Inter-American
Development Bank

Does Gamification in Education Work?

Experimental Evidence from Chile

Roberto Araya
Elena Arias Ortiz
Nicolas Bottan
Julian Cristia

Department of Research and
Chief Economist

DISCUSSION
PAPER Nº
IDB-DP-683

May 2019

Does Gamification in Education Work?

Experimental Evidence from Chile

Roberto Araya*
Elena Arias Ortiz**
Nicolas Bottan***
Julian Cristia**

* Universidad de Chile
** Inter-American Development Bank
*** Cornell University

May 2019

IDB
Inter-American
Development Bank

http://www.iadb.org

The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the Inter-American Development Bank, its Board of Directors, or the countries they represent.

Does Gamification in Education Work? Experimental Evidence from Chile[1]

Roberto Araya
Universidad de Chile

Elena Arias Ortiz
Inter-American Development Bank

Nicolas Bottan
Cornell University

Julian Cristia
Inter-American Development Bank

Abstract

Gamification, or the introduction of game elements to non-game contexts, has potential to improve education but there is little rigorous evidence about its effectiveness. In this paper, we experimentally evaluate an innovative technology program that uses gamification to spur student motivation and increase math learning in low-performing primary schools in Chile. The *ConectaIdeas* program involves two weekly sessions in a computer lab in which students use an online platform to solve problem sets. The platform tracks individual students advances, compares these advances to those from classmates and features different types of group competitions to promote student motivation. Results indicate large positive effects on math learning, of about 0.27 standard deviation, on the national standardized exam (no effects were found on language). The program also affected several non-academic outcomes. On one hand, it increased students' preference towards using technology for math learning and promoted the idea among students that study effort can raise intelligence. On the other, the program increased math anxiety and reduced students' preference towards teamwork. These effects suggest that gamification could be an important tool to boost learning, but that it may bring unintended consequences.

[1] The project would not have been materialized without the support and commitment provided by Ben Petrazzini. We want to especially thank for their significant contribution to Nora Cuevas, Alcaldesa de San Bernardo and by Ruth Carrillo, Dina Herrera, José Sotelo, and Julio Torres. We also thank Maritza Moraga, Jimena Romero, Obed Ulloa and Anastasiya Yarygina for their outstanding research assistance. The team at the Centro de Investigacion Avanzada en Educacion of the Universidad de Chile, including Carlos Aguirre, Raúl Gormaz, Paulina Jaure, and Arturo Prat, played a critical role in the implementation of this project. We are grateful to María Paz Espinosa and Camila González Nakada from MIDE UC and Cecilia Lecaros from APTUS for their support in the data collection activities. We acknowledge funding from the International Development Research Center, the Inter-American Development Bank, and PIA-CONICYT Basal Funds for Centers of Excellence Project FB0003. The views expressed in this paper are those of the authors and should not be attributed to the Inter-American Development Bank.

1. Introduction

The introduction of game elements to non-game contexts -known as gamification- has become an increasingly common strategy used in education, health, and business to motivate individuals to undertake desired behaviors. For example, the device "Fitbit" tracks the number of steps that a person takes in a day, provides a congratulatory message when a targeted number of steps is achieved, and enables competitions among users to further spur motivation. As this case exemplifies, the basic idea behind gamification is that the introduction of simple game elements, such as points, badges, and leaderboards, can transform a dull task in an engaging activity.

In Education, gamification can play an important role considering that student motivation has long been recognized as critical for learning (Weiner, 1990). However, there may be drawbacks to its use related to potential reductions in intrinsic motivation, increases in anxiety, and short-lived effects on engagement. In spite of these potential drawbacks, gamification in education is a flourishing industry. Fueled by the worldwide rise in access to internet-connected devices, companies such as Duolingo and Khan Academy support more than 10 million of students every month. Research in this topic has also increased markedly. According to Google Scholar, the number of papers published in a year that contain the words "gamification" and "education" jumped from 140 in 2010 to 3,570 in 2014 and reached 9,570 in 2018. But, does gamification in education work? That is, do educational programs that introduce game elements to spur motivation generate large learning gains? Unfortunately, in spite of the large number of studies on gamification in education, there is a dearth of rigorous empirical evidence addressing this important question.

This paper seeks to contribute to filling this gap by experimentally evaluating a program that uses gamification intensely to improve academic achievement. The program, called ConectaIdeas, aims to improve math learning among fourth-grade students in Chile. Participating students practice math exercises in an online platform during two weekly 90-minute learning sessions that took place in regular school time. The program employs an array of gamification strategies to promote high use of the learning platform. First, the platform shows each student how many accumulated exercises he or she has performed and compare this figure with the average of the class. Moreover, students can observe the number of exercises completed by each student in the class. Second, personalized "ads" are shown regularly to motivate students by stressing the notion that intelligence can be improved by exerting effort while studying. Third, whole sections of students participate in weekly competitions with sections in other schools based on the average number of exercises completed on the platform. Fourth, sections also participate in inter-school “live” tournaments every two months in which students are paired to compete in solving mathematical problems embedded in an online game.

Based on these features ConectaIdeas seems to be a promising program, but does it impact student learning? To answer this question, we conducted a randomized controlled trial in 24 public primary schools, attended by low-income students in Santiago, Chile. Students were not only socioeconomically disadvantaged but also lagged in learning: they scored, on average, about 0.7 standard deviations below the math national average. We randomly selected one fourth-grade section within each of these schools and assigned it to the treatment group and assigned the other section in that grade to the control group. We collected baseline data in March 2017, and the program was implemented immediately after and until November 2017 (the school year in Chile runs from March to December).

Our primary outcome is obtained from the Chilean national standardized examination applied in November 2017 (after 7 seven months of program exposure). This is a yearly assessment implemented in all schools by the Ministry of Education to monitor learning in math and language. Measuring effects on this test is important because evidence shows that effect sizes vary considerably between different types of tests. In fact, Hill et al. (2008) reviewed experimental evaluations in education and documented that the average effect on broad standardized tests was 0.07 standard deviations, compared to an average effect of 0.23 for narrow standardized test and to 0.44 for a specialized test developed for a specific intervention. Hence, using a broad national standardized examination as the primary evaluation outcome allows estimating how a potential program scale-up may impact the main assessment used by a Ministry of Education to monitor learning quality and equity.

Results indicate that ConectaIdeas generated a large statistically significant improvement in math learning. Our preferred specification shows an effect of 0.27 standard deviations. The estimated effects on math under alternative specifications range from 0.22 to 0.29 standard deviations. The effects on math achievement are similar across different sets of students defined by gender, mother’s education, and baseline achievement. Even though the program aimed to improve learning in math, it could have generated spillovers to other subjects. Nevertheless, estimated effects on language are close to zero and not statistically significant.

To benchmark the effects, we compare them with those from other educational evaluations that have also analyzed effects on the Chilean national standardized exam. One important evaluation is the one that assessed the effects of extending the school day from 4 to 7 hours a day in schools in Chile. This landmark program, which entailed a massive increase in educational spending, generated increases in math and language of about 0.06 standard deviations (Bellei, 2009). In turn, a program that provided lesson plans to teachers and teacher materials improved math and language test scores in about 0.07 and 0.09 standard deviations, respectively (Bassi et al. 2016). Hence, the math effects of ConectaIdeas are about four times larger compared with those from these two studies.

We also analyze whether the program affected students' perceptions in different areas. On the positive side, we find evidence that the program increased students' preference towards using computers for math instruction which may be important in a context of rising access and use of technology across life domains. We also find some evidence that the program increased the likelihood that students believe that exerting effort while studying can increase intelligence. We find no evidence of effects on math intrinsic motivation or in math self-efficacy. On the negative side, we find that the program increased anxiety associated with studying math and also reduced preferences toward collaborating in teams.

We exploit individual-level granular data recorded on the learning platform to document how much, when, and where students used the online platform. We find that virtually all students used the platform and that the average student used the platform for 28 hours during the school year. A key question is whether the positive academic effects can be partly explained by students practicing math at home. However, the evidence is unequivocal on this point: home use accounts for a mere 2 percent of the logged in time and, hence, it cannot explain the results found. We also document that the time spent on the platform remains largely flat during the seven-month period of program exposure. This finding contrasts with the sharp decreasing use over time seen in programs that provided laptops or internet for home use (Malamud et al., 2018). Therefore, the ConectaIdeas program was able to deal with the challenge of strong novelty effects found in these other programs.

The experimental estimates correspond to the implementation of the program during the 2017 school year. But is the large effect documented just a one-off result, or instead does it represent the typical effect of the program? To explore this issue, we generate non-experimental estimates exploiting the implementation of ConectaIdeas in 11 schools in Santiago, Chile between 2011 and 2016, together with school-level longitudinal data from the national standardized examinations. Using a difference-in-differences framework, we find that ConectaIdeas generated positive and statistically significant effects of between 0.19 to 0.22 standard deviations on math and no statistically significant effects on language. These results suggest that the large experimental estimates described earlier can be representative of the typical effect of the program.

Our study is related to a large literature from Education and Computer Science that has analyzed different aspects related to the use of gamification in education. Studies have theoretically analyzed the potential advantages and disadvantages of different models of gamification in education, documented examples of its introduction in particular contexts, and provided some qualitative and quantitative evidence regarding its effects on student outcomes. Reviews of this literature have generally concluded that incorporating gamification can increase student motivation and engagement (Lister, 2015; COMPLETE). However, there is little rigorous empirical evidence on the effects of educational interventions that use gamification on academic achievement.

To the best of our knowledge, there are no studies from the Economics literature that have rigorously evaluated the effects of a program that used gamification intensely to improve learning outcomes. However, there are two strands of the Economics literature that are linked to our study. The first strand includes evaluations of interventions that used monetary incentives to increase student motivation. Studies that have evaluated the effects of providing monetary incentives to students have found, in general, positive though modest effects on academic achievement (Bettinger, 2012; Fryer 2011). One exception to this finding is the study by Li et al. (2014) that reports that when incentives were provided to individual students, the learning effects were small, but that when the incentives where provided to promote group competitions (and within-group collaboration), then the learning effects were large. The second strand includes experimental studies that evaluated the learning effects of computer-assisted instruction programs. Experimental evaluations implemented in India (Banerjee et al., 2007 and Muralidharan et al., 2019), China (Lai et al., 2013; Mo et al., 2013; Lai et al., 2015) and the US (Dynarski et al., 2007; Wijekumar et al., 2009; Rutherford et al., 2014) have showed positive learning effects of these interventions though the effects for programs implemented in the US have been considerably smaller.

