assigns an overall score between zero and 10.[17]

In Table 2 we compare the characteristics of the schools that participated in the evaluation broken down by municipality. In particular, schools in municipalities #3 and #4 report better academic performance based on test results and higher grade completion rates and IDEB scores when compared to the other municipalities selected for the study, and also when compared to the average of schools in the country. In contrast, schools in municipalities #1 and #2 show the worst results when compared to the sample of study and to the country average. Looking at the characteristics of schools and students, we can see that schools with better academic performance have better access to infrastructure and technology. They also have fewer afro-descendant and indigenous population students and parents are more educated. In terms of teacher characteristics, a higher percentage of teachers complete 80% or more of the syllabus in municipalities with better test results (municipalities #3 and #4).

[16] Level I (0-20), Level II (20-40), Level III (40-48), Level IV (48-56), Level V (56-65), Level VI (65-76), Level VII (76-84), and Level VIII (84-100). For more information see: http://download.inep.gov.br/informacoes_estatisticas/indicadores_educacionais/2015/nota_tecnica/nota_tecnica_inep_inse_2015.pdf

[17] At the national level, the Brazilian government has the goal to reach a score of 6.0 by 2022. (http://inep.gov.br/web/guest/basic-education-assessments).

Table 2: Descriptive Statistics Evaluated Schools (Treated and Control at Baseline)

	Muni #1 (1)	Muni #2 (2)	Muni #3 (3)	Muni #4 (4)	Average (5)	Brazil (6)
Education						
Math score (standardized)	-0.9	-0.8	0.2	1.0	-0.3	0.0
Portuguese score (standardized)	-0.8	-0.7	0.4	1.0	-0.2	0.0
5th grade completion rates	80.7	79.4	96.7	99.1	86.2	90.9
IDEB initial years	4.2	3.9	5.7	6.5	4.8	5.2
School characteristics						
Students per course	31.1	23.9	30.3	30.9	27.7	27.4
INSE	47.2	46.6	53.6	52.6	49.0	49.3
School assets	9.8	9.8	9.4	8.8	9.5	6.7
Student characteristics (%)						
Female	49.7	49.2	51.5	52.4	50.3	48.8
Black, brown and indigenous	64.8	77.3	63.0	66.6	70.3	65.2
Have failed a class or more	41.9	41.5	18.9	8.7	31.3	26.7
Bathrooms >= 1	95.8	98.4	100.0	99.9	98.5	97.6
Bedrooms >= 1	98.4	98.6	99.3	98.9	98.7	98.5
Has computer	48.9	62.4	71.2	79.3	64.6	54.1
Mother & father finished primary education	60.7	63.3	78.1	85.5	69.7	63.2
Mother or father finished high school	38.1	44.8	58.8	70.1	50.8	46.0
Parents encourage you to do homework	90.8	91.7	97.2	96.9	93.5	94.6
Teacher characteristics (%)						
Female	63.6	76.7	83.3	82.0	76.5	80.2
Black, brown and indigenous	56.0	48.0	61.0	37.0	49.8	53.1
Salary > 2000 reais	46.4	64.4	79.3	82.0	67.0	47.3
Works as teacher > 10 years	93.3	64.5	52.7	76.3	69.7	66.5
Works in the school > 5 years	69.8	38.8	14.0	56.0	43.0	47.3
Works in the same classroom > 5 years	86.7	45.0	15.7	49.0	47.9	40.8
Fulfill >= 80% syllabus	29.8	27.9	72.8	92.0	47.7	48.7

3.3 Balance

To show that pair randomization was successful in constructing a control group, Table 3 presents the balance between schools in the treatment and control groups across multiple baseline characteristics (column 5). In addition, to understand the external validity of our results, we compare treatment schools with the average observed in the five municipalities where the evaluation was conducted (column 6) and also against country averages (column 7).[18] The Table shows that there are no significant differences between treated and control schools, which validates the randomization process. When we compare the sample of study with schools in the same municipalities and with national averages, we see differences across

[18] When comparing with municipality and country-level averages we only include in those samples schools that never participated in a MindLab program before.

several dimensions: (i) treated schools have lower performance according to fifth grade completion rates and IDEB scores; (ii) treated schools seem to have a higher level of assets than the rest of schools;[19] (iii) treated schools have a higher proportion of students that have failed one class or more and a lower percentage of indigenous or black students than the rest of schools in the municipalities that never received the program; and (iv) treated schools have a higher proportion of students with access to a computer and a higher proportion of teachers with salaries above the national average of 2,000 reais (approx. US$500).

Table 3: Balance

	Never Treated		Control	Treated	Difference	Difference	Difference
	All	4 muni.			(3 - 4)	(2 - 4)	(1 - 4)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Education							
Math score (standardized)	0.0	-0.3	-0.2	-0.3	0.1	0.0	0.3
Portuguese score (standardized)	0.0	-0.2	-0.1	-0.2	0.1	0.0	0.2
5th grade completion rates	91.0	90.9	86.2	86.2	0.0	4.7**	4.7**
IDEB initial years	5.2	5.0	4.9	4.8	0.1	0.2	0.4**
School characteristics							
Students per course	27.4	28.1	29.1	27.7	1.4	0.4	-0.3
INSE	49.3	49.2	49.7	49.0	0.7	0.1	0.3
School assets	6.7	8.5	9.1	9.5	-0.4	-1.0***	-2.9***
Student characteristics (%)							
Female	48.8	49.4	49.7	50.3	-0.6	-0.9	-1.5
Black, brown and indigenous	65.3	69.6	68.5	70.3	-1.8	-0.7	-5.1
Have failed a class or more	26.7	25.2	30.5	31.3	-0.8	-6.1**	-4.6
Bathrooms >= 1	97.5	98.7	98.5	98.5	0.0	0.2	-0.9
Bedrooms >= 1	98.5	98.8	98.4	98.7	-0.4	0.0	-0.2
Has computer	53.8	62.2	66.4	64.6	1.7	-2.5	-10.8**
Mother & father finished primary education	63.1	68.6	71.4	69.7	1.7	-1.1	-6.5
Mother or father finished high school	45.9	51.1	53.3	50.8	2.5	0.2	-4.9
Parents encourage you to do homework	94.6	94.7	94.5	93.5	1.0	1.2	1.1
Teacher characteristics (%)							
Female	80.1	83.6	82.1	76.5	5.7	7.1	3.6
Black, brown and indigenous	53.2	62.7	60.2	49.8	10.4	12.9**	3.3
Salary over 2000 reais	47.1	60.1	70.7	67.0	3.7	-6.9	-19.9***
Works as teacher > 10 years	66.5	71.6	74.6	69.7	4.9	1.9	-3.1
Works in the school > 5 years	45.7	37.8	42.8	43.0	-0.2	-5.2	2.7
Works in the same classroom > 5 years	40.8	45.2	48.6	47.9	0.7	-2.7	-7.1
Fulfill >= 80% syllabus	48.7	42.7	42.5	47.7	-5.2	-5.0	1.1

Note: *p<.1; **p<.05; ***p<.01

[19] The school assets variable was constructed as an index that adds the amount of assets that schools have, assigning a score from 0 to 12. A score of 0 means that the school does not have any of the following assets: Water Supply, Energy Supply, Sewage, Computer Lab, Science Lab, Sports Court, Library, Pantry, Auditorium, Projector, Computers, Internet; a score of 12 means that the school has all of these assets; and scores in between vay based on the number of assets. The MindLab program does not require the school to have specific infrastructure characteristics. Schools that participated in the evaluation were previously selected by the Secretary of Education of each municipality. Of this list, those schools that had received the MindLab program before were taken out of the sample.

3.4 Estimation of Impacts

Given the pair randomization experiment, the average treatment effect within the pair for a given outcome is given by:

$$\tilde{\tau}_p = \frac{1}{n_{p,T=1}} \sum_{i=1}^{n_{p,T=1}} Y_{i,p,T=1} - \frac{1}{n_{p,T=0}} \sum_{i=1}^{n_{p,T=0}} Y_{i,p,T=0} \tag{2}$$

Where $\tilde{\tau}_p$ is the average causal effect of the treatment for all units i within pair p. T is an indicator variable that takes the value of 1 for units that are treated and 0 for controls. The overall average effect is estimated as the average over the within-pair estimates:

$$\tilde{\tau} = \frac{1}{N/2} \sum_{p=1}^{N/2} \tilde{\tau}_p \tag{3}$$

Where N denotes the total number of schools in the sample and $N/2$ denotes the number of pairs in the sample. We estimate program impacts using the following OLS regression:

$$Y_{ispm} = \alpha + \delta T_{spm} + \phi_{pm} + \mu_{ispm} \tag{4}$$

Where Y_{ispm} is the performance of the student i, in school s, pair p, and municipality m for a given test. T_{spm} is a dummy variable that takes the value of 1 if school s in pair p and municipality m is treated with the MindLab program and 0 otherwise. Since randomization was stratified within pair and municipality, we also include pair-municipality fixed effects in the model given by ϕ_{pm}. μ_{ispm} is the error term. As treatment was given at the school level, standard errors are clustered at the school level and wild bootstrapped to take into account the relatively small number of clusters in the sample (Cameron *et al.*, 2008).

Since the program was randomly assigned to schools and therefore independent of baseline characteristics, the inclusion of observable baseline characteristics as control variables in equation (2) could improve the precision of the estimated treatment effect, without introducing bias to the estimated coefficients. As we will show later, given that we observe balance in baseline covariates we do not include any additional covariates in the estimation.

3.5 Outcomes of Interest

The evaluation focuses on four main outcomes: (i) reasoning skills, (ii) cognitive skills, (iii) metacognition, and (iv) socioemotional skills. In this section, we further explain each of the four outcomes, and Appendix A provides grater detail on the tests and variables used to measure each outcome. In addition, and to the extent that improvement in these outcomes could also influence academic learning, we also explore the impacts on standardized tests of

math and Portuguese.

To measure the impact on the main four outcomes mentioned above it was important to identify adequate indicators that were relevant for the Brazilian context and available in Portuguese. We worked in partnership with Instituto Ayrton Senna (IAS), a Brazilian company specialized in education, that developed the questionnaires and carried out the fieldwork. Experts from IAS carefully analyzed MindLab's program to identify four outcomes that could potentially be affected and that could be measured by the tests available in Portuguese. They selected which test(s) could be used to measure each outcome and, when necessary, they adapted the tests to students' needs and/or time and financial restrictions[20].

The first outcome is focused on measuring three types of reasoning: (i) Abstract (AR); (ii) Logical (LR); and (iii) Spatial (SR). Abstract Reasoning is the ability to quickly identify relationships, patterns and trends. Logical reasoning consists of the ability to analyze and evaluate written material and reason with the information obtained. Finally, Spatial Reasoning helps to visualize three-dimensional images in our minds and mentally manipulate these images and twist and turn them into the shapes we want. All the items and questions were taken from published tests[21] and were chosen for their adequacy for fifth grade students.

