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 the potential to improve learning by increasing student motivation but there is little rigorous evidence about its effectiveness. In this paper, we experimentally evaluate an innovative technology program that uses gamification to 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 math exercises. The platform tracks how many exercises students perform in the platform and features different types of individual and group competitions to promote student motivation. Results indicate large positive effects on math learning, of about 0.27 standard deviation, on the Chilean 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 hand, the program increased math anxiety and reduced students' preference towards teamwork. These results suggest that gamification could be an important tool to boost student learning, but that it may bring unintended consequences.

[1] This 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 Analía Jaimovich, Nora Cuevas (Mayor of San Bernardo), 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 Investigación Avanzada en Educación 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 acknowledge excellent comments and suggestions by Paulo Bastos, Samuel Berlinski, Marcelo Cabrol, Alejandrina Cristia, Colleen Eddy, Sebastián Espinoza, Alfonso Flores-Lagunes, Sebastián Gallegos, Ana María Ibañez, Mike Lovenheim, Julián Messina, Doug Miller, Francisco Pardo, Pamela Reyes-Santander, Amy Schwartz, Emiliana Vegas, Jouni Viiri, Ana Yarygina and seminar participants at the Inter-American Development Bank, the 2019 Pacific Conference for Development Economics, the 2019 annual conference of the Society for Research on Educational Effectiveness (SREE), the 2019 Research on Improving Educational Systems (RISE) conference, and the 2019 Daniel Patrick Moynihan Syracuse/Cornell Summer Workshop in Education and Social Policy. 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. The authors acknowledge funding from the International Development Research Center and the Inter-American Development Bank. Roberto Araya acknowledges PIA-CONICYT Basal Funds for Centers of Excellence Project FB0003 as well as inital funding from the Corporación de Fomento de la Producción (CORFO) and from the Fondo Nacional de Desarrollo Tecnológico y Productivo (FONTEC) that supported the development of the program. 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 central 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 (Barata et al., 2013; McDaniel et al., 2012; Hanus and Fox, 2015). 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 students every month.[2] Research on this topic has also increased markedly. According to Google Scholar, the number of papers published annually 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 measuring its effects on learning.

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 take place in regular school time. The program employs an array of gamification strategies to promote intensive use of the learning platform. First, the platform shows each student how many accumulated exercises she has completed and compares this figure with the

[2] Duolingo is an app for language acquisition. The app has 25 million users every month and its market valuation stood at 700 million dollars in 2017 (Buchanan, 2018). Khan Academy provides content and exercises in math, science, history, and other subjects. The website had 12 million of users every month in 2016 (https://khanacademyannualreport.org/).

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 promoting the notion that intelligence can be improved by exerting effort while studying. Third, whole class 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 math problems embedded in an online game.

Does ConectaIdeas 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, 0.7 standard deviations below the math national average. We randomly assigned one fourth-grade section within each of these schools 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 exam applied in November 2017 (after 7 seven months of program exposure). This is a paper-based assessment implemented yearly in all schools to monitor math and language learning. 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 in the US 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 specialized tests developed for specific interventions. Hence, using a broad national standardized exam 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 effects on math achievement are similar across subsamples 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, increased math and language achievement by 0.06 standard deviations (Bellei, 2009). In turn, a program that provided lesson plans and materials to teachers improved math and language test scores by 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 interventions.

To provide a comprehensive assessment of the program, we also examine effects on non-academic outcomes. On the positive side, we find 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-concept. 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 in the treatment group used the platform and that, on average, students used the platform for 27 hours. 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 constant during the seven-month period of program exposure. This finding contrasts with the sharp decrease in use over time documented in programs that provide laptops or internet for home use (Malamud et al., 2018). Therefore, the ConectaIdeas program was able to avoid the strong novelty effect 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, 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 exams. 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 documented are 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 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 (Denny 2013; Domínguez et al., 2013; Mekler et al., 2017). Reviews of this literature have generally concluded that incorporating gamification can increase student motivation and engagement (Lister, 2015; Alsawaier, 2018).[3] However, there is little rigorous empirical evidence on the effects of educational interventions that use gamification on academic achievement (Markopoulos et al., 2015).

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 individual incentives were provided to 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

[3] There has also been long-standing and recent research on how to incorporate games and tournaments in the classroom, without using technology, to boost motivation (Edwards et al. 1972; Slavin, 2010).

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 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 exam. Fourth, the study also reports program effects on intrinsic motivation, self-concept, 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 effects on academic achievement and section 5 reports evidence on potential mechanisms. Section 6 presents robustness checks including non-experimental estimates of program effects. Finally, section 7 concludes.

2. The ConectaIdeas Program

The ConectaIdeas program was developed by a team led by the researcher Roberto Araya, now 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 during 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 math instructional time.[4] In a typical session, all students worked solving the same set of 20 to 30 exercises assigned to them that were aligned to the topics covered in regular math instruction and included in the national curriculum. 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 at 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 program includes several gamification strategies. 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 comparisons with their 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 proven to be important in different domains such as energy conservation and worker effort (Cialdini et al., 2007; DellaVigna and Pope, 2018).[5]

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, that is, by having a "growth mindset" (Dweck, 2006). To that end, the platform shows students personalized "ads" emphasizing 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 a piano and a written message stating that

[4] In Chile, the large majority of students, including those participating in the study, attend school for about 7 hours. The regular schedule includes mandatory instructional time that schools have to assign to specific subjects but also some time that schools can allocate to any subject. Schools typically allocate these hours to math and language because these are the subjects assessed in the national standardized exam. The additional math session of the ConectaIdeas program used part of the time that schools can allocate to any subject.

[5] Another screen shows the ranking of individual students in the class who are ordered by the number of accumulated exercises completed in the year, the past week, or the current session.