The main contribution of this study is that it presents a comprehensive experimental assessment of the effects of an educational program that uses gamification intensely. In particular, the study presents a number of advantages that are summarized next. First, it presents unbiased and highly precise estimates due to the within school-randomization design and the large number of students participating in the study (about 1,100). Second, it evaluates a program that is implemented in public schools during regular school time, which is relevant for considering future scale-up. Third, it measures effects on academic achievement using a broad national standardized examination. Fourth, the study also reports program effects on intrinsic motivation, self-efficacy, anxiety, growth mindset, and preferences for teamwork and towards the use of technology at school. Finally, the study complements the one-year experimental estimates with non-experimental estimates from several years to provide a more definitive assessment of program effects.

The paper is organized as follows. Section 2 describes the ConectaIdeas Program. Section 3 details the experimental design, data, identification strategy, and documents baseline balance. Section 4 presents the main effects on learning measures and on non-academic outcomes while section 5 presents additional results including robustness checks and non-experimental estimates of the program effects. Finally, section 6 concludes.

2. The ConectaIdeas Program

The ConectaIdeas program was developed by a team led by the researcher Roberto Araya at the Centro de Investigación Avanzada en Educación at the Universidad de Chile. The team aimed to design a program that could generate large increases in math learning among low socioeconomic students. The guiding principle behind the project was that the introduction of game elements to math instruction, facilitated by the use of technology, could boost student motivation, and lead to fast learning. After years of small-scale development, the ConectaIdeas program was implemented from 2011 to 2016 in 11 schools in the community of Lo Prado in Santiago, Chile. During this period, the team streamlined the design and developed detailed implementation protocols.

The program implemented in the 2017 experimental evaluation entailed providing students two weekly 90-minute math learning sessions in the computer lab. One of these sessions replaced traditional math instruction in the classroom while the other session represented additional instructional time. In a typical session, all students worked solving the same set of 20 to 30 exercises assigned to them that are aligned to the topics covered in regular math instruction. When solving these problems, students received automatic feedback regarding whether their answers were correct or not. Lab coordinators, hired and supervised by the team at the Centro de Investigación Avanzado en Educación, were responsible for conducting the learning sessions in the computer lab, in collaboration with regular classroom teachers. Lab coordinators were former teachers who received a one-day training and ongoing supervision from the implementation team (teachers did not receive formal training but learning-by-doing was promoted). The infrastructure needed to implement the program included a computer lab with one computer (or tablet) per student and one broadband internet connection that was shared among students. Because, the program did not involve the use of videos, the internet requirements were limited.

The program includes several gamification strategies to ensure high use of the learning platform. Figure 1 contains a screenshot of the platform that depicts several of these strategies. The first strategy is centered on motivating students by keeping track of their advances and making comparison between the student and her classmates. As Figure 1 shows, the student is presented with a graph that plots the number of accumulated exercises that she has completed by each week (the dark blue line in the graph). Showing this information is intended to motivate the student by making her effort visible and concrete. Moreover, the graph also includes a line for the class average (the light blue line in the graph in Figure 1). Presenting this information seeks to activate the motivational effects embedded in social comparisons that have shown to be important in different domains such as energy conservation and worker effort (Cialdini et al., 2007; DellaVigna and Pope, 2018). More detailed social comparisons are also presented in a different screen that can be accessed by students which shows the ranking of students in the class ordered by the number of accumulated exercises completed. Students can choose to see this ranking using data for the entire year, the past week, or the current session. Providing these different time horizons for making comparisons could provide a motivation to students that may be far behind in the annual accumulated statistic but that could fare better in the past week or the current session.

The second strategy seeks to motivate students by conveying the idea that intelligence is malleable and that it can be improved by exerting effort while studying. To that end, the

ConectaIdeas platform presents to students personalized “ads” that emphasize this message. Figure 1 shows an example of one of these ads. In this case, the student is presented with an image of a child playing the piano and a written message stating that “Effort, and only effort, Student name, is the road to perfection” (where Student name is replaced by the actual name of the student logged in that computer). These images and the written messages are presented for 20 seconds after the student enters this screen and they are accompanied by a computer-generated audio of the message. The images and messages presented are rotated from a library of 10 examples that promote the general theme of the importance of practice and effort. For example, another message states that “the brain is like a muscle, the more you practice, Student name, the stronger it becomes.”

In contrast to the first two strategies, the third strategy focus on group motivation rather than on individual motivation. In particular, competitions are set in which sections of students participating in the program try to outperform other sections in terms of the average number of exercises completed each week. Returning to Figure 1, we see that on the right hand side of the screen, pictures of different sections of students are shown. This is the ranking of sections participating in the program that are ordered from top to bottom by the average number of exercises completed in the week. Note that only some sections are shown in this ranking. That is, the section shown in the middle (i.e. in the fourth position counting from top to bottom) corresponds to the photo of the section of the student logged in the computer. The top three sections shown correspond to the three sections that are just above her section in terms of number of completed exercises and the bottom three sections are those that are immediately below in the ranking. During the learning session, the ranking of sections is updated and, hence, the section of the student will continue to be shown in the fourth position but will be surpassing other sections, which are typically inactive at that moment, whose pictures will move down in the ranking. The photo of the section of the student logged in could attain one of the top three positions if it achieves one of these positions among all sections participating in the program. Finally, students can click any of the pictures shown in this ranking to know more details about the competing sections such as the name of the school, grade and section identification.

The fourth strategy also seeks to motivate students by activating social dynamics and within-class collaboration. To that end, “live” tournaments are organized every two months in which students compete to solve math exercises embedded in an online game. For this tournament, a time is scheduled in which all participating sections in the ConectaIdeas program should be connected to the platform. Then, each student in a section is paired with another student in a different school. The two paired students play the “spiral game,” shown in Figure 2, in which they take turns at solving math exercises and they move “tokens” with the objective of placing all of them at the center of the spiral (the cell numbered 143). In the screen shown, Student 1, has to solve the question displayed on the right lower corner (“what is the number than when subtracted from 30 the result is 18?”) by moving one of her tokens to the cell number 12. Next, Student 2, will have her turn at solving a math exercise moving her token. The game has other features that provide additional positive or negative payoffs to the players and also require students considering different strategies in their plays (basically, which token to move in each play). Every 5 minutes the points collected by each player are recorded and average for each section and the ranking of the participating schools is shown. During the competition, a program staff acts as the “announcer” and informs students about the ranking, makes comments about how the competition is evolving and tries to drive excitement among participants.

As the platform is used by students, it generates data that can be used to monitor student learning and support instruction. In particular, the platform provides teachers and lab coordinators a dashboard with real-time information on individual student advances. In this dashboard students are ranked from those that are in most need of support (those that have answered few questions and/or that have received a low grade in the questions answered) to those that need less support. The platform also presents a dashboard that shows the fraction of correct responses by question (to help identifying questions for which all students need support). Finally, the system also generates reports that are emailed to lab coordinators, teachers, and principals to provide an overall snapshot of how the class is advancing.

3. Research Design

3.1. Design and Sample Selection

We implemented a randomized controlled trial to assess the causal effect of the ConectaIdeas program. The ConectaIdeas team was tasked with recruiting 24 public schools that satisfied certain initial eligibility requirements. First, schools had to be located in the Santiago metro to simplify logistics given that the ConectaIdeas team is based on that city. Second, because of the importance of finding programs that can improve learning for disadvantage populations, eligible schools had to be classified in the two lowest socioeconomic status categories out of the five categories determined by the Ministry of Education. Third, the school had to be urban and count with at least two fourth grade classrooms (to randomize one section to the treatment group and one to the control group).

The recruitment process began in the end of January 2017 (three months before the start of the school year) and consisted of three steps. First, the ConectaIdeas team identified 22 school districts (*comunas*) that had sufficient schools satisfying the criteria described above. An email invitation was sent to the directors of these 22 school districts, which were followed up by calls. The team then visited 11 districts that replied and expressed interest. The second step consisted of arranging information sessions with the district director and directors of schools that satisfied the eligibility criteria. These sessions were conducted with 9 school districts.

In the third and final step, the ConectaIdeas team conducted school technical visits to verify whether the school counted with the minimum required technology to accommodate the program. This consists of a computer lab with at least 30 computers, and adequate internet access. Even though this was a requirement, it was not completely binding because the ConectaIdeas program could supplement the necessary equipment in some instances if there were infrastructure deficits (e.g., unreliable internet connection). The technical visits were scheduled to be conducted in 31 schools in 6 school districts. These target schools were not only identified because of all the requirements listed above, but also by proximity to each other in order to reduce transport costs. In the end, there were three schools in one district that were never visited because the ConectaIdeas team had already confirmed the 24 participating schools that were located across 4 school districts. All target schools in the districts of La Pitana and Quinta Normal participated. Additionally, nine out target schools in the San Bernardo district and six out of seven schools in the Maipu district participated in the study (the two target schools that were not included in the study were dropped because they did not meet the technical requirements). In conclusion, the recruitment procedure did not involve decisions from individual schools to either participate or not to the program (i.e. there were no possibilities for schools to "self-select" into the program).