The second outcome is cognitive skills, more specifically we focus on executive function, which is defined as the set of cognitive skills that allow people to control and coordinate their thoughts and behavior (Shallice, 1982). The majority of instruments available today require an individualized application by a trained psychologist. Due to time and budget constraints, we did not apply these instruments to all kids in our sample, but we randomly selected a sub-sample of 10 kids per school that were subject to these tests. As explained in more detail in Appendix A, two tests were used to capture the executive function dimension: the Stroop Test and the Trail Making Test.

The third outcome we explore is metacognition. The literature distinguishes between two components (a) metacognitive knowledge and (b) metacognitive control (Pintrich, 2002). Metacognitive knowledge concerns knowledge of general strategies that can be used for different assignments, recognition of the conditions under which these strategies might be used, judgment of the extent to which the strategies are effective, and self-awareness. Metacognitive control is the ability to monitor, control, and regulate our cognition and learning. In other words, metacognitive knowledge refers only to knowledge of cognitive strategies, not the actual use of those strategies, whereas metacognitive control involves well represented tasks such as checking, planning, and executing. In order to measure the impact of the program on metacognition, we use two outcomes: the EMETA Scale and an adaptation of the

[20] Some tests were very long in their full version, so shorter versions had to be prepared to accommodate time and budget restrictions.

[21] For reasoning we use the BRT-5 tests (Primi *et al.*, 2012). See more details in Appendix A.

Motivated Strategies for Learning Questionnaire (MSLQ).

The fourth outcome evaluated are socioemotional skills, which are related to an individual's emotions and relationship with society. Socioemotional and cognitive skills are inextricably linked and together provide the foundation for developing many other skills (Busso *et al.*, 2017). Measuring socioemotional skills is complex because all available instruments depend on subjective assessments. We selected five dimensions to measure the impact of the MindLab program on socioemotional skills: "Frustration-tolerance", "Assertiveness", "Respect", "Active listening", and "Growth Mindset", which refers to the belief that you are in control of your own ability and can learn and improve.

The development of cognitive and socioemotional skills is expected to promote learning in the various areas of knowledge (Corcoran *et al.*, 2018). To start, more skilled students find it easier to learn, and thus, with the same effort, they are able to learn more. Second, learning becomes more enjoyable and less painful as more emotionally skillful learners are better able to deal with some aspects intrinsic to the learning process: anxiety, frustration, the expectations of others and of oneself, the unexpected, and uncertainty. Finally, more skilled students tend to devote more effort and engage more in school activities. As a rule, greater confidence in the ability to learn and the fact that learning becomes more enjoyable dominate the fact that learning becomes easier and more skilled students tend to strive harder to learn more. Therefore, students with better cognitive and socioemotional skills may be expected to learn more and achieve greater proficiency in all areas of knowledge. To the extent that MindLab improves cognitive and socioemotional skills, we also test whether it influences academic learning as reported in standardized tests of math and Portuguese coming from the Prova Brazil.

4 Results

4.1 Overall Program Outcomes

We start by assessing whether or not students in schools assigned to the MindLab program did, in fact, perform better in terms of the outcomes presented in the previous section: "Reasoning", "Metacognition", "Socioemotional", and "Executive Function". Then, we test the effect on their performance in math and Portuguese tests conducted as part of the Prova Brasil. Table 4 presents the results for the entire sample (all municipalities evaluated[22]) and also results for each municipality individually.

[22] The questionnaire was collected from a total sample of 2,198 students, but after cleaning the sample (i.e., keeping only the students that had at least one answer in each evaluated dimension), this leaves a sample of 1,743 students.

Columns (1) to (3) in the first section of Table 4 show that, on average, the program had no detectable effects on reasoning and metacognition and a negative and significant effect on socioemotional skills. However, the impact on socioemotional skills is small, with the mean score for treated students being 2% lower than that of the control group (15% of a standard deviation). Column (4) shows no average effect on the "Executive Function". As mentioned earlier, given the complexities and specialized expertise required to measure this dimension, a random sample of 10 kids was selected per school, which gives a smaller sample than the one used to measure the previous outcomes (527 students). Although we have lower statistical power here due to the sample size, the estimated coefficient is close to zero indicating that, if there would be any significant effect, this would be almost null.

Columns (5) and (6) present the average results for the math and Portuguese scores, respectively, showing no impacts. It is important to mention that the results of the Prova Brasil at the student level are anonymized; therefore, we cannot match our evaluation sample with the sample in Prova Brasil. In addition, there are some schools in our sample whose Prova Brasil results were not available because they did not meet the disclosure requirements.[23] If one of the schools in a pair was left out of the Prova Brasil database, we remove the whole pair from the estimation,[24] this leaves a sample of 2,532 students and 40 schools.

4.2 Results by Municipality

When we compare the results across municipalities, we find notable heterogeneities. In particular, in Column (1) we see that there is a positive and significant effect on "Reasoning" in municipality #3, where the mean score for the treatment group on this test is 19% higher than the control mean (49% of a standard deviation). In Column (2) we do not distinguish any significant effects on "Metacognition" for any municipality and in Column (3) we observe that the negative overall program effect is driven by municipalities #1 and #4, where the scores for the treated students are 4% and 3% lower than the control group, respectively (i.e., 28% and 26% of a standard deviation).

For "Executive Function", the sample is too small to allow us to estimate the results at the municipality level. For academic achievement, using scores from the Prova Brasil, the

[23] Even though the test is mandatory for all public schools, Article 18 of Ordinance 447 establishes that the results are available only for the schools that cumulatively meet the following criteria: (i) report having at least 10 students present at the time of the application of the instruments; and (ii) attain a participation rate of at least 80% of the students enrolled, according to data declared by the school to the Census of Basic Education.

[24] As expected, it is important to note that the schools that did not meet the requirements are different to those that did have Prova Brasil (see Table B1 in Appendix B). As a robustness check, in Table B6 Appendix B we estimate baseline regressions with and without the sample of schools that have Prova Brasil and results remain unchanged.

results show positive and statistically significant results, both for math and Portuguese, for municipality #3. Specifically, we observe large increases in math test scores equal to 52% of a standard deviation and 53% in Portuguese. In contrast, even though we did not find effects on other outcomes, we see negative and significant results for municipality #2, for both math and Portuguese scores (39% and 45% of a standard deviation, respectively).

It is important to highlight that while we do not observe significant changes in reasoning, metacognition and socioemotional tests in certain municipalities, it is plausible to have direct effects on academic achievement. Analysis by Durlak *et al.* (2011) shows that some programs that were ineffective in developing students' socioemotional skills had a positive impact on academic performance. The authors argue that this result may indicate that the positive effect on students' academic performance does not stem from the impact on socioemotional skills but rather from the improvement in the relationship between the teachers and students developed by the program.

In our case we observe the opposite effect: students in municipality #3 show lower scores in math and Portuguese when compared to the control group while we do not see any negative and significant effects in their socioemotional skills. Toda *et al.* (2018) provide an overview of the negative effects of gamification in education, showing how it can lead to loss of performance and other undesired behaviors due to demotivating effects caused by excessive competition or frustration for not completing all the required tasks. The authors highlight the importance of game design and adequate planning for the right deployment of games within each learning context.

Table 4: Main Program Effects

	Questionaire			Personalized Test		Prova Brazil	
	(1) Reasoning (0-1)	(2) Metacognition (1-5)	(3) Socioemotional (1-5)		(4) Executive function (0-1)	(5) Math (standardized)	(6) Portuguese (standardized)
All				**All**		**All**	
Treated	0.002	-0.027	-0.067**	Treated	-0.002	Treated	-0.015	-0.034
	(0.010)	(0.028)	(0.018)		(0.014)		(0.053)	(0.055)
Control group mean	0.470	4.155	3.415	Control group mean	0.153	Control group mean	0.169	0.198
Standard Deviation	(0.190)	(0.659)	(0.462)	Standard Deviation	(0.142)	Standard Deviation	(0.990)	(0.972)
Observations	1,743	1,743	1,743	Observations	527	Observations	2,532	2,532
Pair FE	Yes	Yes	Yes	Pair FE	Yes	Pair FE	Yes	Yes
Muni. #1						**Muni. #1**		
Treated	-0.022	-0.075	-0.137***			Treated	0.038	0.187
	(0.027)	(0.071)	(0.018)				(0.065)	(0.119)
Control group mean	0.389	4.138	3.417			Control group mean	-0.469	-0.419
Standard Deviation	(0.178)	(0.742)	(0.485)			Standard Deviation	(0.894)	(0.899)
Observations	294	294	294			Observations	272	272
Pair FE	Yes	Yes	Yes			Pair FE	Yes	Yes
Muni. #2						**Muni. #2**		
Treated	-0.016	0.001	-0.017			Treated	-0.322**	-0.398**
	(0.018)	(0.054)	(0.036)				(0.069)	(0.087)
Control group mean	0.455	4.145	3.314			Control group mean	-0.205	-0.091
Standard Deviation	(0.182)	(0.687)	(0.444)			Standard Deviation	(0.823)	(0.886)
Observations	727	727	727			Observations	771	771
Pair FE	Yes	Yes	Yes			Pair FE	Yes	Yes
Muni. #3						**Muni. #3**		
Treated	0.083***	-0.089	-0.061			Treated	0.442**	0.464***
	(0.019)	(0.058)	(0.046)				(0.108)	(0.059)
Control group mean	0.439	4.182	3.432			Control group mean	0.045	0.041
Standard Deviation	(0.171)	(0.601)	(0.467)			Standard Deviation	(0.858)	(0.879)
Observations	311	311	311			Observations	367	367
Pair FE	Yes	Yes	Yes			Pair FE	Yes	Yes
Muni. #4						**Muni. #4**		
Treated	-0.009	0.000	-0.111***			Treated	0.028	-0.007
	(0.010)	(0.027)	(0.017)				(0.068)	(0.063)
Control group mean	0.563	4.165	3.576			Control group mean	0.552	0.542
Standard Deviation	(0.187)	(0.599)	(0.429)			Standard Deviation	(0.964)	(0.922)
Observations	411	411	411			Observations	1,122	1,122
Pair FE	Yes	Yes	Yes			Pair FE	Yes	Yes

Note: SE clustered by school and wild bootstrapped to correct possible bias due to small number of clusters are presented in parentheses. *p<.1; **p<.05; ***p<.01.

4.3 Drivers of Reasoning Results in Municipality #3

As we saw in Column (1) of Table 4, there is a positive and statistically significant effect on "Reasoning" but only for municipality #3. Given that the reasoning outcome is constructed as the average of "Logical", "Abstract", and "Spatial" reasoning, we evaluate each component separately to better understand the mechanisms explaining this effect[25]. Table 5 shows the impact on the average score or proportion of correct responses. These variables range from zero to one, were values of one indicate that all questions in the test were answered correctly. We can see that the program had a positive and significant effect in all three reasoning tests. The largest impacts seem to come from improvements in spatial reasoning, with an increase in the average score of treated students of 35% (56% of a standard deviation), although the coefficient is marginally significant. The impacts on logical reasoning are quite significant

[25] To minimize the number of tables we do not report these results for municipalities #1, #2 and #4, but they are all non-significant and can be made available upon request.

and show that the average score for treated students is 14% higher than the control group mean (26% of a standard deviation).