Table 1 presents statistics obtained from the 2016 national standardized examination (known in Chile as "Sistema de Medición de Calidad de la Educación" or *SIMCE*) that allow understanding how the sample construction process unfolded. Column (1) presents means for the universe of the schools in the country and columns (2) to (5) presents means for samples of schools that result from restricting the sample progressively to include the eligibility requirements. In particular, column (2) restricts to schools in the Santiago metropolitan area and column (3) further restricts the sample to schools in the bottom two categories in terms of socioeconomic status. Next, column (4) further restricts the sample to schools with at least two sections in fourth grade and column (5) presents the sample of schools participating in the study. The table shows that the makeup of the study sample is quite similar to the sample of low socioeconomic status schools in Santiago with two exceptions: enrollment in the study schools is larger (due to the two-section restriction) and their students perform even worse in math and language. In fact, students in the study sample underperform the average student in the country by 0.60 standard deviations in language and 0.68 in math.[2]

We adopted a within-school, section-level randomization design. Within each of the 24 participating schools, we randomly assigned one of the two fourth-grade sections to the treatment group. These sections participated in the ConectaIdeas program. The other sections were assigned to the control group and received traditional math instruction. For the three schools in the sample that had more than two sections, we only included the first two (i.e., A and B sections) in the evaluation. The randomization was conducted before baseline data was collected, and schools were informed of the treatment status of each section *after* the baseline was collected in March 2017. Subsequently, we documented perfect compliance of program assignment to treatment. That is, all sections assigned to treatment participated in ConectaIdeas and none of the sections assigned to the control group participated in the program.

3.2. Identification Strategy

Evaluating the program effects is straight forward due to random assignment of schools to treatment within schools. The advantage of this design is that we are able to account for school characteristics that may influence both sections by including school fixed-effects. Additionally, because the intra-cluster correlation at the section level is close to zero, once school fixed-effects are added, our design is almost as precise as a design featuring individual-level randomization.

We estimate the effects of the program under two main specifications. The first specification involves estimating the following equation:

[2] Test scores in the national standardized exam are normalized using the nationwide mean and standard deviation.

$$y_{ics}^{post} = \alpha_1 + \beta_1 * Treatment_{cs} + \phi_s + \varepsilon_{ics} \qquad (1)$$

where y_{ics}^{post} is the outcome variable in the post period (e.g., the math test score measured in the national standardized exam) for student i, in section c , in school s. $Treatment_{cs}$ is an indicator variable that equals one if the section was assigned to the treatment group and zero if not. ε_{ics} is the error term, which should be uncorrelated with the treatment assignment because of random assignment, and ϕ_s are school fixed effects. Our main coefficient of interest, β, estimates the average treatment effect of the program on the outcome variable (e.g., the average difference in math test score between students in the treatment group compared to those in the control group).

Because learning is strongly correlated over time, controlling for the baseline test scores can increase statistical precision. Moreover, doing so can account for potential differences in baseline test score levels. Consequently, the second specification that we use is similar to the first one but also controls for the baseline value of the outcome:

$$y_{ics}^{post} = \alpha_2 + \beta_2 * Treatment_{cs} + \gamma_2 * y_{ics}^{pre} + \phi_s + \varepsilon_{ics} \qquad (2)$$

where y_{ics}^{pre} is the baseline test score in the respective subject. That is, when estimating effects on math, we control for the baseline math test score and when estimating effects in language we control for the baseline test score in that subject. This is our preferred specification because it controls for potential baseline differences in outcomes and because it should generate more precise estimates.

Finally, all estimates presented throughout the paper will include heteroscedasticity-robust standard errors that are clustered at the section level (the unit of randomization). One potential concern is that because we are clustering standard errors among 48 sections, we might be overstating the precision of our estimates due to a relatively modest number of clusters (Cameron and Miller, 2015). Consequently, in the robustness section we show additional results using alternative strategies to compute appropriate standard errors.

3.3 Data

Our analysis relies on a combination of administrative records from the Chilean national standardized exam, survey data, and administrative program data from the ConectaIdeas platform. The survey data was collected in 3 different waves during the academic year: i) a baseline survey conducted in March 2017; ii) a midline survey conducted in August 2017; and iii) an endline survey conducted in November 2017.

The main outcome of the study corresponds to math test scores constructed from the national standardized exam applied in November 7 and 8, 2017. Effects on language on this assessment were also analyzed to explore potential spillovers on this subject. The national standardized exam is conducted annually since 1998 among all fourth grade students at the end of the academic year and it is reported and widely used for monitoring educational outcomes.[3] Moreover, these tests are important for teachers and principals because they are linked to monetary incentives and low scores can trigger administrative actions including visits to schools and the introduction of changes in how schools are managed.

The baseline tests on math and language were conducted as part of this study right before the program started. The objective of this baseline assessment was to document students' learning levels in math and language.[4] These test scores are used to check baseline balance, are useful to improve the precision of the estimated effects (McKenzie 2012) and makes it possible to explore heterogeneous program effects by baseline academic achievement.

In the endline survey, also conducted as part of this study, students completed a questionnaire designed to capture program effects on the following areas: math self-concept, math intrinsic motivation, preference for having math lessons in the computer lab (as opposed to the regular classroom), having a growth mindset (Claro et al. 2016), and preference for teamwork. These primary data on students' perceptions were complemented with secondary data from a questionnaire included in the national standardized exam that explores whether students' math self-concept and whether students have anxiety related to math tests, grades, and homework.

Finally, we analyze data from the ConectaIdeas platform to document how computers were used for math instruction. The unit of observation of these data is an exercise solved by a

[3] The national standardized exams are run by the Agency for Education Quality, which is external and independent of the Ministry of Education, since 2012. The agency tests students every year in fourth grade and every two years in sixth grade.

[4] The tests used in the baseline examination were developed by the firm APTUS.

student. These data include the time spent in each exercise, when this happened, whether the exercise was answered correctly and also whether the student requested help from a classmate and whether the student did received help. Using these data, we are able to provide a comprehensive snapshot about how students used the ConectaIdeas platform.

3.4 Randomization and Balance

In this section we analyze whether the randomization generated similar treatment and control groups. To that end, Table 2 presents means for the treatment and control groups (in columns 1 and 2, respectively) for baseline test scores and student characteristics. In turn, column (3) presents estimated differences between the treatment and control groups controlling for school fixed-effects. Finally, column (4) presents the sample size for each variable analyzed.

Baseline test scores, collected in March 2017 just before the program started, allow analyzing whether the academic achievement of treatment and control students was similar before the program was implement. Results indicate no statistically significant differences in language test scores across groups. However, students in the treatment group on average underperformed those in the control group by 0.08 standard deviations in the math test and this difference is significant at the 10 percent level. Though this is a modest difference in performance, still these results provide additional motivation to control for baseline academic achievement when estimating treatment effects on academic achievement.

To further explore the similarity of the treatment and control groups, Table 2 presents statistics for student characteristics constructed using data from the questionnaire applied together with the national standardized exam in November 2017. Results indicate that the composition of the treatment and control groups are quite similar. The differences in the analyzed characteristics are small and only statistically significant at the five percent level for the variable regarding mothers' education (47 percent of students in the treatment group reported that their mothers completed secondary education compared to 53 percent in the control group).[5]

[5] The main sample in the paper includes students that participated in the baseline math exam and in the 2017 fourth-grade national standardized math exam. Consequently, this is the sample that is analyzed when exploring differences in student characteristics in Table 2.

4. Main results

4.1. Evidence on platform use

We start by documenting how much, where and when the ConectaIdeas platform was used. To explore these questions, we exploit rich individual-level longitudinal data on platform use. We focus on the use between the end of March, when the intervention was started, and November 6 (right before the national standardized exam, which took place on November 7 and 8, 2017).

Log data shows that the platform was intensively used: the average student used it for about 27 hours. However, it is not the case that students used the platform every day for a few minutes. Rather, the average student was connected to the platform 43 days (in a period of about 210 days) and each time she used it for 39 minutes. And the use was heavily concentrated at school, which accounted for 98% of the total platform use.[6] These results are further corroborated by Figures 1 and 2 which show that the use of the platform was heavily concentrated from Monday to Friday and during the times of the day when schools were opened. These results point to the central role that school use played in explaining the results.

Now, if the gamification features built-in ConectaIdeas generated intensive use of the platform use at school, why they did not produce a high use at home? The low use at home can be considered a design issue. Because in the ConectaIdeas platform students can only work in exercises that have been assigned to them by their teachers, if students are not assigned exercises to do during the weekend, they cannot use the platform to practice. And lab coordinators and teachers were not instructed to assign exercises to students during the weekend to practice. Hence, in future implementations of the ConectaIdeas program, it would be interesting to explore whether platform use at home can also contribute to improved learning by assigning exercises to students as homework or as a supplementary voluntary activity.

Using data from the platform we also document that there was little heterogeneity across schools in terms of the numbers of technology sessions implemented. The average school implemented 49 math technology sessions and the 10th and 90th percentiles stand at 42 and 55 sessions. Finally, Figure 3 presents the distribution of platform use by month. Results indicate

[6] We classify use as "in school" if it took place in days in which schools were opened (i.e. weekdays that were not holidays or vacation) and between the times that schools were open.

that platform use was similar throughout the school year, excluding months that were special in some way.[7]

4.2. Effects on Academic Achievement

This subsection answers the main question of the study: did ConectaIdeas affect student learning? To answer this question, Table 3 presents program effects on math academic achievement measured in the 2017 national standardized exam. Results indicate that the ConectaIdeas generated large effects on math achievement. In the first specification, which does not control for baseline math achievement, the estimated effect is 0.22 standard deviation. In our preferred specification, which controls for baseline math achievement, the effect is slightly larger at 0.27 standard deviations. In either case, the estimated effects are statistically significant at the one percent level.

Even though the program focused exclusively on math, it could have generated spillover effects into language. For instance, the program could have motivated students to study more overall, or it could have discouraged students in language class if they shifted study time to math instead. Results indicate that the program did not affect language achievement. In both specifications, the estimated effects on language are small and not statistically significant.

The documented math effects seem large not only when compared with those from other educational evaluations conducted in Chile (as discussed in the introduction) but also when compared with common policy benchmarks. One policy benchmark relates to achievement gaps between students from different socioeconomic background (Hill et al., 2008). In particular, Chilean fourth graders taking the math national standardized exam whose mothers finished secondary school outperform their counterparts whose mothers did not finish this education level by 0.51 standard deviations. Hence, ConectaIdeas could close about 50% of this socioeconomic gap (0.27/0.51). A second commonly used benchmark relates to comparing the effects with the usual learning progression that students experience in one year. Unfortunately, we do not count with data from Chile about how much students improve their academic achievement in math in one year. However, Hill et al. (2008) document that fourth graders in the US improve their learning in 0.52 standard deviations in a year. Assuming that student academic progression in

[7] Use in March and November is minimal because these months were just only partially included in the time windows for this analysis. And use was low in April because schools were entering the program during the month and in July because of the 2-week winter vacation.

Chile is similar to the U.S., we can think that students that participated in ConectaIdeas advanced about 50% more than their counterparts in the control group (0.27/0.52).

We now turn to whether the ConectaIdeas program generated different effects on sub-samples defined by gender, mothers' education and baseline academic achievement. Table 4 presents these results. In what follows, we focus the discussion of effects on math scores because this is subject targeted by the program. To start with, effects are slightly larger for boys than for girls (0.29 versus 0.24 standard deviations) but these effects are not statistically significantly different. When exploring effects by mothers' education we find that these are almost equal (0.28 versus 0.29 standard deviations). This pattern of similar effects across subsamples is also present when we divide the sample by baseline academic achievement. That is, effects for students that scored below the median at the baseline math test are identical to those that scored above the median at baseline. To sum up, results indicate the the positive effects of ConectaIdeas were experienced by different subpopulations of students defined by gender, mother's education and baseline academic achievement.

4.3. Effects on non-academic outcomes

As discussed in the first section, introducing game elements to learning activities can generate effects on a range of outcomes beyond math and language academic achievement. In fact, it has been recognized that gamification may directly affect other outcomes such as engagement, intrinsic motivation, and anxiety. Moreover, considering the particular gamification features embedded in ConectaIdeas, we could expect potential effects on a number of dimensions. Armed with data from the endline student survey that we collected as well as from the questionnaire that was applied together with the national standardized examination, we provide evidence on this issue.

To that end, we construct indices based on responses to questions measuring relevant outcomes. For example, we construct an index for intrinsic motivation using 9 items included in the endline student survey that were translated to Spanish from the scale used in the TIMSS math fourth-grade examination from 2015. All items are transformed into dummy variables that equal 1 if the student agrees with a statement, standardized using the mean and the standard deviation, average across items for the same construct and later standardized again for easier interpretation.[8] Table 5 presents the estimated effects obtained running regressions of these indices on a treatment dummy and school-fixed effects (i.e. estimating equation 1).

Results indicate positive effects on two areas that are well aligned to prior expectations. To start with, the basis of gamification is producing a more engaging and attractive experience and indeed 79% of students in the treatment group report preferring doing math sessions in the computer lab compared to only 59% of students in the control group. This result translates to a treatment effect of 0.41 standard deviations in students' preferences towards doing math lessons in the computer lab. In addition, one of the ConectaIdeas features involved presenting personalized ads to students to motivate the adoption of a growth mindset. And we document a positive effect of 0.09 standard deviations in this area.

In contrast, there are two areas in which we do not find statistically significant effects though some effects could have been expected. The first one is on intrinsic motivation, that is, the inherent enjoyment of learning math per se. Because ConectaIdeas emphasizes doing math exercises to increase scores and fare better in individual and group competitions, it may reduce math intrinsic motivation, but this hypothesis is supported by the reported results. The second one is on math self-efficacy or the self-perception regarding students' own abilities to solve math exercises. Because ConectaIdeas produced large increases in math achievement, we could expect positive effects on this area, but this is backed by the analysis.

Finally, there are two areas in which we find effects that can be considered as undesirable. In particular, we found positive statistically significant effects on math anxiety of 0.13 standard deviations that could be linked to the social comparisons and individual and group competitions that are built in ConectaIdeas. Moreover, we document negative statistically significant effects on a scale of preference for teamwork of 0.21 standard deviations.[9] This result can be surprising considering that ConectaIdeas promoted within-class collaboration by setting up group competitions. One potential explanation for this unexpected result is that some students may notice the downsides of working in teams (e.g. the weaker link between own performance and final outcomes) when participating repeatedly in team competitions.

[8] All data definitions available in the Appendix.

[9] In this analysis we checked effects on six different outcomes. Because were doing multiple hypothesis tests, the probability of finding an statistically significant result is an outcome is heightened. Hence, we follow ... and compute q-values which are analogous to p-values but that incorporate multiple hypothesis adjustments. All results described are still statistically significant after this adjustment with the exception of the positive effects on growth mindset that have an associated q-value of 0.16.

The main conclusion of this analysis is that ConectaIdeas did not only affected academic outcomes but also generated a range of effects on different areas. On the positive side, we document increases in students' preferences towards using ConectaIdeas for math lessons and also an increase in students' likelihood of adopting a growth mindset. In turn, we do not find statistically significant effects on math intrinsic motivation and math self-efficacy. On the negative side, we document that ConectaIdeas generated an increase in math anxiety and a decrease in preferences towards teamwork.

5. Additional results

5.1. Robustness checks

For the main results on academic achievement presented in section 4.2, we compute standard errors clustered at the section level. Because the number of clusters may seem limited (48), it can be the case that standard formulas used to compute the standard errors may generate conservative estimates. To tackle this issue, we have computed alternative standard errors following a number of different specifications. To start with, we compute wild-t bootstrapped standard errors at the section level following Cameron and Miller (2014). In addition, we compute standard errors clustered at the school level (for our base specification and also when computing wild-t bootstrapped errors). Moreover, we compute standard errors aggregating outcomes (adjusted for baseline levels) at the section level and running a regression at this level including school-fixed effects (as suggested by Bertrand, Duflo and Mullainathan, 2004). Finally, we follow the methodology described in Ibragimov and Muller (2010), where the main model is estimated separately for each school, and then we perform a t-test on the distribution of estimated treatment coefficients. In all cases, the findings presented in our main analysis remain unaltered: we find statistically significant effects at the 1 percent level for math achievement and no effects for language achievement (results are presented in Table A.1).

A second methodological issue relates to potential spill-over effects that the implementation of the program in treatment sections may have generated on control sections in participating schools. That is, because in the main analysis we compare treatment sections with control sections in the same schools, it is possible that the introduction of the program may have affected the behavior of teachers and students in the control sections. In that case, the difference in academic achievement between students in treatment sections and those in control sections do not reflect the real causal effects of ConectaIdeas. Though spill-over effects within schools can play a role in certain interventions (e.g. in interventions that involve information provision), in this context this possibility may be attenuated. This is because, the implementation team controlled the ConectaIdeas platform and did not allow students in control sections to access it.

Still, we empirically explored this possibility by generating non-experimental estimates of the effects of the program on *control* sections. To that end, we used data from the national standardized exam and kept the control sections in the 24 schools participating in the experimental evaluation as well as a data for sections A and B in a sample of comparison schools. This set of comparison schools included those that satisfied these restrictions: located in Santiago, classified in the bottom two categories in terms of socioeconomic status by the Ministry of Education in 2017, that had two or three sections in 2017 and that participated in the national standardized exam in 2016 and 2017. To create a more balanced comparison group, we estimated the propensity score of being a school that participated in the experiment in 2017 using these covariates: student age, gender, kindergarten attendance and mother's education. Table A.2. presents the estimated spill-over effects on control sections when using the baseline specification and also when using propensity-score reweighting. In all cases, we do not find evidence that the implementation of the program in treatment sections generated effects on the academic achievement of students in control sections.

5.2. Effects measured by academic tests applied as part of the study

In addition to the effects estimated using our primary outcome measure (the national standardized examination), we also measured effects on math and language using tests developed and administered by a testing company.[10] These midline and endline tests, applied after 4 and 7 months of exposure, were applied as a backup in case we did not gain access to the data from the national standardized exams.

Table A.3 presents these results. Panel A reports that at midline the program generated effects of 0.18 standard deviations in math learning in our preferred specification. These results seem to be in line with the effects of 0.27 standard deviations documented on the standardized national exam considering that the midline study test was applied fourth months after the program started and national standardized exam was applied after 7 months of program

[10] This company is Centro de Medicion MIDE UC (https://www.mideuc.cl/).

exposure. In contrast, the results from the endline exam applied as part of the study show smaller effects of 0.13 standard deviations.

Though we cannot provide a definite explanation for the lower documented effects in the endline study test, the difference between the effects documented in both study tests may be related to how the tests were developed. For the midline test, the testing company surveyed teachers at the study schools and assessed students in the curriculum areas that had been covered during the first semester in these schools. In contrast, for the endline test, the testing company used a standard exam that it is routinely used to measure learning advances to schools interested in documenting how much their students are learning in each year. These schools tend to include more private, high-performing schools compared to the national student population in Chile. The topics covered in fourth grade in these schools can be quite different to those covered in the schools participating in the study (mainly public, low-performance schools). Hence, it may be the case that important skills that were emphasized in the intervention (and that were covered in the national standardized test) were not adequately covered in the endline exam applied by the testing company.[11]

5.3. Non-experimental estimates

In this section we further explore the robustness of the effectiveness of ConectaIdeas in improving learning in math. One potential concern is that our experimental evaluation could have influenced the quality of the implementation of the program. Therefore, it is important to gauge the effectiveness of the program under more normal circumstances. We do so by exploiting the program's initial implementation over the course of six years in other schools prior to the experimental evaluation in 2017. In particular, ConectaIdeas had been implemented across 11 schools in the district of Lo Prado in Santiago from 2011 until 2016.[12] We build on work in

[11] In line with this explanation, we document that while there is a strong overlap between the learning objectives that students practiced and that were assessed in the midline study test this was not the case for the endline study test. From the six top learning objectives in terms of students practice in the platform, the endline test did not assessed two of them and included only one question for other two learning objectives. In addition, the endline test included several items on the eight learning objectives that accounted the least practice in the platform. In contrast, these problems of lack of coverage are minimized with the national standardized exam that employs a rotated form application by which different students solve different subsets of questions (in total 175 items are included). A final piece of evidence that suggests that the results from the study endline exam may be less reliable compared to the national standardized exam is that the correlation between the study endline exam and the baseline exam was lower than the correlation between the national standardized exam and the baseline exam (0.59 versus 0.68). And a similar pattern is found when checking the correlations with the midline exams (0.66 and 0.76, respectively).

[12] Three of those schools did not receive the program in 2013 and 2014.

Araya (2018), who evaluates the effectiveness of ConectaIdeas using a before-after approach within program schools from 2009 to 2016.