Table 5: Reasoning in Municipality #3 (Average Score)

	(1) Logical	(2) Abstract	(3) Spatial	(4) Total
Treated	0.069***	0.061*	0.120*	0.083***
	(0.012)	(0.020)	(0.035)	(0.019)
Control Group Mean	0.470	0.490	0.356	0.439
Standard Deviation	(0.264)	(0.218)	(0.213)	(0.171)
Observations	311	311	311	311
Pair FE	Yes	Yes	Yes	Yes

Note: SE clustered by school and wild bootstrapped to correct possible bias due to small number of clusters are presented in parentheses. *p<.1; **p<.05; ***p<.01.

We also test the effect of MindLab's program on the number of mistakes that children make in each reasoning test separately, as well as on the total number of mistakes in all tests. Although the average score and the number of mistakes might seem to be two sides of the same coin, they could be capturing slightly different aspects (i.e., responding incorrectly to a question is different than leaving a question blank). There is no a priori expectation on whether students impacted by the MindLab program should be more cautious during testing by leaving questions blank when they do not know the answer or if they would be more inclined to take risks and guess. Table 6 shows that the students that participated in the program report reductions in the number of mistakes in all the reasoning tests. The overall reduction is 13% when compared to the control mean and the largest reduction is observed in spatial reasoning mistakes (16%).

Table 6: Reasoning in Municipality #3 (# Mistakes)

	(1) Logical	(2) Abstract	(3) Spatial	(4) Total
Treated	-0.567***	-0.561*	-0.943**	-0.690***
	(0.288)	(0.289)	(0.479)	(0.157)
Control Group Mean	4.848	4.689	5.974	5.170
Standard Deviation	(2.369)	(1.984)	(1.830)	(1.555)
Observations	311	311	311	311
Pair FE	Yes	Yes	Yes	Yes

Note: SE clustered by school and wild bootstrapped to correct possible bias due to small number of clusters are presented in parentheses. *p<.1; **p<.05; ***p<.01.

As the probability of getting a significant result simply due to chance increases with the number of hypotheses being tested, we take two approaches. First, as presented earlier, we aggregate several sub-tests into one single score. Second, even though the number of outcomes or sub-tests within a given test is not extremely large in our case, we correct for multiple hypothesis testing as a robustness check. As shown in Table B2 in Appendix B, our results are robust to both Family-wise Error Rate and False Discovery Rate corrections. In summary, we observe positive, statistically significant and robust results on all "Reasoning" tests for the students from municipality #3 who participated in MindLab's program.

4.4 Drivers of Changes in Socioemotional Skills

Regarding socioemotional skills, Table 4 shows negative results for all municipalities, but they are only statistically significant for municipality #1 and #4. In this section, we disaggregate the "Socioemotional" outcome in all its components to better understand the drivers behind this result. Table 7 suggests that the overall negative effect is mainly driven by two components: "Active Listening" and "Respect". However, when we analyze municipalities #1 and #4 separately, we see that the score on "Respect" is significantly lower for treated students in municipality #4, while for municipality #1 although point estimates for all the dimensions are negative, they are not precisely estimated.

Table 7: Socioemotional

	(1) Frustration-tolerance (1-5)	(2) Assertiveness (1-5)	(3) Active listening (1-5)	(4) Respect (1-5)	(5) Growth Mindset (1-5)	(6) Total (1-5)
All						
Treated	-0.033	-0.009	-0.144**	-0.094**	-0.053	-0.067**
	(0.037)	(0.030)	(0.044)	(0.031)	(0.030)	(0.018)
Control group mean	2.974	3.356	3.416	3.812	3.520	3.415
Standard Deviation	(0.821)	(0.735)	(0.903)	(0.731)	(0.667)	(0.462)
Observations	1,743	1,743	1,743	1,743	1,743	1,743
Pair Control	Yes	Yes	Yes	Yes	Yes	Yes
Muni. #1						
Treated	-0.198	-0.113	-0.147	-0.165	-0.064	-0.137***
	(-0.073)	(0.068)	(0.058)	(0.065)	(0.070)	(0.018)
Control group mean	3.075	3.357	3.356	3.877	3.422	3.417
Standard Deviation	(0.857)	(0.688)	(0.973)	(0.752)	(0.654)	(0.485)
Observations	294	294	294	294	294	294
Pair Control	Yes	Yes	Yes	Yes	Yes	Yes
Muni. #4						
Treated	-0.076	-0.030	-0.192	-0.151***	-0.105	-0.111***
	(0.062)	(0.047)	(0.065)	(0.023)	(0.050)	(0.017)
Control group mean	3.147	3.286	3.699	4.034	3.715	3.576
Standard Deviation	(0.859)	(0.752)	(0.782)	(0.644)	(0.574)	(0.429)
Observations	411	411	411	411	411	411
Pair Control	Yes	Yes	Yes	Yes	Yes	Yes

Note: SE clustered by school and wild bootstrapped to correct possible bias due to small number of clusters are presented in parentheses. *p<.1; **p<.05; ***p<.01.
To minimize the number of tables we do not report the results for municipalities #2 and #3 here, but they are all non-significant and can be made available upon request.

Table B3 in Appendix B presents the results for several Family-Wise Error Rate and False Discovery Rate corrections. In bold we mark the outcomes that were significant in Table 7 and that remain significant after the multiple hypothesis corrections. We find that the overall negative effect on "Active Listening" is robust to all adjustments as well as the negative effect on "Respect" observed in municipality #4.

Although these findings suggest negative program effects, it is important to mention that when self-perception instruments are being used what could be considered as a positive or negative effect is not always straightforward. West (2014) discusses the limitations of self-reported measures of non-cognitive skills identifying two important biases. First, the social desirability bias by which a responder may be inclined to choose a higher rating to appear more attractive. To the extent that this bias is randomly distributed across groups it should not affect impact measurement. Second, the reference bias, which occurs when responses are influenced by differing standards of comparison, which could in fact be affected by the program. For example, in the case of socio-emotional skills, the estimated results could also be understood as evidence that students exposed to the MindLab program are

more aware of these skills and set a higher bar when assessing these dimensions. In other words, the student is able to look at him/herself more rigorously, more assertively, and more questioningly. Therefore, what is called a "negative effect" may actually be the expression of a positive effect, as it points to an increase in student self-awareness, which is also one of the aims of the methodology. Further robustness checks are conducted to confirm whether we observe similar results looking at parents' responses.

4.5 Parent Perceptions of Children's Improvements

To complement measures of socioemotional development, we asked parents (or guardians) about their perceptions of their children's abilities and socioemotional behaviors. We constructed a test using parts of the Early Adolescent Temperament Questionnaire (EATQ-R) (Ellis, 2002). The EATQ-R measures 12 aspects of children's socioemotional development. Based on expert advice, we picked four outcomes (i.e., "Inhibitory control", "Attention", "Frustration", and "Aggressiveness") with six items each (24 questions in total)[26]. In addition, we added some questions for parents regarding their socioeconomic level (i.e., level of education and household's appliances).[27]

The information coming from parents' responses is not only valuable to understand their view about the improvements/changes they perceive in their children, but also serves to check the robustness of the results presented before. Some of the changes observed in children through the tests should also be visible by parents in their behavior at home. Table 8 shows no average effects detected by parents on the entire sample, but there is heterogeneity across municipalities. In particular, there are negative and significant effects on "Attention" in municipality #4, while parents in municipality #3 perceive positive changes in their kids across several dimensions. Specifically, there are significant effects on inhibitory control (5% higher than the control group mean, 22% of a standard deviation), frustration control (8% higher, 23% of a standard deviation), and aggressiveness behavior (7% higher, 24% of a standard deviation).

Overall, these results confirm that positive changes have happened in municipality #3 and that they can be attributed to the MindLab program. These positive effects are observed both in children's tests and perceived by their parents and are robust to the multiple hypothesis

[26] Complete EATQ-R also includes: "Activation Control", "Affiliation", "Perceptual Sensitivity", "Pleasure Sensitivity", "High Intensity Pleasure", "Fear", "Shyness", and "Depressive mood".

[27] Three complementary strategies were used to obtain the parents' responses. First, the parents were invited to a meeting at the school where they were asked to complete this questionnaire. If they did not attend the meeting or did not complete the questionnaire, they received a copy of the questionnaire at home the day after the meeting. For the children whose parents or guardians did not return the completed questionnaire, we tried to contact them by phone and fill out the questions through a telephone interview.

testing corrections (see Table B4 in Appendix B). In contrast, no detectable impacts are observed in the rest of the municipalities and, if any, there are some negative impacts in municipality #4, mostly coming from a reduction in students' respect and their attention or capacity to focus.

Table 8: Parents

	(1) Inhibitory control (1-5)	(2) Attention (1-5)	(3) Frustration (1-5)	(4) Aggressiveness (1-5)	(5) Total (1-5)
All					
Treated	0.017	-0.020	-0.023	-0.015	-0.010
	(0.037)	(0.046)	(0.047)	(0.041)	(0.034)
Control group mean	3.454	3.221	3.129	3.793	3.399
Standard Deviation	(0.769)	(0.786)	(0.992)	(0.723)	(0.604)
Observations	1,236	1,236	1,236	1,236	1,236
Pair Control	Yes	Yes	Yes	Yes	Yes
Muni. #1					
Treated	0.161	0.164	-0.055	0.001	0.068
	(0.087)	(0.108)	(0.116)	(0.050)	(0.044)
Control group mean	3.295	3.096	3.037	3.672	3.275
Standard Deviation	(0.858)	(0.811)	(1.023)	(0.772)	(0.633)
Observations	212	212	212	212	212
Pair Control	Yes	Yes	Yes	Yes	Yes
Muni. #2					
Treated	-0.129	-0.034	-0.065	-0.102	-0.082
	(0.069)	(0.075)	(0.064)	(0.053)	(0.051)
Control group mean	3.473	3.228	3.130	3.778	3.402
Standard Deviation	(0.759)	(0.777)	(0.984)	(0.692)	(0.576)
Observations	430	430	430	430	430
Pair Control	Yes	Yes	Yes	Yes	Yes
Muni. #3					
Treated	0.181*	0.081	0.239***	0.268*	0.192*
	(0.049)	(0.099)	(0.045)	(0.074)	(0.050)
Control group mean	3.414	3.191	3.049	3.713	3.342
Standard Deviation	(0.823)	(0.871)	(1.031)	(0.794)	(0.683)
Observations	252	252	252	252	252
Pair Control	Yes	Yes	Yes	Yes	Yes
Muni. #4					
Treated	-0.016	-0.180**	-0.144	-0.128	-0.117
	(0.056)	(0.054)	(0.086)	(0.087)	(0.067)
Control group mean	3.534	3.292	3.223	3.917	3.491
Standard Deviation	(0.688)	(0.721)	(0.958)	(0.667)	(0.552)
Observations	342	342	342	342	342
Pair Control	Yes	Yes	Yes	Yes	Yes

Note: SE clustered by school and wild bootstrapped to correct possible bias due to small number of clusters are presented in parentheses. *p<.1; **p<.05; ***p<.01.