Though the central elements of ConectaIdeas have remained unaltered over the years, there are some differences between the version implemented in 2011-2016 and the 2017 version evaluated experimentally. First, fourth graders participating in ConectaIdeas in 2011-2016 were expected to use the platform weekly for 135 minutes compared to the 180 minutes for the version evaluated experimentally.[13] Second, in the period 2011 to 2014, third grade students also were exposed to the program, having one 45-minute session each week. Finally, the platform underwent some minor modifications and fine-tuning over the years.

Our identification strategy in this section will rely on exploiting the temporal variation in program implementation at these 11 schools, using other similar schools as a counterfactual, following a difference-in-differences strategy. We obtained data of the fourth grade SIMCE evaluations spanning the years 2005 to 2016 at the student level.

To obtain a valid comparison group of schools we take a number of steps. First, we restrict the sample to urban schools in the Santiago metropolitan area that are of SES similar to those that received the program (in the three bottom categories). Additionally, we keep schools that consistently participate in the national standardized exam during the period and that are not too small in the pre-program period (fourth grade enrollment is always above 8 students in each year). Next, using school-level characteristics from the pre-program period, we estimate the propensity score of receiving the program. The score is a function the average combined national standardized exam score in math and language and the proportion of students that attended kindergarten. Finally, we can use the predicted propensity score to generate school-weights and use propensity score re-weighting in our estimations.[14]

Table 7 presents summary statistics for average school pre-program characteristics for our sample of treatment and comparison schools. Comparison schools are the subset of schools in Santiago that satisfy the sample restrictions mentioned above. At a first glance, the program and comparison schools look quite different (column 3). For instance, treatment schools underperform comparison schools in both math and language. Additionally, a lower share of

[13] Fourth graders participating in ConectaIdeas during 2011-2016 had about 45 minutes of additional math instruction per week whereas those participating in the experimental evaluation had about 90 additional minutes of math instruction.

[14] The weight for the control group is given by: $\frac{pscore}{1-pscore}$ while the weight for program schools equals 1.

students' mothers has secondary education, and a higher share of their mothers are indigenous. These differences are not only statistically significant, but also large in magnitude. In column (4) we restrict the sample to those schools for which there is overlap in the propensity scores (i.e., the propensity score lies between the minimum and maximum score in the treatment group). By applying this restriction, the differences between the program and comparison samples shrink considerably and, in most cases, become statistically insignificant. Finally, in column (5) we show differences between treatment and comparison schools after applying propensity score re-weighting.

We estimate the following model to exploit the non-experimental variation and assess the effect of ConectaIdeas on learning in fourth grade:

$$y_{ist} = \alpha + \beta Treatment_{st} + \tau_{\mathrm{t}} + \phi_s + X_{ist} + \varepsilon_{ist}$$

where $Treatment_{st}$ equals 1 in for school s that participated in the program in year t and 0 otherwise, τ_{t} are year fixed effects, and ϕ_s are school fixed effects, and X_{it} are student characteristics such as gender, dummy variable for attending kindergarten, family income, parental education, and class cohort size. Finally, β is the parameter of interest and estimates the average effect of participating in ConectaIdeas on math or language scores. Standard errors are clustered at the school level in all regressions.

Results are presented in Table 8. Columns 1 and 2 present the simple difference-in-differences estimates using the entire sample without propensity score re-weighting, while columns 3 and 4 restricts the sample to the common support and employs propensity score re-weighting. We find that regardless of the specification used, the estimate of participating in ConectaIdeas on math test scores is highly statistically significant (p-value<0.01) and economically significant. Students that received ConectaIdeas experience gains of around 0.19 to 0.22 standard deviations in math scores. At the same time, the program did not seem to affect language test scores. These results are similar to those obtained in our experimental design though slightly smaller (though this may be related to the lower time intensity that the program had during the 2011-2016 period compared to the version evaluated in 2017). These results provide supporting evidence regarding the robustness of the experimental effects of ConectaIdeas presented in section 4.

6. Conclusion

We conducted a randomized controlled trial among 24 primary, low-performing schools in Santiago, Chile to evaluate the effectiveness of ConectaIdeas–a math program that incorporates several gamification features to spur student motivation. We find that the program was effective in increasing learning in math by around 0.27 standard deviations as measured in the Chilean national standardized exam. We do not find any significant spillovers to language test scores. The program also affected other non-academic outcomes. On the positive side, the program increased students' preference towards using computers in math instruction and promoted the adoption of a growth mindset. On the negative side, the program generated increases in math anxiety and reduced students' preference towards teamwork.

It is important to consider some characteristics of the program when thinking about extrapolating the results to other contexts or scaling it up. The program targeted 24 Chilean schools with students from poorer backgrounds and low average performance, where the margin for improvement in math learning might be higher than in schools with higher levels of achievement. Also, although the program relies mainly on existing resources from the school (teachers follow the same curriculum and require minimal training, makes use of existing computer labs, conducted during school hours), it does require some basic infrastructure such as a reliable internet connection that might not be available in rural settings, and a lab coordinator provided by the program. Still, the substantial positive academic effects documented suggests that programs that adopt gamification features may be a promising strategy to increase student achievement.

References

Arias, E., and Cristia, J. 2014. “The IDB and Technology in Education: How to Promote Effective Programs?” IDB Technical Note 670. Inter-American Development Bank, Washington D.C.

Araya, R. 2018. “Teacher Training, Mentoring or Performance Support Systems?” *International Conference on Applied Human Factors and Ergonomics*: 306-315. Springer, Cham.

Banerjee, A., Cole, S., Duflo, E., and Linden, L. 2007. “Remedying Education: Evidence From Two Randomized Experiments in India.” *The Quarterly Journal of Economics* 122 (3):1235-1264.

Barrera-Osorio, F., and Linden, L. 2009. “The Use and Misuse of Computers in Education: Evidence from a Randomized Controlled Trial of a Language Arts Program.” C*ambridge, MA: Abdul Latif Jameel Poverty Action Lab (JPAL). www. leighlinden. com/Barrera-Linden,* 20.

Barrow, L., Markman, L., and Rouse, C. 2009. “Technology's Edge: The Educational Benefits of Computer-Aided Instruction.” A*merican Economic Journal: Economic Policy* 1(1): 52-74.

Bassi, Marina, Costas Meghir, and Ana Reynoso. "Education Quality and Teaching Practices." National Bureau of Economic Research Working Paper no. 22719 (2016).

Bellei, Cristián. "Does lengthening the school day increase students’ academic achievement? Results from a natural experiment in Chile." *Economics of Education Review* 28, no. 5 (2009): 629-640.

Beuermann, D., Cristia, J., Cueto, S., Malamud, O., and Cruz-Aguayo, Y. 2015. “One Laptop per Child at Home: Short-Term Impacts from a Randomized Experiment in Peru.” *American Economic Journal: Applied Economics* 7(2): 53-80.

Bos, M., Ganimian, A., and Vegas, E. 2013. “América Latina en PISA 2012: ¿Cómo le Fue a la Región?” Inter-American Development Bank, Washington DC.

Cameron, A., Gelbach, J., and Miller, D. 2008. “Bootstrap-Based Improvements for Inference with Clustered Errors.” *Review of Economics and Statistics* 90(3): 414-427.

Cameron, A. C., & Miller, D. L. (2015). A practitioner’s guide to cluster-robust inference. Journal of Human Resources, 50(2), 317-372.

Cheung, A., and Slavin, R. 2016. "How Methodological Features Affect Effect Sizes in Education." *Educational Researcher* 45(5): 283-292.

Cristia, J., Ibarrarán, P., Cueto, S., Santiago, A. and Severín, E. 2017. "Technology and Child Development: Evidence from the One Laptop per Child Program." *American Economic Journal: Applied Economics* 9(3): 295-320.

Dynarski, M., et al. 2007. "Effectiveness of Reading and Mathematics Software Products: Findings from the First Student Cohort." DIANE Publishing.

Fairlie, R., and Robinson, J. 2013. "Experimental Evidence on the Effects of Home Computers on Academic Achievement among Schoolchildren." *American Economic Journal: Applied Economics* 5(3): 211-40.

Gneezy, U., Niederle, M., and Rustichini, A. 2003. "Performance in Competitive Environments: Gender Differences." *The Quarterly Journal of Economics* 118(3): 1049-1074.

Hahn, Y., Islam, A., Patacchini, E., and Zenou, Y. 2017. "Do Friendship Networks Improve Female Education?" IZA Discussion Paper no. 10674.

Hambleton, R., and Traub, R. 1974. "The Effects of Item Order on Test Performance and Stress." *The Journal of Experimental Education* 43(1): 40-46.

Hill, C., Bloom, H., Black, A., and Lipsey, M. 2008. "Empirical Benchmarks for Interpreting Effect Sizes in Research." *Child Development Perspectives* 2(3): 172-177.

Lai, F et al. 2013. "Computer Assisted Learning as Extracurricular Tutor? Evidence from a Randomised Experiment in Rural Boarding Schools in Shaanxi." *Journal of Development Effectiveness* 5(2): 208-231.

Lai, F., Luo, R., Zhang, L., Huang, X., and Rozelle, S. 2015. "Does Computer-Assisted Learning Improve Learning Outcomes? Evidence From a Randomized Experiment in Migrant Schools in Beijing." *Economics of Education Review* 47: 34-48.

Linden, L. 2008. "Complement or Substitute? The Effect of Technology on Student Achievement in India". Unpublished manuscript. Columbia University.

Malamud, O., Cueto, S., Cristia, J., and Beuermann, D. 2018. "Do Children Benefit from Internet Access? Experimental Evidence from a Developing Country." Unpublished manuscript. Northwestern University.

McEwan, P. 2015. "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments." *Review of Educational Research* 85(3): 353-394.

McKenzie, D. 2012. "Beyond Baseline and Follow-Up: The Case for More T in Experiments." *Journal of Development Economics* 99(2): 210-221.

Mo, D., Swinnen, J., Zhang, L., Yi, H., Qu, Q., Boswell, M., and Rozelle, S. 2012. "Can One Laptop per Child Reduce the Digital Divide and Educational Gap? Evidence from a Randomized Experiment in Migrant Schools Ii Beijing." *World Development* 46: 14-29.