5 Explaining Heterogeneous Effects

The main question arising from the heterogeneous results observed across municipalities is whether there are any specific municipality or school-level characteristics that may explain the differences in outcomes. Table 2 presents the characteristics of the municipalities. Municipalities #1 and #4, which perform worse, are actually quite different from each other. For example, while in municipality #4 a larger percentage of parents have primary (86%) and high school education (70%), these numbers are much lower in municipality #1 (64% and 38%, respectively). If there is any common characteristic across these two municipalities, it is that a higher percentage of teachers have more than 10 years of teaching experience (93% in municipality #1 and 76% in municipality #4) when compared to other municipalities (average of the rest is 59%).

When we look at municipality #3, which reports positive impacts, some characteristics stand out from the other municipalities. It has a lower percentage of indigenous and afro-descendant students (63% vs. 70% for the others) and a higher proportion of teachers that are indigenous or afro-descendant (61% vs. 47% for the others). It also has a lower proportion of teachers with more than 10 years of experience (53% vs. 78%) and a lower proportion of teachers who have worked more than five years in the school (14% vs. an average of 55% for the sample).

To measure heterogeneous effects we estimate the baseline model interacting treatment with the covariates that, based on this suggestive analysis, seem to be unique or distinct for municipalities performing worse and those performing better. More specifically, we test for differences across the number of students per group, the overall experience of teachers, the time teachers have been working for the same school, and the proportion of students in the class that are indigenous or afro-descendant. Table B5 shows no statistically significant differences, suggesting that teachers, students or school characteristics do not explain the differential impacts observed.

Another important dimension of program success can be related to how teachers and school staff perceive and adopt the program. To better understand this, a survey was collected from a sample of 73 teachers and directors from schools participating in the program. The survey was collected at the same time students were being tested. It included questions in three areas: (i) how the program was implemented, (ii) if they believe the program had an impact, and (iii) staff engagement in the program.

The results show that 79% of respondents indicate that the program generates a positive impact, but there are differences across municipalities. Municipality #3 is the only one where 100% of the school staff surveyed believe there are positive impacts. In contrast, only 53%

of respondents in municipality #4 believe there was an impact. Regarding responses on staff engagement in the program, we also see differences across the best and worst-performing municipalities. In municipality #3, 57% of teachers and directors were positive in terms of their engagement in the program, 14% were negative, and 29% were neutral. On the other hand, in municipality #4, only 33% were positive, 33% were negative, and 33% were neutral. In municipality #1, only 14% were positive, while 28% were negative, and 57% were neutral. This information suggests that differences in the quality of implementation may be driving heterogeneous effects across municipalities. However, our data does not allow us to empirically test this hypothesis and more research would be needed to ascertain this conclusion more credibly in the context of this program.

Table 9: Does the Program Have an Impact?

	Muni #1 (1)	Muni #2 (2)	Muni #3 (3)	Muni #4 (4)	Total (5)
Yes	11	25	14	8	58
No	0	0	0	1	1
Don't know	3	5	0	6	14
Total	14	30	14	15	73

6 Robustness and Placebo Tests

In addition to correcting for multiple hypothesis testing, we perform some robustness and placebo tests. First, we re-estimate effects on reasoning, cognitive, and socioemotional outcomes excluding those schools that do not have the Prova Brasil. This way we are using the same sample of schools as the one used for math and Portuguese scores. As presented in Table B6 in Appendix B, all the results remain unchanged.

In a second robustness test we define the outcomes and estimate a linear probability model where the dependent variable takes the value of one if the outcome score is in the top 25% of the score distribution within each municipality and zero otherwise. Results are reported in Table B7 in Appendix B and show that the positive effect observed in municipality #3 for reasoning remains, and the negative effect observed in the full sample and in municipality #1 on socioemotional skills is also stable. In this model, we no longer see the statistically negative and significant effect on socioemotional tests for municipality #4 but a negative effect in metacognition emerges.

Finally, we implement a placebo test exploiting the information from Prova Brasil in 2015, two years prior to our baseline. If treatment is orthogonal to school characteristics, we

should expect to see no effect in years prior to the implementation of the MindLab program. As shown in Table B8 in Appendix B, there are no significant differences between treatment and control groups in math and Portuguese results for municipality #3 in the baseline, which further supports the conclusion that the effects that have been identified can be attributed to the MindLab program. In contrast, we do see differences in baseline scores for treatment and control groups for municipality #2, but these differences are actually in the opposite direction. Treated schools were performing worse than control schools in the baseline. In a robustness check we re-estimated Table 4 columns (5) and (6) for municipality #2 controlling for the average score in math in 2015 and the results remained unchanged.[28]

7 Conclusions

There has been an increasing number of school programs dedicated to promoting the development of socioemotional skills. This follows widely accepted evidence on the importance of these skills for academic performance (Corcoran *et al.*, 2018) and for accessing the labor market and securing a job (Bassi *et al.*, 2016). Despite the increased popularity of socioemotional learning programs, the literature has been divided in terms of their effectiveness in achieving the expected objectives. Multiple studies have shown large positive impacts (Durlak *et al.*, 2011), but several rigorous experimental studies with large samples have shown smaller impacts or no impacts (Corcoran *et al.*, 2018). Educational systems have also seen an increase in the use of games to promote learning. The very limited causal research on gamification and learning has shown positive effects but also several unintended consequences (Araya *et al.*, 2019; Dominguez *et al.*, 2013), thus raising the need for further evidence to better understand when these approaches work.

We conducted a large-scale randomized control trial to evaluate the impacts of a game-based learning program dedicated to building socioemotional skills. More specifically, we evaluated the impacts of the MindLab program in Brazil, a globally-recognized methodology, which is based on reasoning games, metacognitive methods, and uses the teacher as a mediator who provokes reflection. We designed a matched-pair cluster randomized experiment with a sample of 56 schools (28 treated and 28 controls) in four municipal networks and a sample of 2,532 fifth-grade students. We worked with local experts in education to identify and develop a series of tests aimed at measuring cognitive, social, emotional and ethical skills. In addition, we explored the impacts on academic achievement by looking at Brazilian standardized tests.

Our results show no significant program impacts at the average level and we observe small

[28] To minimize the number of tables we do not report the results here, but they can be made available upon request.

negative effects, if any, on children's socioemotional outcomes. The treatment group's test scores on this dimension are 15% of a standard deviation lower than those of the control group. Despite the almost null average effect, there is important heterogeneity across municipalities, with one municipality showing robust positive results for reasoning and academic performance, while other municipalities present negative and significant effects in socioemotional outcomes and academic performance. Positive results on socioemotional skills are detected both with student-level measurements and parent-level surveys, while negative results are not consistently observed across different types of measures.

An analysis of the mechanisms driving the effects shows that no specific school, teacher or student characteristics available in our dataset play a role in explaining the differential effects. However, administrative program data and surveys conducted among school staff reveal that, in successful cases, teachers were clearly involved and had a positive perception of the program and its methods. In cases where the program was unsuccessful, opposite views were evident. Therefore, differences in the quality of implementation may be driving heterogenous effects across different municipalities. However, our data does not allow us to empirically test this hypothesis and more research will be needed to ascertain this conclusion more credibly.

One of the key conclusions emerging from our analysis is that game-based learning approaches can work well in some settings but can also have limited impacts in others, or even have unexpected consequences. This is consistent with what previous studies have also suggested. In addition, it is important to keep in mind the limitations of self-reported measures of non-cognitive skills. There is a possibility that the negative estimated results found for some outcomes could also be understood as evidence that students exposed to the program are more aware of these skills and set a higher bar when assessing these dimensions. Therefore, what is called a "negative effect" may actually be the expression of a positive effect, as it points to an increase in student self-awareness, which is also one of the aims of the methodology.

Overall, these programs need to be carefully designed and implemented. Our evidence highlights how important it is for teachers and school staff to fully understand and "buy into" the program. Therefore, providing teachers with the right support during training and ensuring clear communication around the value of these new methods is key. Game design elements and planning are also critical. In some municipalities, our results point towards negative effects on academic performance, measured by standardized tests, even though no effects were observed on socioemotional and cognitive skills. Previous research has shown how gamification can lead to loss of performance and other undesired behaviors due to demotivating effects caused by excessive competition or frustration for not completing

all required tasks (Toda *et al.*, 2018).

Although our results offer valuable evidence on the effects of socioemotional learning programs that use gamification, they need to be taken with caution considering their external validity. As shown, the schools that participated in the evaluation sample have some different characteristics than the universe of schools in the municipalities that were selected for the study and from other schools in the country. Thus, these results cannot be generalized to the entire municipality or country, but only to those schools that share similar characteristics and institutional contexts as those that participated in the evaluation.

Finally, it is important to mention that with the results obtained in this study, MindLab has been making adjustments and improvements to its program and implementation procedures. First, they have created training courses for the teams of managers in municipalities, recognizing the importance of the support received by professionals from the Department of Education (technical teams from the Department of Education, principals and coordinators of the school units) to better organize the program and make it more viable in the network. Regarding teachers, several measures are being taken. First, MindLab is adjusting its calendar to the teachers' routine to avoid delays at the beginning of the school year as much as possible. Second, whenever possible, MindLab holds an opening event in each municipality with a lecture, workshop or other activity to prepare teachers for the training and showcase the value of the program for both students' education and teachers' professional development. Additionally, MindLab has created a more flexible curriculum that adapts the number of classes to the characteristics of each municipality. Finally, the materials delivered to teachers have been improved to make them more attractive and self-explanatory. Student books and materials prepared to integrate families into the process have also been expanded. Once all these improvements are fully in place it would be important to rigorously re-evaluate the program to confirm the effectiveness of these measures in order to keep guiding its expansion.

References

Ajello, A. M., Di Marco, C., & Marchi, S. (2012). Acquisire life skills mediante il progetto mind lab: una verifica sperimentale nelle scuole primarie delle provincie di trento e vicenza.

Araya, R., Arias Ortiz, E., Bottan, N., & Cristia, J. (2019). Does gamification in education work? *IDB Working Paper Series*, IDB-WP-982.

Azevedo, V., Paes de Barros, R., Franco, S., Garcia, B. S., & Muller Machado, L. (2019). *Avaliação não experimental de impacto do programa menteinovadora*. Technical report.

Bassi, M., Busso, M., Urzua, S., & Vargas, J. (2016). *Disconnected: Skills, Education, and Employment in Latin America*. Inter-American Development Bank. Education.

Blakemore, S.-J. & Choudhury, S. (2006). Development of the adolescent brain: implications for executive function and social cognition. *Journal of child psychology and psychiatry*, 47(3-4), 296–312.

Bolfer, C. P. M. (2009). Avaliação neuropsicológica das funções executivas e da atenção em crianças com transtorno do déficit de atenção/hiperatividade (tdah).

Bolfer, C. P. M. (2014). Avaliação neuropsicológica da funções executivas e da atenção antes e depois do uso do metilfenidato em crianças com transtorno de déficit de atenção/hiperatividade.

Bourguignon, F., Ferreira, F. H., & Menéndez, M. (2007). Inequality of opportunity in brazil. *Review of Income and Wealth*, 53(4), 585–618.

Bruhn, M. & McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American economic journal: Applied Economics*, 1(4), 200–232.

Busso, M., Cristia, J., Hincapie, D., Messina, J., & Ripani, L. (2017). Learning better: Public policy for skills development.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3), 414–427.

Corcoran, R. P., Cheung, A. C., Kim, E., & Xie, C. (2018). Effective universal school-based social and emotional learning programs for improving academic achievement: A systematic review and meta-analysis of 50 years of research. *Educational Research Review*, 25, 56–72.

Dominguez, A.and Saenz-de-Navarrete, J., de Marcos, L., Fernandez-Sanz, L., Pages, C., & Martinez-Herraiz, J. (2013). Gamifying learning experiences: Practical implications and outcomes. *Computers and Education*, 63, 380–392.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, 82(1), 405–432.

Ellis, L. K. (2002). *Individual differences and adolescent psychosocial development*. PhD thesis, University of Oregon Eugene.

Garcia, S. & Abed, A. (2009). Impacto do desenvolvimento de habilidades por meio da aplicação da metodologia do projeto menteinovadora: um estudo em alunos de 5º ano do ensino fundamental. *Mind Lab Brasil & INADE*.

Garcia, S. & Abed, A. (2010). A metodologia do projeto menteinovadora no desenvolvimento de habilidades em alunos de 5º ano do ensino fundamental: o professor faz a diferença nas demandas do século xxi. *Mind Lab Brasil & INADE*.

Garcia, S. & Abed, A. (2013). Contribuições da metodologia mind lab na transformação dos protagonistas da escola do século xxi. *Mind Lab Brasil & INADE*.

Garcia, S., Abed, A., Soares, T., & Ramos, M. (2012). O prazer de ensinar e de aprender: contribuições de uma metodologia no aprimoramento das práticas pedagógicas. *São Paulo: Mind Lab Brasil & INADE*.

Garcia, S. R. R., Abed, A. L. Z., Soares, T. M., & Donnini, S. (2011). Saltos de aprendizagem: o percurso de uma metodologia inovadora em educação. *São Paulo: Mind Lab Brasil & INADE*.

Green, D. & Gendelman, D. (2003). Teaching children to think strategically: Results from a randomized experiment. *Unpublished manuscript, Institution for Socialand Policy Studies at Yale University*.

Green, D. & Gendelman, D. (2004). Can a curriculum that teaches abstract reasoning skills improve standardized test scores? *Unpublished manuscript, Institution for Socialand Policy Studies at Yale University*.

Hromek, R. & Roffey, S. (2009). Promoting social and emotional learning with games. *Simulation & Gaming*, 40(5), 626–644.

Imai, K., King, G., & Nall, C. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the mexican universal health insurance evaluation. *Statistical Science*, 24(1), 29–53.

Konishi, C. & Wong, T. (2018). Relationships and school success: From a social-emotional learning perspective.

McCormick, M., Neuhaus, R., O'Connor, E., White, H., Horn, E., Harding, S., Cappella, E., & McClowry, S. (2020). Long-term effects of social-emotional learning on academic skills: Evidence from a randomized trial of insights. *Journal of Research on Educational Effectiveness*, DOI: 10.1080/19345747.2020.1831117.

MindLab (2018). A mindlab. https://www.mindlab.com.br/a-mind-lab/. Accessed: 02-06-2021.

Mishima, N., Kubota, S., & Nagata, S. (2000). The development of a questionnaire to assess the attitude of active listening. *Journal of Occupational Health*, 42(3), 111–118.

Mondi, C. & Reynolds, A. (2020). Socio-emotional learning among low-income prekindergarteners: The roles of individual factors and early intervention. *Early Education and Development*, DOI: 10.1080/10409289.2020.1778989.

Novella, R., Repetto, A., Robino, C., & Rucci, G. (2018). *Millennials in Latin America and the Caribbean: to work or study?* Inter-American Development Bank. Espacio Publico. IDRC/CRDI. Canada.

Otfried, S. & Strauss, E. (1998). A compendium of neuropsychological tests.

Pintrich, P. R. (1991). A manual for the use of the motivated strategies for learning questionnaire (mslq).

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory into practice*, 41(4), 219–225.

Primi, R., Couto, G., Almeida, L. S., Guisande, M. A., & Miguel, F. K. (2012). Intelligence, age and schooling: Data from the battery of reasoning tests (brt-5). *Psicologia: Reflexão e Crítica*, 25(1), 79–88.

Sandoval, L. (2012). The effect of education on brazil's economic development. *Global Majority E-Journal*, 3(1), 4–19.

Santos, D. & Primi, R. (2014). Desenvolvimento socioemocional e aprendizado escolar: Uma proposta de mensuração para apoiar políticas públicas.

Shallice, T. (1982). Specific impairments of planning. *Phil. Trans. R. Soc. Lond. B*, 298(1089), 199–209.

Sklad, M., Diekstra, R., Ritter, M. D., Ben, J., & Gravesteijn, C. (2012). Effectiveness of school-based universal social, emotional, and behavioral programs: Do they enhance students' development in the area of skill, behavior, and adjustment? *Psychology in the Schools*, 49(9), 892–909.

Toda, A., Isotani, S., & Dias Valle, P. (2018). The dark side of gamification: An oiverview of negative effects of gamification in education. *Mimeo*.

West, M. (2014). The limitations of self-report measures of non-cognitive skills. *Brookings Institutions Report*.

Zainuddin, Z., KaiWah Chu, S., Shujahat, M., & C.J., P. (2020). The impact of gamifiication on learning and instruction: A systematic review of empirical evidence. *Educational Research Review*, 30.

A Appendix A: Description of Outcomes of Interest

The main objective of the MindLab program is the development of cognitive, social, emotional and ethical skills.The following sections provide more details on the different outcomes explored in the paper.

A.1 Reasoning

The "BRT-5" identifies five types of reasoning: (i) Abstract (AR); (ii) Logical (LR); (iii) Numerical (NR); (iv) Spatial (SR); and (v) Mechanical (MR) (Primi *et al.*, 2012). MindLab's curriculum fosters different types of cognitive abilities in each grade. For fifth grade they focus on Logical and Abstract Reasoning, but we decided to also measure Spatial Reasoning given the possibilities to find spillover effects on that outcome. Abstract Reasoning is the ability to quickly identify relationships, patterns and trends and we measure it by nine abstract analogies of geometric figures. Logical Reasoning consists of the ability to analyze and evaluate written material and reason with the information obtained, by for example analyzing relationships among component parts of sentences or recognizing relationships among words and concepts. Therefore, to assess the effect on Logical Reasoning we use nine verbal analogies. Finally, Spatial Reasoning helps you visualize three-dimensional images in your mind and to mentally manipulate these images and twist and turn them into the shape you want. We measure it by nine spatial series related to the rotation of the six faces of a cube. For each reasoning dimension we use nine items to evaluate or generate the score. All these items or questions were taken from published tests and were chosen considering their adequacy for fifth grade students.

A.2 Executive Function

The term executive function is used to describe the set of cognitive skills that allows us to control and coordinate our thoughts and behavior (Shallice, 1982). These skills include (i) planning, (ii) decision-making, (iii) problem solving, and (iv) resource management. Each of these executive functions has a role in cognitive control, such as, leaving aside unimportant information, keeping in mind a plan to carry out in the future, and controlling impulses (Blakemore & Choudhury, 2006). MindLab aims to foster these abilities by emphasizing in each game: (i) the importance of considering different strategies; (ii) encouraging students to compare their options in terms of pros and cons and picking the best choice in each case; (iii) emphasizing the importance of understanding the nature of the problem or the situation; and (iv) keeping in mind that time and tries are scarce and finite.

The measurement of executive function abilities is not an easy task. The majority of instruments available today require an individualized application by a trained psychologist. Due to time and budget constraints, we did not apply these instruments to all kids in our sample, but we randomly selected a sub-sample of 10 kids per school who were subjected to these tests. This decision may affect the statistical power of our estimated impacts, as discussed later.

Two tests were used to capture the executive function dimension: The Stroop Test, developed by John Ridley Stroop (1935) and the Trail Making Test (TMT) developed by Partington and Leiter in 1938.

The Stroop Test is used to evaluate inhibitory control (impulsiveness), selective attention, and flexibility. The test assumes that people have difficulties in processing simultaneous information with conflicting meaning, even when one of them has no relevance to the task. We used the Victoria Stroop Test (Otfried & Strauss, 1998), a brief version of the Stroop test that consists of three cards, all using 24 printed stimuli arranged in four columns in standard colors: green, red, blue and orange. In the first card, small rectangles are presented, and the student must name as fast as he/she can the color in which they are printed. On the second card, there are common words (i.e., "each", "nothing", "never", and "everything") colored in the same standard colors and the student must name the colors in which these words are printed. On the last card, there are color names printed in a different color (e.g. "red" is printed in blue) and the task is to name the color in which the word is printed and not the names of the printed colors.

The Trail Making Test includes two parts. Part A consists of two sheets, one with the first 12 letters of the alphabet randomly arranged and the other with a random sequence of numbers from 1 to 12. The task is to connect the elements in order (the letters in the first sheet and the numbers in the second sheet) without removing the pencil from the paper. Part B consists of a page with the first 12 letters and the first 12 numbers randomly arranged and the participant's task is to link items alternately to alphabetic and numeric sequences without removing the pencil from the paper (i.e., A-1-B-2-C-3, ..., M-12). For both parts (A and B), there is a time limit for the execution of the task. In this evaluation, we apply only Part B of the Track Test in a short version with eight numbers and seven alternate letters developed by Bolfer (2009, 2014).

A.3 Metacognition

As Pintrich (2002) points out, metacognition concerns knowledge about cognition in general, as well as awareness of and knowledge about one's own cognition. The literature distinguishes between two components: (a) metacognitive knowledge and (b) metacognitive

control. Metacognitive knowledge concerns knowledge of general strategies that can be used for different assignments, recognition of the conditions under which these strategies might be used, judgment of the extent to which the strategies are effective, and self-awareness. Metacognitive control is the ability to monitor, control, and regulate one's cognition and learning. In other words, metacognitive knowledge refers only to knowledge of cognitive strategies, not the actual use of those strategies, whereas metacognitive control involves well represented tasks such as checking, planning, and executing.