Mo, D., et al. 2014. "Integrating Computer-Assisted Learning into a Regular Curriculum: Evidence from a Randomised Experiment in Rural Schools in Shaanxi." *Journal of development effectiveness*, 6(3): 300-323.

Muralidharan, K., Singh, A., and Ganimian, A. Forthcoming. "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India." *American Economic Review*.

Nicol, D., and Macfarlane-Dick, D. 2006. "Formative Assessment and Self-Regulated Learning: A Model and Seven Principles of Good Feedback Practice." *Studies in Higher Education* 31(2): 199-218.

Ñopo, H. 2012. *New Century, Old Disparities: Gender and Ethnic Earnings Gaps in Latin America and The Caribbean.* The World Bank.

OECD. 2013. "PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy." PISA, OECD Publishing, Paris.

Rivas, A. 2015. *América Latina después de PISA: Lecciones Aprendidas de la Educación en Siete Países (2000-2015).* Fundación Cippec.

Weiner, B. (1990). History of motivational research in education. Journal of educational Psychology, 82(4), 616.

Wijekumar, K., Hitchcock, J., Turner, H., Lei, P., and Peck, K. 2009. "A Multisite Cluster Randomized Trial of the Effects of CompassLearning Odyssey [R] Math on the Math Achievement of Selected Grade 4 Students in the Mid-Atlantic Region." Final Report. NCEE 2009-4068. *National Center for Education Evaluation and Regional Assistance.*

Table 1: Sample Construction - Pre-Treatment Year (2016)

		Additional Sample Restrictions			
	All Schools	In Santiago	Low SES	Two or more Classrooms	Participated in Evaluation
	(1)	(2)	(3)	(4)	(5)
Test Scores (Normalized with Whole Country)					
Math	0.00	0.07	-0.37	-0.25	-0.68
Language	0.00	0.02	-0.38	-0.32	-0.60
Student Characteristics					
Female	0.50	0.50	0.48	0.50	0.48
Age	9.61	9.63	9.70	9.68	9.83
Attended Kindergarten	0.98	0.99	0.99	0.99	0.98
Mother with secondary education	0.72	0.76	0.52	0.55	0.48
Father at home	0.60	0.61	0.54	0.55	0.50
Indigenous mother	0.11	0.07	0.11	0.12	0.11
Number of students	*217,034*	*84,972*	*27,048*	*14,675*	*1,366*
School Characteristics					
Enrollment in 4th grade	29.35	47.90	37.41	67.94	56.92
Rural	0.39	0.07	0.14	0.06	0.04
Low SES	0.64	0.41	1.00	1.00	0.96
Number of schools	*7,395*	*1,774*	*723*	*216*	*24*

Notes: This table presents means for different groups of schools. Data from the 2016 fourth-grade national standardized exam are used. All test scores have been normalized subtracting the mean and dividing by the standard deviation of the sample that includes all students in the country. SES stands for socio-economic status. Column (1) presents means for students in all schools in the country, column (2) restricts the sample to those in the Santiago metropolitan area, column (3) further restricts the sample to schools in the two bottom categories (out of five) in terms of SES, column (4) further restricts the sample to schools with two or more classrooms, and column (5) further restricts the sample to schools participating in the study.

Table 2: Balance in Baseline Test Scores and Student Characteristics - Treatment Year (2017)

	Treatment (1)	Control (2)	Difference (3)	*N* (4)
Baseline Test Scores (Normalized with Control Group)				
Math	-0.09	0.00	-0.08	*1,089*
			(0.05)*	
Language	-0.05	0.00	-0.04	*1,057*
			(0.07)	
Student Characteristics				
Female	0.48	0.47	0.02	*1,089*
			(0.02)	
Age	9.74	9.76	-0.02	*1055*
			(0.03)	
Attended Kindergarten	0.98	0.99	-0.01	*788*
			(0.01)	
Mother with secondary education	0.47	0.53	-0.06	*837*
			(0.02)**	
Father at home	0.53	0.54	-0.01	*873*
			(0.02)	
Indigenous mother	0.16	0.14	0.03	*737*
			(0.02)	
Has internet	0.81	0.82	-0.02	*840*
			(0.02)	

Notes: This table presents means and estimated differences between the treatment and control groups. Results on baseline test scores are constructed using data from the baseline exam implemented as part of the study. Results on student characteristics are constructed using data from the 2017 fourth-grade national standardized exam. The sample used to analyze baseline math test scores and student characteristics includes students that participated in the baseline math exam and in the 2017 fourth-grade national standardized math exam. The sample used to analyze baseline language test scores includes students that participated in the baseline language exam and in the 2017 fourth-grade national standardized language exam. All test scores have been normalized subtracting the mean and dividing by the standard deviation of the control group. Columns (1) and (2) present means for treatment and control groups, respectively. Column (3) presents differences between the treatment and control groups controlling for school fixed effects. Column (4) presents the number of students in each sample. Standard errors, reported in parentheses, are clustered at the section level. Significance at the one, five, and ten percent levels is indicated by ***, **, and *, respectively.

Table 3: Effects on Academic Achievement

	Treatment	Control	Difference	Adjusted Difference	*N*
	(1)	(2)	(3)	(4)	(5)
Math	-0.39	-0.61	0.22	0.27	*1,089*
			(0.05)***	(0.04)***	
Language	-0.61	-0.59	-0.04	-0.01	*1,057*
			(0.05)	(0.04)	

Notes: This table presents estimated effects of Conecta Ideas on test scores in Math and Language. Data from the 2017 fourth-grade national standardized exam are used. Labels in rows correspond to dependent variables. Column (1) and (2) present means for treatment and control groups, respectively. Column (3) presents differences between the treatment and control groups controlling for school fixed effects. Column (4) presents adjusted differences controlling for school fixed effects and baseline value of the outcome. Column (5) presents the number of students in each sample. The sample used to analyze math test scores includes students that participated in the baseline math test and in the 2017 fourth-grade national standardized math exam. The sample used to analyze baseline language test scores includes students that participated in the baseline language test and in the 2017 fourth-grade national standardized language exam. All test scores have been normalized subtracting the mean and dividing by the standard deviation of the sample that includes all students in the country. Standard errors, reported in parentheses, are clustered at the section level. Significance at the one, five, and ten percent levels is indicated by ***, **, and *, respectively.

Table 4: Heterogeneous Effects on Academic Achievement

	Gender		Mother with Secondary Education		Baseline Score	
	Boys	Girls	Yes	No	Low	High
	(1)	(2)	(3)	(4)	(5)	(6)
Math	0.29	0.24	0.28	0.29	0.26	0.26
	(0.05)***	(0.06)***	(0.07)***	(0.05)***	(0.06)***	(0.05)***
N	*571*	*518*	*434*	*439*	*510*	*579*
Language	0.03	-0.05	0.05	-0.03	-0.03	0.05
	(0.05)	(0.05)	(0.06)	(0.05)	(0.06)	(0.05)
N	*565*	*492*	*420*	*427*	*530*	*527*

Notes: This table presents estimated effects of Conecta Ideas on test scores in Math and Language for differente sub-samples of students. Data from the 2017 fourth-grade national standardized exam are used. Each cell corresponds to one regression. Labels in rows correspond to dependent variables. The column titles indicate the sample included in the estimation. The sample used to analyze math test scores includes students that participated in the baseline math test and in the 2017 fourth-grade national standardized math exam. The sample used to analyze baseline language test scores includes students that participated in the baseline language test and in the 2017 fourth-grade national standardized language exam. All test scores have been normalized subtracting the mean and dividing by the standard deviation of the sample that includes all students in the country. Standard errors, reported in parentheses, are clustered at the section level. Significance at the one, five, and ten percent levels is indicated by ***, **, and *, respectively.

Table 5: Effects on Students' Perceptions

	Treatment	Control	Raw Difference	*N*
	(1)	(2)	(3)	(4)
Math Self Concept	0.06	0.00	0.10 (0.07)	*706*
Math Intrinsic Motivation	0.09	0.00	0.10 (0.08)	*797*
Math Anxiety	0.15	0.00	0.13 (0.05)**	*883*
Prefers Math Lessons in Lab	0.42	0.00	0.40 (0.06)***	*787*
Growth Mindset	0.06	0.00	0.10 (0.05)*	*790*
Teamwork	-0.20	0.00	-0.21 (0.06)***	*827*

Notes: This table presents estimated effects of Conecta Ideas on indices representing students' perceptions. Labels in rows correspond to dependent variables. Column (1) and (2) present means for treatment and control groups, respectively. Column (3) presents differences between the treatment and control groups controlling for school fixed effects. Standard errors, reported in parentheses, are clustered at the section level. Significance at the one, five, and ten percent levels is indicated by ***, **, and *, respectively.