MindLab's program intends to promote metacognition in three different ways. First, by boosting kids' awareness of their thinking process by stimulating them to play in a conscious way. Second, by broadening their repertoire of reasoning and learning methods and strategies (provided that they are able to transcend the use of the game environment where they have experienced and learned these strategies). Finally, by improving their metacognitive control to: (i) choose what seems to be the best strategy in each situation, (ii) assess to what extent the chosen strategy is leading to expected results, and (iii) re-adjust the strategy in order to increase its effectiveness.

In order to measure the impact of the program on metacognition, we use two outcomes: the EMETA Scale, developed by Pascualon and Schelini, and an adaptation of MindLab's Motivated Strategies for Learning Questionnaire (MSLQ).

EMETA is a self-report questionnaire originally based on 70 items. We use a shorter version composed of six items that evaluate metacognitive knowledge and ten items that evaluate cognitive self-regulation. This short version of EMETA is a four-point Likert scale ranged from "never" to "always".

The Adapted MSLQ, like MindLab's version[29], is a self-report questionnaire that uses a seven-point Likert scale. However, this adaptation uses only four themes: "Planning", "Study Strategy", "Self-regulation", and "Execution", each one with 10 items.

A.4 Socioemotional

Socioemotional skills help people to identify and manage their own and others' emotions to improve productivity, including the ability to work in groups. Socioemotional and cognitive skills are inextricably linked and together provide the foundation for developing many other skills (Busso *et al.*, 2017). However, measuring socioemotional skills is complex because all available instruments depend on subjective assessments.

MindLab asserts that its methodology helps students develop various regulatory functions,

[29] MindLab's MSLQ is an adaptation of the in Pintrich (1991) test. The original version included 81 items, organized in a fifteen-point scale; MindLab's test uses 70 items, organized in a seven-point scale with 10 items each.

including the capacity to: (i) deal with uncertainty, loss, and success, (ii) regulate their anxiety, frustration, enthusiasm, and euphoria, and (iii) regulate their inhibitory control and impulsivity. The program also encourages a friendly and respectful relationship among students; develops students' ability to work together as a team, cooperate, collaborate and act positively for the common good; and expands their ability to work in healthy competitive environments.

We selected five dimensions to measure the impact of the MindLab program on socioemotional skills: "Frustration-tolerance", "Assertiveness", "Respect", "Active listening", and "Growth Mindset". The first three dimensions were captured using questions from an instrument developed by Instituto Ayrton Senna called SENNA (Social and Emotional or Noncognitive Nationwide Assessment) (Santos & Primi, 2014). To evaluate Active Listening, we construct a six-item test based on the traditional Adult Active Listening Attitude Scale (ALAS) developed by Mishima *et al.* (2000), that uses a five-point Likert scale in order to facilitate the application in conjunction with items taken from SENNA 2.0.[30] Finally, Growth Mindset was measured using an adapted version of the scale adopted by a set of six California educational districts operating through a nonprofit organization called the California Office to Reform Education (CORE). It is a six-item self-report test that uses a five-point Likert scale ranged from "nothing" to "totally".

A.5 Math and Portuguese

The development of cognitive and socioemotional skills is expected to promote learning in various areas of knowledge (Corcoran *et al.*, 2018). In the first place, more skilled students find it easier to learn, and thus, with the same effort, they are able to learn more. Second, learning becomes more enjoyable and less painful as more emotionally skillful learners are better able to deal with some aspects intrinsic to the learning process: anxiety, frustration, the expectations of others and of oneself, the unexpected and uncertainty. Finally, more skilled students tend to devote more effort and engage more in school activities. As a rule, greater confidence in the ability to learn and the fact that learning becomes more enjoyable dominate the fact that learning becomes easier and more skilled students tend to strive harder to learn more. Therefore, students with better cognitive and socioemotional skills may be expected to learn more and achieve greater proficiency in all areas of knowledge. To the extent that MindLab improves cognitive and socioemotional skills, we also test whether it influences academic learning as reported in standardized tests of math and Portuguese coming from the Prova Brasil.

[30] ALAS is a self-report test, originally formed by 40 items on a four-point Likert scale.

The following tables summarize the outcomes of interest and the different sources of information.

Table A1: Summary Outcomes of Interest

Outcome	Dimensions	Source
Reasoning	Logical	BRT-5
	Abstract	BRT-5
	Spatial	BRT-5
Metacognition	Metacognitive knowledge	EMETA
	Cognitive self-regulation	EMETA
	Planning	Adapted MSLQ
	Study Strategy	Adapted MSLQ
	Self-regulation	Adapted MSLQ
	Execution	Adapted MSLQ
Socioemotional	Frustration-tolerance	SENNA 2.0
	Assertiveness	SENNA 2.0
	Respect	SENNA 2.0
	Active listening	ALAS
	Growth Mindset	CORE
Executive Function	Inhibitory control	The Stroop Test
	Flexible Thinking	TMT
Learning	Math	Prova Brazil
	Portuguese	Prova Brazil

For a detailed description of the indicators see Table A2 in Appendix A.

Table A2: Outcomes of Interest Description

Outcome	Dimension	Scale	Questions or description	Test	Source
Reasoning	Logical	0-1	Average of correct responses in nine verbal analogies.	BRT-5	Primi *et al.* (2012)
	Abstract	0-1	Average of correct responses in nine abstract analogies of geometric figures.		
	Spatial	0-1	Average of correct responses in nine rotations of the six faces of a cube.		
Metecognition	Meta. knowledge	1=Never, 4=Always	When I finish reading a book, I know what I understood and what I did not.	EMETA	Pascualon and Schelini
		1=Never, 4=Always	When I study the same thing several times, I remember it more easily.		
		1=Never, 4=Always	I can tell you how much I understood about something I studied.		
		1=Never, 4=Always	I know when I understood the history of a book.		
	Cognitive self-reg.	1=Never, 4=Always	I know there are easier and harder ways to solve problems.		
		1=Never, 4=Always	I can learn in different ways, depending on the situation.		
		1=Never, 4=Always	When I'm doing a task, I usually stop it a few times to see if I understood it correctly.		
		1=Never, 4=Always	I think why it is important to learn something before studying about it.		
		1=Never, 4=Always	When I'm studying something new, I think about how I'm doing.		
		1=Never, 4=Always	After finishing a task, I wonder if I learned important things.		
		1=Never, 4=Always	When I need to remember several pieces of information together, I order them.		
		1=Never, 4=Always	While I try to solve a problem, I ask myself questions.		
		1=Never, 4=Always	When I solve a problem, I wonder if I'm thinking of all the possibilities to solve it.		
		1=Never, 4=Always	While performing a task that someone has asked me to do, I stop a few times to see if I am doing it right.		
		1=Never, 4=Always	To better understand one thing I use what I have learned by solving similar things before.		
		1=Never, 4=Always	While I am working on a task, I wonder if I can answer it or if I have to work on it.		
	Planning	1=Totally false, 7=Totally true	As I study, I try to identify the problem and work out a plan to solve it.	Adapted MSLQ	MindLab's Motivated Strategies for Learning Questionnaire (MSLQ)
		1=Totally false, 7=Totally true	When I have more than one tasks to do, I think of what to do first to better organize myself.		
		1=Totally false, 7=Totally true	I find it important to analyze a problem to try to locate the most important information and assess the situation.		
		1=Totally false, 7=Totally true	I try to organize myself to fulfill all my tasks and have time to play.		
		1=Totally false, 7=Totally true	I can organize myself well to study, even when I have a lot of tasks to do.		
		1=Totally false, 7=Totally true	If I have two tests on the same day, I will have difficulty preparing myself well for both.*		
		1=Totally false, 7=Totally true	I arrange to study only one day or two before the exam.*		
		1=Totally false, 7=Totally true	I find it difficult to draw up a study plan.*		
	Study Strategy	1=Totally false, 7=Totally true	In general, when I have to solve a problem, I think of the best way to solve it.		
		1=Totally false, 7=Totally true	When I have plenty of time to do a school assignment, I'd rather leave it to do it later.*		
		1=Totally false, 7=Totally true	I think I learn better when I explain the story to colleagues.		
		1=Totally false, 7=Totally true	I find it hard to remember what I learned in class using the notes in my notebook.*		
		1=Totally false, 7=Totally true	I find it hard to write down what I am learning in class.*		
		1=Totally false, 7=Totally true	I often feel that I am studying the wrong way.*		
		1=Totally false, 7=Totally true	When I read my notes in class, I remember exactly what was taught and what I learned.		
		1=Totally false, 7=Totally true	I study all the subjects in the same way, but sometimes I don't think that works.*		
		1=Totally false, 7=Totally true	When I study, I review the readings and notes that I made in class to appropriate the most important ideas.		
		1=Totally false, 7=Totally true	I try to memorize keywords to remember the most important concepts.		
		1=Totally false, 7=Totally true	When I study, I make a brief summary of the main ideas and concepts given in class.		
		1=Totally false, 7=Totally true	When I study, I try to relate what is in the books with what was explained in class.		
		1=Totally false, 7=Totally true	In class, I often miss important points because I'm thinking in other things.*		
	Self-regulation	1=Totally false, 7=Totally true	When the task is not interesting, I finish it as soon as possible to get rid of it soon.*		
		1=Totally false, 7=Totally true	When I'm reading, sometimes I go back to some previous section to better understand the content.		
		1=Totally false, 7=Totally true	When I do not understand a story, I seek help to resolve my doubts.		
		1=Totally false, 7=Totally true	In most classes, I think about other things and miss the teacher's explanation.*		
		1=Totally false, 7=Totally true	I often ask myself questions to make sure I already know the content of the test.		
		1=Totally false, 7=Totally true	In class, I think differently about what the teacher is explaining.		
		1=Totally false, 7=Totally true	When I'm studying, I get very focused and hardly get distracted by anything else.		
		1=Totally false, 7=Totally true	If I pause in my studies to do something else, then I can remember exactly where I left off.		
		1=Totally false, 7=Totally true	In the school, I am easily distracted by something else and I do not conclude what I was doing.*		
		1=Totally false, 7=Totally true	During my studies, I try to understand the importance of the contents and the purpose of the task.		
		1=Totally false, 7=Totally true	During class, when the subject is difficult, I give up paying attention and taking notes.*		
		1=Totally false, 7=Totally true	During class, I try to understand what the teacher is explaining, even if it does not make much sense to me.		
	Execution	1=Totally false, 7=Totally true	I don't review my notes before the test, because I rarely find time.*		
		1=Totally false, 7=Totally true	During class, even when it's difficult, I try to learn.		
		1=Totally false, 7=Totally true	When I do the tasks, I start with the most important ones.		
		1=Totally false, 7=Totally true	During my studies, I relate what I am learning with other subjects and situations.		
		1=Totally false, 7=Totally true	I often find it hard to keep my study plan.*		
		1=Totally false, 7=Totally true	I can usually do all my tasks.		
		1=Totally false, 7=Totally true	I often can not complete class assignments and homework.*		