Table 6: Non-Experimental Balance during pre-Treatment Period (2005-2011)

			Difference		
	Treatment	Comparison	No adjustments	With Common Support	With Common Support and Reweighting
	(1)	(2)	(3)	(4)	(5)
Test Scores (Normalized with Whole Country)					
Math	-0.47	-0.20	-0.27	-0.04	0.04
			(0.04)***	(0.04)	(0.05)
Language	-0.47	-0.21	-0.27	-0.04	0.04
			(0.04)***	(0.04)	(0.04)
Student Characteristics					
Female	0.49	0.48	0.01	0.02	0.03
			(0.01)	(0.01)	(0.02)*
Age	10.60	10.81	-0.21	-0.43	-0.48
			(0.20)	(0.20)**	(0.21)**
Attended Kindergarten	0.81	0.79	0.02	0.01	0.01
			(0.02)	(0.02)	(0.02)
Mother with secondary education	0.36	0.52	-0.16	-0.01	0.03
			(0.02)***	(0.02)	(0.03)
Father at home	0.22	0.22	-0.00	0.00	0.01
			(0.01)	(0.01)	(0.01)
Indigenous mother	0.15	0.08	0.07	0.03	0.00
			(0.02)***	(0.02)**	(0.02)
Has internet	0.69	0.66	0.03	0.04	0.01
			(0.06)	(0.06)	(0.06)
Number of schools	*11*	*999*	*1,010*	*429*	*429*

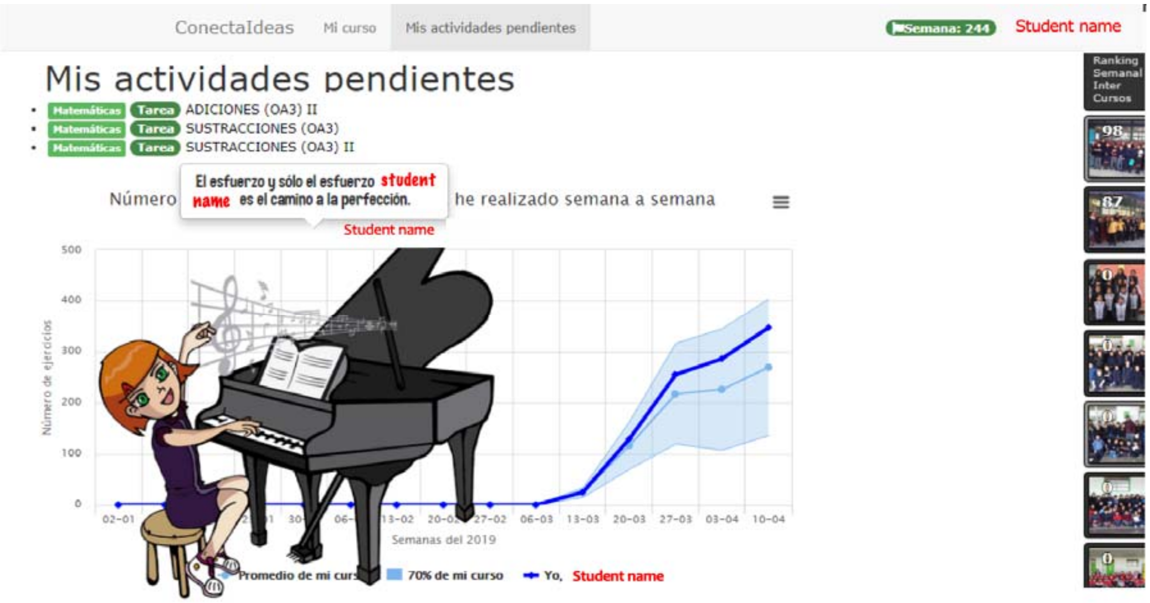
Notes: This table presents means and estimated differences between the treatment and comparison groups used for the non-experimental analysis. Data from the fourth-grade national standardized exam for 2005 to 2010 are used. All test scores have been normalized subtracting the mean and dividing by the standard deviation of the sample that includes all students in the country, for each year. Columns (1) and (2) present means for the treatment and comparison schools, respectively. Column (3) to (5) present differences between the treatment and comparison groups. Standard errors, reported in parentheses, are clustered at the school level. Significance at the one, five and ten percent levels is indicated by ***, **, and *, respectively.

Table 7: Non-Experimental Estimates - Effects on Academic Achievement

	Differences-in-Differences (DID)		DID + Propensity Score Reweighting	
	(1)	(2)	(3)	(4)
Math	0.22***	0.22***	0.19***	0.19***
	(0.06)	(0.06)	(0.07)	(0.07)
Language	0.06	0.07	0.02	0.03
	(0.04)	(0.04)	(0.04)	(0.05)
Number of students	*655,072*	*655,072*	*239,312*	*239,312*
Number of schools	*1,010*	*1,010*	*429*	*429*

Notes: This table present non-experimental difference-in-difference estimates on test scores in Math and Language. Data from the fourth-grade national standardized exam for 2005 to 2016 are used. The unit of observation is a school-year. Each cell corresponds to one regression. Each regression includes a treatment dummy, school fixed-effects, and year fixed-effects. Labels in rows correspond to dependent variables. Columns (1) and (2) include urban schools in the Santiago metropolitan area that are in the bottom three categories (out of five) in terms of SES and that had a minimum enrollment in fourth grade of 8 students in the 2005-2010 period. Columns (3) and (4) further restrict the sample to schools for which there is overlap in the propensity scores. Regression results presented in columns (2) and (4) also include time-varying controls. All test scores have been normalized subtracting the mean and dividing by the standard deviation for all students in the country, for each year. The number of schools and students presented in the table corresponds to those included to estimate effects on math test scores. 654,365 students in 1,010 schools are included to estimate effects on language test scores presented in columns (1) and (2). 239,182 students in 429 schools are included to estimate effects on language test scores presented in columns (3) and (4). Standard errors, reported in parentheses, are clustered at the school level. Significance at the one, five, and ten percent levels is indicated by ***, **, and *, respectively.

Figure 1: Screenshot of student dashboard

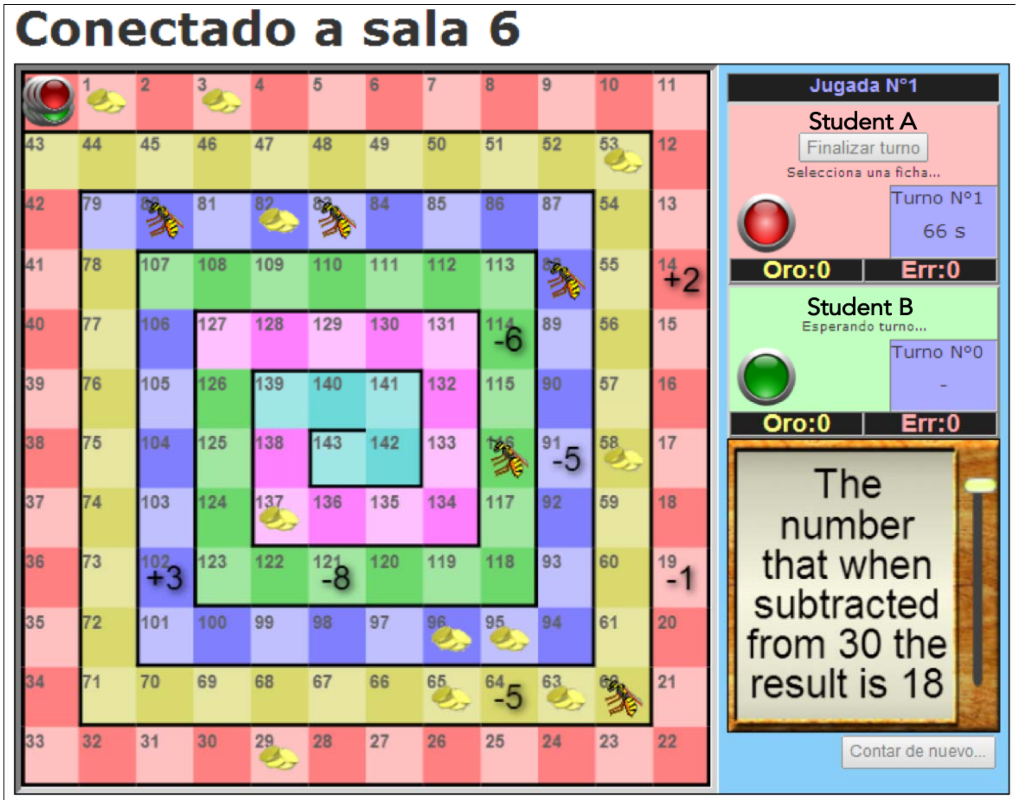


Mis actividades pendientes

- Matemáticas Tarea ADICIONES (OA3) II
- Matemáticas Tarea SUSTRACCIONES (OA3)
- Matemáticas Tarea SUSTRACCIONES (OA3) II

El esfuerzo y sólo el esfuerzo student name es el camino a la perfección.

Figure 2: Screenshot of tournament game



Conectado a sala 6

Jugada N°1

Student A

Selecciona una ficha...

Turno N°1

66 s

Oro:0 Err:0

Student B

Esperando turno...

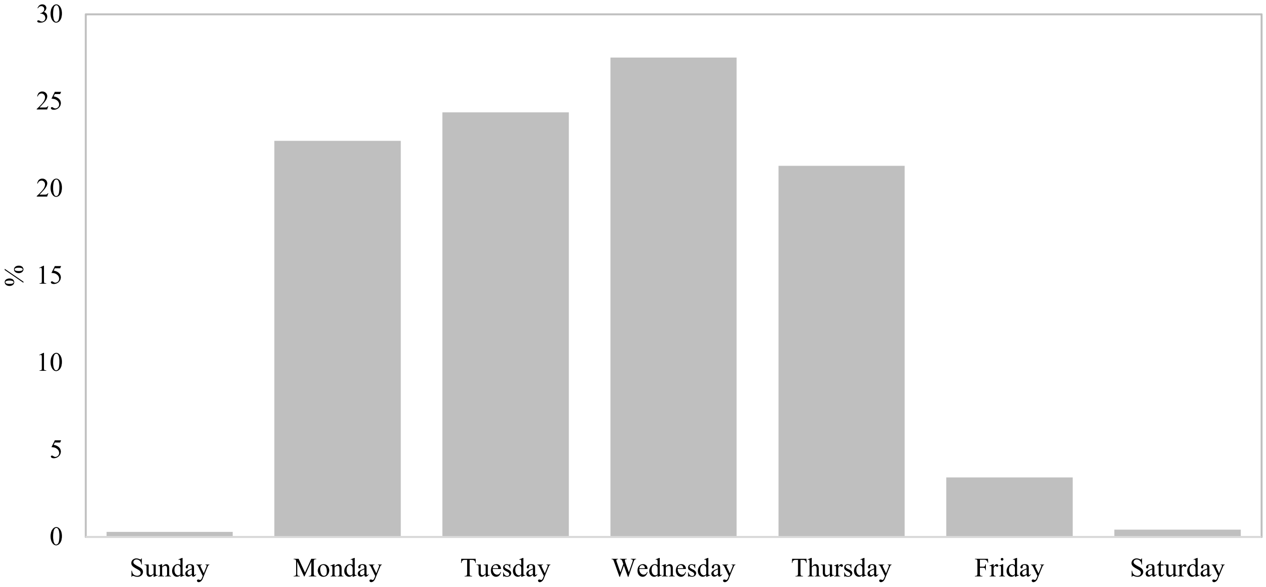
Turno N°0

-

Oro:0 Err:0

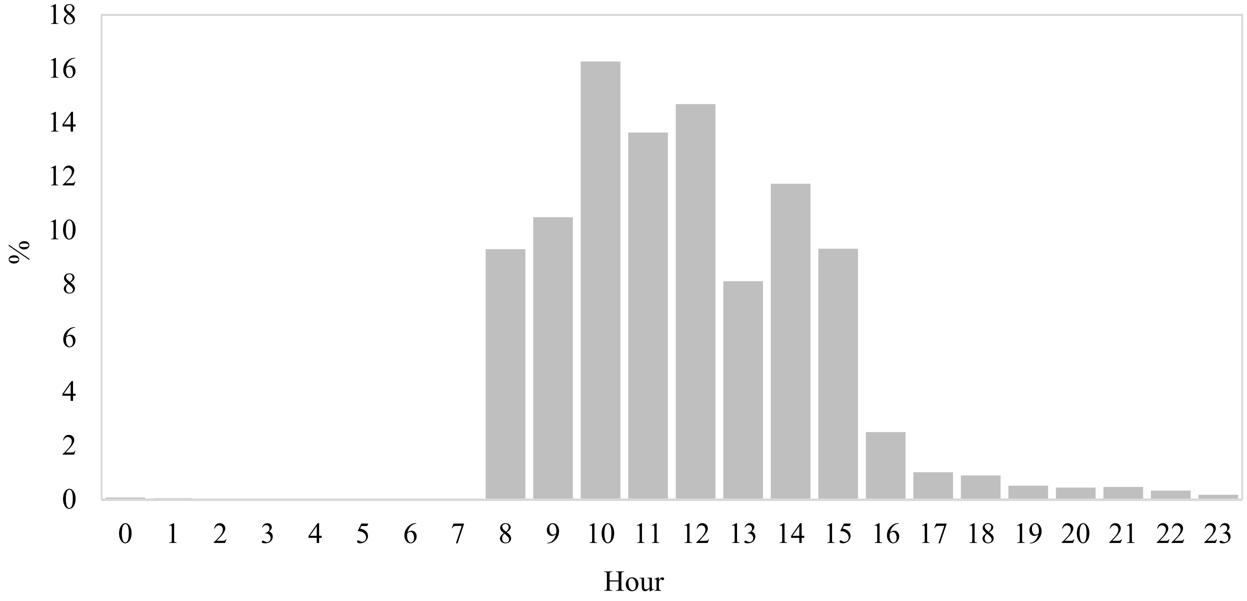
The number that when subtracted from 30 the result is 18

Figure 3: Platform Use by Day of the Week



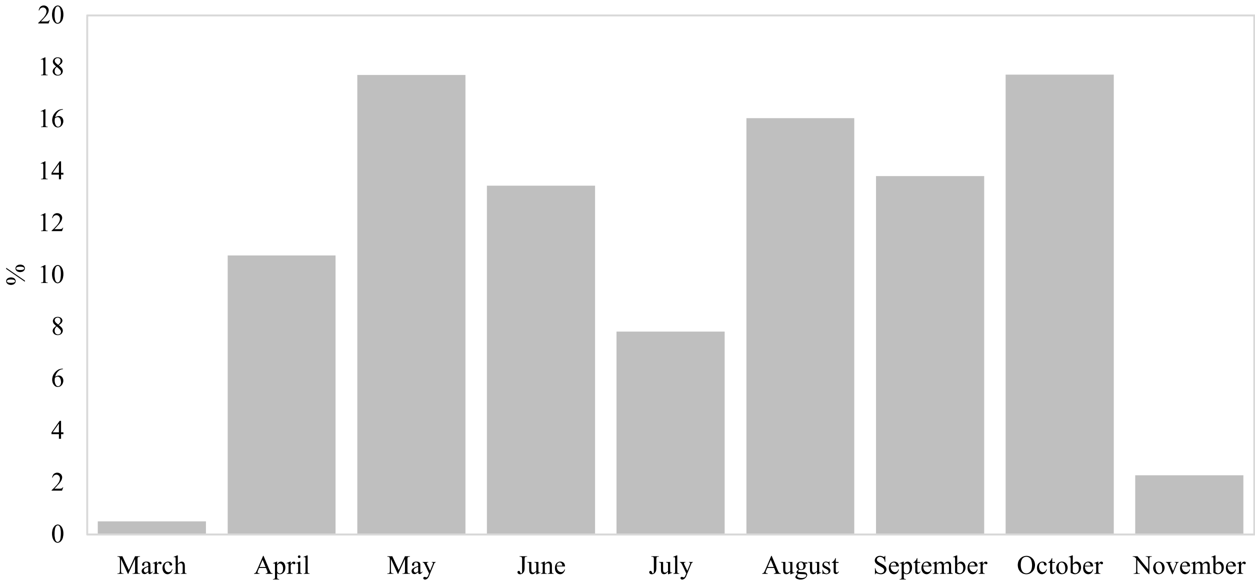
Notes: This figure presents the distribution of platform use by day of the week. Statistics correspond to the period from March 28, 2017 to November 6, 2017.

Figure 4: Distribution of Platform Use by Hour of the Day



Notes: This figure presents the distribution of platform use by hour of the day. Statistics correspond to the period from March 28, 2017 to November 6, 2017.

Figure 5: Platform Use by Month



Notes: This figure presents the distribution of platform use by month. Statistics correspond to the period from March 28, 2017 to November 6, 2017.

Table A.1: Robustness to Alternative Standard Errors

	Section-level		School-level			
	Standard Cluster	Wild Bootstrap	Standard Cluster	Wild Bootstrap	Bertrand et al. (2004)	Ibragimov and Muller (2010)
	(1)	(2)	(3)	(4)	(5)	(6)
Math	0.27	0.27	0.27	0.27	0.28	0.27
	(0.04)***	(0.06)***	(0.06)***	(0.06)***	(0.06)***	(0.07)***
Language	-0.01	-0.01	-0.01	-0.01	0.00	0.03
	(0.04)	(0.06)	(0.06)	(0.06)	(0.05)	(0.06)

Notes: This table presents estimated effects of Conecta Ideas on test scores in Math and Language. Data from the 2017 fourth-grade national standardized exam are used. Labels in rows correspond to dependent variables. Columns (1) through (4) use our main specification (adjusted differences) controlling for school fixed effects and baseline value of outcome. Columns (1) and (3) use conventional clustering at the classroom and school levels. Columns (2) and (4) use clustered wild-t bootstrap (Cameron and Miller, 2014) at the classroom and school levels. Column (5) employs the strategy proposed by Bertrand, Duflo and Mullainathan (2004), where outcomes (adjusted for baseline levels) are aggregated at the classroom level, and then our main specification is estimated using the aggregated data. Finally, column (6) follows Ibragimov and Muller (2010), where the main model is estimated separately for each school, and then we perform a t-test on the distribution of estimated treatment coefficients. The sample used to analyze math test scores includes students that participated in the baseline math test and in the 2017 fourth-grade national standardized math exam. The sample used to analyze baseline language test scores includes students that participated in the baseline language test and in the 2017 fourth-grade national standardized language exam. All test scores have been normalized subtracting the mean and dividing by the standard deviation of the sample that includes all students in the country. Significance at the one, five, and ten percent levels is indicated by ***, **, and *, respectively.

Table A.2: Exploring Spillover Effects on Control Sections

	(1)	(2)	(3)	(4)
Math	0.12	0.11	0.01	0.01
	(0.09)	(0.10)	(0.10)	(0.10)
Language	-0.01	-0.03	-0.06	-0.06
	(0.08)	(0.08)	(0.08)	(0.08)
Controls	N	Y	N	Y
Propensity score reweighting	N	N	Y	Y
Number of students	*23,040*	*22,895*	*17,883*	*17,786*
Number of schools	*218*	*218*	*178*	*178*

Notes: This table present difference-in-difference estimates on test scores in Math and Language. Data from the fourth-grade national standardized exam for 2016 and 2017 are used. Each cell corresponds to one regression. Each regression includes a treatment dummy, student characteristics (age, girl, kinder, mother completed secondary), school fixed-effects, and year fixed-effects. Labels in rows correspond to dependent variables. Columns (1) and (2) include urban schools in the Santiago metropolitan area that are in the bottom two categories (out of five) in terms of SES and that had 2 or 3 classrooms in 2016. Columns (3) and (4) further restrict the sample to schools for which there is overlap in the propensity scores estimated based on 2016 characteristics. Regression results presented in columns (2) and (4) also include time-varying controls. All test scores have been normalized subtracting the mean and dividing by the standard deviation for all students in the country, for each year. Standard errors, reported in parentheses, are clustered at the school level. Significance at the one, five, and ten percent levels is indicated by ***, **, and *, respectively.

Table A.3: Effects on Academic Achievement - Alternative Exams

	Treatment	Control	Difference	Adjusted Difference	*N*
	(1)	(2)	(3)	(4)	(5)
Panel A: Midline - Study Exams					
Math	0.06	0.00	0.11 (0.06)*	0.18 (0.05)***	*903*
Language	-0.07	0.00	-0.05 (0.07)	-0.03 (0.06)	*844*
Panel B: Endline - Study Exams					
Math	0.07	0.00	0.09 (0.06)	0.13 (0.05)***	*923*
Language	-0.02	0.00	-0.03 (0.06)	0.00 (0.05)	*882*

Notes: This table presents estimated effects of Conecta Ideas on test scores in Math and Language using data from exams implemented as part of the study. Panel A reports results generated from the midline study exam. Panel B reports results generated from the endline study exam. Labels in rows correspond to dependent variables. Column (1) and (2) present means for treatment and control groups, respectively. Column (3) presents differences controlling for school fixed effects. Column (4) presents adjusted differences controlling for school fixed effects and baseline value of the outcome. Column (5) presents the number of students in each sample. The sample used to analyze math test scores includes students that participated in the baseline math test and in the 2017 fourth-grade national standardized math exam. The sample used to analyze language test scores includes students that participated in the baseline language test and in the 2017 fourth-grade national standardized language exam. All scores have been standardized subtracting the mean and dividing them by the standard deviation of the control group. Standard errors, reported in parentheses, are clustered at the section level. Significance at the one, five, and ten percent levels is indicated by ***, **, and *, respectively.