Outcome	Dimension		Scale	Questions or description	Test	Source
Socioemotional	Frustration-tolerance	Identity Self-efficacy	1=Nothing, 5=Totally	Control your anger when people make you angry.	SENNA 2.0	Santos & Primi (2014)
			1=Nothing, 5=Totally	Control your anger when something happens that you do not want to happen.		
			1=Nothing, 5=Totally	Avoid getting nervous.		
			1=Nothing, 5=Totally	I get offended easily.*		
			1=Nothing, 5=Totally	I get very angry and I usually lose my temper.*		
			1=Nothing, 5=Totally	I am calm and control my stress well.		
			1=Nothing, 5=Totally	They do not take me seriously		
			1=Nothing, 5=Totally	I often explode with anger.*		
			1=Nothing, 5=Totally	I'm don't get upset easily		
	Assertiveness	Identity Self-efficacy	1=Nothing, 5=Totally	Ask the teacher to repeat it if I did not understand an explanation.		
			1=Nothing, 5=Totally	Ask the teacher questions during class		
			1=Nothing, 5=Totally	Ask for help from teachers when have difficulties.		
			1=Nothing, 5=Totally	I do not say anything when my colleagues say something I do not agree with.*		
			1=Nothing, 5=Totally	I keep quiet in the classroom even when I have something important to ask the teacher.*		
			1=Nothing, 5=Totally	I'm ashamed to ask questions during class.*		
			1=Nothing, 5=Totally	I'm not afraid to say the things I think		
			1=Nothing, 5=Totally	I certainly give my opinions in group discussions.		
			1=Nothing, 5=Totally	I take the lead in group work		
	Respect	Identity Self-efficacy	1=Nothing, 5=Totally	Listen respectfully to others' opinions		
			1=Nothing, 5=Totally	Avoid discussions with other people.		
			1=Nothing, 5=Totally	Treat people you do not like well and respectfully.		
			1=Nothing, 5=Totally	I ask for things with education and gratefulness		
			1=Nothing, 5=Totally	I apologize to the people I hurt.		
			1=Nothing, 5=Totally	I like to provoke others.*		
			1=Nothing, 5=Totally	Cursed people.*		
			1=Nothing, 5=Totally	Respect authorities (teachers, directors, etc.).		
			1=Nothing, 5=Totally	I make threats to get what I want.*		
	Active listening		1=Nothing, 5=Totally	When someone is talking to me I get distracted fast.*	ALAS	Mishima *et al.* (2000)
			1=Nothing, 5=Totally	I listen to what others are talking about without distracting me.		
			1=Nothing, 5=Totally	I start talking before my friend has finished.*		
			1=Nothing, 5=Totally	I always hustle for people to speak faster.*		
			1=Nothing, 5=Totally	When I realize it, instead of listening to the other person, I'm arguing with her.*		
			1=Nothing, 5=Totally	I start talking before my friend has finished.*		
			1=Nothing, 5=Totally	When I realize it, instead of listening to the other person, I'm arguing with her.*		
	Growth Mindset		1=Nothing, 5=Totally	No matter who you are, it is always possible to change your intelligence	CORE	CORE
			1=Nothing, 5=Totally	My intelligence is something I can not change much.*		
			1=Nothing, 5=Totally	No matter how much intelligence you have, you can always improve it.		
			1=Nothing, 5=Totally	There are things I can not learn.*		
			1=Nothing, 5=Totally	Challenging me will not make me smarter.*		
			1=Nothing, 5=Totally	If I'm not naturally clever in a subject, I'll never be able to do well in it.*		
Executive function	Inhibitory control		0-1	Average of correct responses in 3 tests of 24 tasks each.	The Stroop Test	Otfried & Strauss (1998)
	Flexible Thinking		0-1	Average of correct responses in 14 exercises.	Trail Making Test (TMT)	Bolfer (2009, 2014)
Learning	Math		Standarized SAEB	Student Proficiency in Math	Prova Brazil	INEP
	Portuguese		Standarized SAEB	Student Proficiency in Portuguese		

*This variable was multiplied by (-1) to give it a possitive connotation.

Table A3: Description of Parents Outcomes of Interest

Outcome		Scale	Questions	Test
Inhibitory	control	1=Never, 5=Always	Has trouble waiting her/his turn to speak when she/he is agitated.*	Early Adolescent Temperament (EATQ-R)
		1=Never, 5=Always	Opens gifts before she/he should.*	
		1=Never, 5=Always	Is likely to try to do something she/he shouldn't even if she/he tries to avoid.*	
Attention		1=Never, 5=Always	Can avoid laughing at inappropriate times.	
		1=Never, 5=Always	Generally able to focus on plans and goals.	
		1=Never, 5=Always	Finds really easy to focus on a problem.	
		1=Never, 5=Always	If he/he is interrupted or distracted, forgets what was saying.*	
		1=Never, 5=Always	Has difficulty concentrating with noises when trying to study.*	
		1=Never, 5=Always	Is good to deal with several different stimuli that are happening around.	
		1=Never, 5=Always	Often stops in the middle of a task and goes out to do something else without ending it.*	
		1=Never, 5=Always	Pay attention when someone tells she/he how to do something.	
Frustration		1=Never, 5=Always	Bothered by the little things that other colleagues do.*	
		1=Never, 5=Always	Gets very annoyed when someone criticizes sh/he.*	
		1=Never, 5=Always	Gets angry when you don't take her/him somewhere she/he wants to go.*	
		1=Never, 5=Always	Gets angry when has to stop doing something she/he enjoys.*	
		1=Never, 5=Always	Hates when people do not agree with her/him.*	
		1=Never, 5=Always	She/he gets very frustrated when she/he makes a mistake in the schoolwork.*	
		1=Never, 5=Always	When angry at someone, say things she/he knows will hurt the feelings of the person.*	
Aggressiveness		1=Never, 5=Always	If she/he is very angry, she/he may hit someone.*	
		1=Never, 5=Always	Tends to be rude to people she/he do not like.*	
		1=Never, 5=Always	Usually tries to blame mistakes on someone else.*	
		1=Never, 5=Always	Hit doors when angry.*	
		1=Never, 5=Always	Enjoys the appearance of other people.*	
		1=Never, 5=Always	Doesn't criticize others.	

*This variable was multiplied by (-1) to give it a possitive connotation.

B Appendix B: Robustness Checks

Table B1: Schools with and without Prova Brazil: Balance

	Without Prova Brazil (1)	With Prova Brazil (2)	Difference (3)
Education			
Math score (standardized)	-0.7	-0.1	-0.6***
Portuguese score (standardized)	-0.6	0.0	-0.7***
5th grade completion rates	79.8	88.8	-9.0**
IDEB initial years	4.2	5.1	-0.9***
School characteristics			
Students per course	26.6	29.1	-2.5
INSE	47.1	50.3	-3.2***
School assets	9.4	9.3	0.1
Student characteristics			
Female	50.1	49.9	0.2
Black, brown and indigenous	73.1	68.0	5.2
Have failed a class or more	41.5	26.9	14.6***
Bathrooms >= 1	97.3	98.9	-1.6**
Bedrooms >= 1	97.7	98.9	-1.2*
Has computer	56.3	69.0	-12.7***
Mother & father finished primary education	65.6	72.4	-6.8
Mother or father finished high school	43.4	55.4	-12.0**
Parents encourage you to do homework	89.3	95.8	-6.5***
Teacher characteristics			
Female	75.8	80.7	-4.9
Black, brown and indigenous	54.6	55.1	-0.5
Salary over 2000 reais	64.3	70.7	-6.4
Works as teacher > 10 years	79.2	69.3	9.9
Works in the school > 5 years	51.0	39.7	11.3
Works in the same classroom > 5 years	65.6	41.4	24.2***
Fulfill >= 80% syllabus	32.8	50.1	-17.3*

Note: *p<.1; **p<.05; ***p<.01

Table B2: Family-wise Error Rate & False Discovery Rate in Municipality #3

	Reasoning	(1) Coefficient	(2)	(3)	(4)	(5)	(6)	(7)
			pvalue				qvalue	
			Original	Bonferroni	Sidak	Westfall-Young*	FDR	FDR Sharp
Mean (0-1)	**Logical**	**0.076**	**0.006**	**0.048**	**0.047**	**0.040**	**0.012**	**0.013**
	Abstract	**0.064**	**0.074**	**0.592**	**0.459**	**0.040**	**0.085**	**0.037**
	Spatial	**0.123**	**0.068**	**0.544**	**0.431**	**0.012**	**0.085**	**0.037**
	Total	**0.090**	**0.006**	**0.048**	**0.047**	**0.012**	**0.012**	**0.013**
Mistakes (#)	**Logical**	**-0.642**	**0.002**	**0.016**	**0.016**	**0.040**	**0.012**	**0.013**
	Abstract	**-0.590**	**0.088**	**0.704**	**0.521**	**0.040**	**0.089**	**0.037**
	Spatial	**-0.970**	**0.022**	**0.176**	**0.163**	**0.012**	**0.036**	**0.018**
	Total	**-2.138**	**0.004**	**0.032**	**0.032**	**0.012**	**0.012**	**0.013**

*Westfall-Young standard errors were run over clustered standard errors, but not wild bootstrapped.

Table B3: Family-wise Error Rate & False Discovery Rate in municipalities #1 and # 4

	Socioemotional	(1) Coefficient	(2)	(3)	(4)	(5)	(6)	(7)
			pvalue				qvalue	
			Original	Bonferroni	Sidak	Westfall-Young*	FDR	FDR Sharp
All	Frustration-tolerance	-0.034	0.552	1.000	0.992	0.695	0.663	0.496
	Assertiveness	0.019	0.888	1.000	1.000	0.695	0.888	0.570
	Active listening	**-0.149**	**0.036**	**0.216**	**0.197**	**0.015**	**0.092**	**0.102**
	Respect	**-0.078**	**0.046**	**0.276**	**0.246**	**0.207**	**0.092**	**0.102**
	Growth Mindset	-0.045	0.242	1.000	0.810	0.560	0.363	0.222
	Total	**-0.056**	**0.020**	**0.120**	**0.114**	**0.118**	**0.092**	**0.102**
Muni. #1	Frustration-tolerance	-0.127	0.100	0.600	0.469	0.595	0.231	0.239
	Assertiveness	-0.067	0.332	1.000	0.911	0.848	0.399	0.301
	Active listening	-0.073	0.154	0.924	0.633	0.848	0.231	0.239
	Respect	-0.166	0.152	0.912	0.628	0.323	0.231	0.239
	Growth Mindset	-0.050	0.618	1.000	0.997	0.848	0.618	0.301
	Total	**-0.088**	**0.002**	**0.012**	**0.012**	**0.409**	**0.012**	**0.013**
Muni. #4	Frustration-tolerance	-0.116	0.540	1.000	0.991	0.249	0.648	0.480
	Assertiveness	-0.028	0.666	1.000	0.999	0.705	0.666	0.500
	Active listening	-0.211	0.126	0.756	0.554	0.056	0.252	**0.202**
	Respect	**-0.138**	**0.002**	**0.012**	**0.012**	**0.174**	**0.012**	**0.013**
	Growth Mindset	-0.122	0.308	1.000	0.890	0.189	0.462	**0.337**
	Total	**-0.123**	**0.004**	**0.024**	**0.024**	**0.037**	**0.012**	**0.013**

*Westfall-Young standard errors were run over clustered standard errors, but not wild bootstrapped.

Table B4: Family-wise Error Rate & False Discovery Rate in Municipality #3

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Coefficient	pvalue				qvalue	
		Original	Bonferroni	Sidak	Westfall-Young*	FDR	FDR Sharp
Inhibitory control	**0.181**	**0.070**	**0.350**	**0.304**	**0.197**	**0.088**	**0.076**
Attention	0.081	0.706	1.000	0.998	0.607	0.706	0.165
Frustration	**0.239**	**0.004**	**0.020**	**0.020**	**0.176**	**0.020**	**0.021**
Aggressiveness	**0.268**	**0.066**	**0.330**	**0.289**	**0.023**	**0.088**	**0.076**
Total	**0.192**	**0.050**	**0.250**	**0.226**	**0.063**	**0.088**	**0.076**

*Westfall-Young standard errors were run over clustered standard errors, but not wild bootstrapped.

Table B5: Heterogeneous Treatment Effects

	Questionnaire			Prova Brazil	
	(1) Reasoning (0-1)	(2) Metacognition (1-5)	(3) Socioemotional (1-5)	(4) Math (standardized)	(5) Portuguese (standardized)
X: Number of students per grade<median					
Treated	0.000	-0.023	-0.071**	-0.018	-0.038
	(0.015)	(0.032)	(0.026)	(0.055)	(0.052)
X	-0.005	0.072	0.030	-0.025	-0.032
	(0.019)	(0.054)	(0.037)	(0.094)	(0.114)
Treated*X	-0.011	0.006	0.017	-0.122	-0.118
	(0.025)	(0.058)	(0.046)	(0.095)	(0.108)
X: Teacher has more than 10 years of experience					
Treated	-0.003	-0.139	-0.033	-0.013	0.041
	(0.040)	(0.093)	(0.062)	(0.204)	(0.221)
X	-0.003	-0.092	-0.008	-0.140	0.027
	(0.042)	(0.109)	(0.077)	(0.229)	(0.257)
Treated*X	-0.004	0.163	-0.043	-0.067	-0.150
	(0.047)	(0.127)	(0.074)	(0.239)	(0.244)
Teacher working in school for more than 5 years					
Treated	0.040	-0.073	-0.033	0.040	0.057
	(0.021)	(0.056)	(0.037)	(0.102)	(0.103)
X	-0.002	0.003	-0.020	-0.162	-0.047
	(0.029)	(0.078)	(0.045)	(0.173)	(0.150)
Treated*X	-0.106	0.127	-0.069	-0.227	-0.291
	(0.038)	(0.115)	(0.065)	(0.194)	(0.189)
Students afrodescendents/indigenous>median					
Treated	-0.009	-0.017	-0.107***	-0.040	-0.061
	(0.016)	(0.035)	(0.017)	(0.043)	(0.046)
X	-0.020	0.021	-0.027	-0.175**	-0.269***
	(0.025)	(0.061)	(0.043)	(0.054)	(0.064)
Treated*X	0.010	-0.007	0.090*	-0.009	0.021
	(0.021)	(0.057)	(0.032)	(0.084)	(0.081)
Teachers afrodescendents/indigenous>median					
Treated	0.008	-0.095	-0.135***	0.071	0.055
	(0.015)	(0.045)	(0.018)	(0.074)	(0.080)
X	0.011	-0.046	-0.101**	-0.008	-0.005
	(0.021)	(0.051)	(0.033)	(0.056)	(0.052)
Treated*X	-0.013	0.147	0.148**	-0.123	-0.127
	(0.036)	(0.085)	(0.046)	(0.075)	(0.078)
Mother or father finished high school>median					
Treated	-0.025	-0.041	-0.034	-0.156	-0.124
	(0.021)	(0.046)	(0.022)	(0.065)	(0.079)
X	0.010	-0.142*	-0.079*	-0.161	-0.072
	(0.018)	(0.050)	(0.029)	(0.076)	(0.088)
Treated*X	0.038	0.050	-0.048	0.192	0.103
	(0.031)	(0.061)	(0.030)	(0.108)	(0.120)
Control group mean	0.473	4.146	3.409	-0.043	-0.067
Observations	2,123	2,123	2,123	3,258	3,258
Pair FE	Yes	Yes	Yes	Yes	Yes

Note: SE clustered by school and wild bootstrapped to correct possible bias due to small number of clusters are presented in parentheses. *p<.1; **p<.05; ***p<.01.

Table B6: Estimations only with Prova Brasil Sample

	Questionnaire		
	(1) Reasoning (0-1)	(2) Metacognition (1-5)	(3) Socioemotional (1-5)
All			
Treated	-0.001	-0.037	-0.077***
	(0.012)	(0.034)	(0.018)
Constant	0.487	4.169	3.438
Standard Deviation	(0.190)	(0.646)	(0.459)
Observations	1,345	1,345	1,345
Pair FE	Yes	Yes	Yes
Muni. #1			
Treated	0.023	-0.165	-0.134
	(0.040)	(0.089)	(0.025)
Constant	0.371	4.245	3.382
Standard Deviation	(0.192)	(0.763)	(0.443)
Observations	138	138	138
Pair FE	Yes	Yes	Yes
Muni. #2			
Treated	-0.054	0.000	-0.041
	(0.018)	(0.070)	(0.035)
Constant	0.476	4.148	3.338
Standard Deviation	(0.178)	(0.682)	(0.455)
Observations	485	485	485
Pair FE	Yes	Yes	Yes
Muni. #3			
Treated	0.083***	-0.089	-0.061
	(0.019)	(0.058)	(0.046)
Constant	0.439	4.182	3.432
Standard Deviation	(0.171)	(0.601)	(0.467)
Observations	311	311	311
Pair FE	Yes	Yes	Yes
Muni. #4			
Treated	-0.009	0.000	-0.111***
	(0.010)	(0.027)	(0.017)
Constant	0.563	4.165	3.576
Standard Deviation	(0.187)	(0.599)	(0.429)
Observations	411	411	411
Pair FE	Yes	Yes	Yes

Note: SE clustered by school and wild bootstrapped to correct possible bias due to small number of clusters are presented in parentheses. *p<.1; **p<.05; ***p<.01.

Table B7: Estimation with Linear Probability Model

	Questionnaire			Personalized Test		Prova Brazil	
	(1) Reasoning (0-1)	(2) Metacognition (1-5)	(3) Socioemotional (1-5)	(4) Executive function (0-1)		(5) Math (standardized)	(6) Portuguese (standardized)
All				**All**		**All**	
Treated	-0.003	-0.050*	-0.045*	Treated	0.051	Treated	-0.014	-0.041
	(0.022)	(0.018)	(0.016)		(0.039)		(0.020)	(0.020)
Control group mean	0.261	0.274	0.268	Control group mean	0.228	Control group mean	0.261	0.272
Standard Deviation	(0.440)	(0.446)	(0.443)	Standard Deviation	(0.420)	Standard Deviation	(0.439)	(0.445)
Observations	1,743	1,743	1,743		527	Observations	2,532	2,532
Pair FE	Yes	Yes	Yes		Yes	Pair FE	Yes	Yes
Muni. #1						**Muni. #1**		
Treated	-0.034	-0.090	-0.109**			Treated	0.014	0.092
	(0.044)	(0.040)	(0.022)				(0.029)	(0.034)
Control group mean	0.268	0.293	0.309			Control group mean	0.248	0.216
Standard Deviation	(0.445)	(0.457)	(0.464)			Standard Deviation	(0.433)	(0.413)
Observations	294	294	294			Observations	272	272
Pair FE	Yes	Yes	Yes			Pair FE	Yes	Yes
Muni. #2						**Muni. #2**		
Treated	-0.056	-0.038	-0.030			Treated	-0.152***	-0.144**
	(0.033)	(0.034)	(0.027)				(0.028)	(0.033)
Control group mean	0.271	0.269	0.261			Control group mean	0.347	0.341
Standard Deviation	(0.445)	(0.444)	(0.440)			Standard Deviation	(0.477)	(0.475)
Observations	727	727	727			Observations	771	771
Pair FE	Yes	Yes	Yes			Pair FE	Yes	Yes
Muni. #3						**Muni. #3**		
Treated	0.161*	-0.058	-0.019			Treated	0.171**	0.139**
	(0.057)	(0.021)	(0.053)				(0.046)	(0.030)
Control group mean	0.199	0.278	0.252			Control group mean	0.166	0.177
Standard Deviation	(0.400)	(0.450)	(0.435)			Standard Deviation	(0.373)	(0.383)
Observations	311	311	311			Observations	367	367
Pair FE	Yes	Yes	Yes			Pair FE	Yes	Yes
Muni. #4						**Muni. #4**		
Treated	-0.011	-0.039	-0.048			Treated	0.010	-0.062*
	(0.023)	(0.028)	(0.027)				(0.018)	(0.019)
Control group mean	0.284	0.270	0.270			Control group mean	0.248	0.279
Standard Deviation	(0.452)	(0.445)	(0.445)			Standard Deviation	(0.432)	(0.449)
Observations	411	411	411			Observations	1,122	1,122
Pair FE	Yes	Yes	Yes			Pair FE	Yes	Yes

Note: The outcome variable takes the value of 1 if the score is in the top 25% of the distribution of grades within the municipality and zero otherwise.
SE clustered by school and wild bootstrapped to correct possible bias due to small number of clusters are presented in parentheses. *p<.1; **p<.05; ***p<.01.

Table B8: Estimations with Prova Brasil 2015

	Prova Brazil	
	(5) Math (standardized)	(6) Portuguese (standardized)
All		
Treated	-0.089	-0.060
	(0.041)	(0.051)
Control group mean	0.139	0.148
Standard Deviation	(1.035)	(1.005)
Observations	2,704	2,704
Pair FE	Yes	Yes
Muni. #1		
Treated	0.029	0.090
	(0.137)	(0.216)
Control group mean	-0.536	-0.451
Standard Deviation	(0.838)	(0.882)
Observations	217	217
Pair FE	Yes	Yes
Muni. #2		
Treated	-0.250**	-0.202
	(0.078)	(0.073)
Control group mean	-0.341	-0.295
Standard Deviation	(0.751)	(0.872)
Observations	696	696
Pair FE	Yes	Yes
Muni. #3		
Treated	-0.037	-0.051
	(0.081)	(0.093)
Control group mean	0.005	0.101
Standard Deviation	(0.911)	(0.907)
Observations	576	576
Pair FE	Yes	Yes
Muni. #4		
Treated	-0.045	-0.013
	(0.061)	(0.083)
Control group mean	0.522	0.466
Standard Deviation	(1.066)	(0.999)
Observations	1215	1215
Pair FE	Yes	Yes

Note: SE clustered by school and wild bootstrapped to correct possible bias due to small number of clusters are presented in parentheses. *p<.1; **p<.05; ***p<.01.