“Effort, and only effort, *Student name*, is the road to perfection” (where *Student name* is replaced by the name of the student). These images are presented for 20 seconds and they are accompanied by computer-generated audio of the message. The images and messages presented are rotated from a library of 10 examples that emphasize the importance of effort while studying.

The third strategy focuses on group motivation rather than on individual motivation. In particular, competitions are set so that 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, photos of different sections of students are shown. This is a subset of 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. The photo shown in the fourth position (counting from the top) corresponds to 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 the average number of completed exercises and the bottom three sections are those that are immediately below in the ranking. Students can click any of these pictures to know the school name of the competing section.

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). Every five minutes the individual scores accumulated by each student are averaged at the class level and a program staff that acts as the “announcer” informs students which schools are doing better and tries to drive excitement among participants.

Besides these gamification features that seek to promote student engagement, the ConectaIdeas platform also includes some tools to facilitate and support the work of teachers. In particular, the platform provides teachers and lab coordinators a dashboard with real-time information about number of questions answered and number of correct answers by each student. In this dashboard, students are ranked from those that are in most need of support (those that have answered few questions or have a low rate of correct responses) to those that need less support. Additionally, the platform also presents a dashboard that shows the rate of correct responses per question to help to identify questions for which all students need support. Moreover, the system generates feedback reports for lab coordinators, teachers, and school principals.

Finally, we present information on the per-student costs associated to implementing ConectaIdeas. These costs include the salaries for a project coordinator, the lab coordinators as well as the costs of different inputs such as computers, internet, software, cloud computing, and general management. The per-student cost of implementing ConectaIdeas for the duration of this intervention stand at 150 dollars of 2017 and they represent a 5% increase in the public expenditure per primary student in Chile (see Appendix 1 for more details).

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 located in Santiago (to simplify logistics) and that had at least two fourth-grade sections. Moreover, the schools should have been classified by the Ministry of Education in the two lowest socioeconomic status categories (out of the five) to test whether ConectaIdeas could also close socioeconomic achievement gaps.

The recruitment process began at the end of January 2017 with the identification of 22 school districts (*comunas*) that had schools satisfying the criteria described above. The district directors were contacted first by email and then by phone calls. The ConectaIdeas team then visited 11 districts directors that replied and expressed interest. After that, information sessions with district directors and school principals were conducted in 9 school districts. In the final step, the team conducted school technical visits to verify technical requirements. The technical visits were scheduled in 31 schools in 6 school districts. However, after visiting 4 school districts (La Pintana, Maipu, Quinta Normal, and San Bernardo) the team confirmed 24 schools that met all the mentioned requirements and hence the recruitment process was stopped. Importantly, the recruitment procedure did not involve individual schools making decisions to self-select into the program.

Table 1 presents statistics obtained from the 2016 national standardized exam (known in Chile as "Sistema de Medición de la Calidad de la Educación" or *SIMCE*) that shows 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 Santiago 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.[6]

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 the collection of baseline data, and schools were informed of the treatment status of each section *after* the baseline was collected in March 2017. There was perfect compliance of program assignment to treatment. That is, all sections assigned to treatment participated in ConectaIdeas and none of the control sections participated in the program. Finally, attrition for the academic achievement outcomes is low and balanced between the treatment and control groups.[7]

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

[7] Attrition rates for estimating effects on math achievement were 10% for the treatment group and 8% for the control group. A regressions of attrition status on a treatment dummy (controlling for school fixed-effects) generates a produces a coefficient of 0.02 and a standard error of 0.01. For language attrition rates were 10% for the treatment group and 9% for the control group. The attrition regression produces a coefficient of 0.00 and a standard error of 0.01.

3.2. Identification Strategy

Evaluating the effect of the program is straightforward due to the random assignment of sections to treatment within schools. The advantage of this design is that it allows accounting 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:

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

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. The coefficient of interest, β, estimates the average treatment effect of the program on the outcome variable.

In a second specification, we also control 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 on language we control for the baseline test score in that subject. Because learning is strongly correlated over time, controlling for the baseline test scores typically increases statistical precision. Moreover, doing so can account for baseline differences in academic achievement. Consequently, this is our preferred specification.

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 estimate standard errors.

3.3 Data

Our analysis relies on a combination of administrative records from the Chilean national standardized exam, survey data, and administrative data from the ConectaIdeas platform.

The main outcome of the study corresponds to math test scores from the national standardized exam applied on 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 widely used for monitoring educational outcomes. 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. We also applied baseline, midline (after 4 months of exposure), and endline (after 7 months of exposure) math and language tests as part of the study.

We also elicit data from students at endline about their 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, 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 log data from the ConectaIdeas platform to document how computers were used for math instruction.

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 that were collected in March 2017. In turn, column (3) presents estimated differences between the treatment and control groups controlling for school fixed-effects and column (4) presents the sample size for each variable analyzed. Results indicate no statistically significant differences in language test scores across groups. However, treatment students underperformed control students 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 it provides additional motivation to control for baseline academic achievement when estimating effects on academic achievement.

Table 2 also 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 similar. The differences in the analyzed characteristics are small and only statistically significant at the five percent level for mothers' education.[8]

4. Main Results

Did ConectaIdeas affect student learning? To answer this question, Table 3 present program effects on math academic achievement measured in the 2017 national standardized exam. Results indicate that the ConectaIdeas program 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 induced shifting study time from language to math. However, results indicate that the program did not affect language achievement.

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. 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 learning 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

[8] 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.

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 Chile is similar to the US, we can state 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 the 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 positive effects of ConectaIdeas were experienced by different subpopulations of students defined by gender, mother's education, and baseline academic achievement.

5. Mechanisms

5.1. Evidence on Platform Use

The ConectaIdeas program generated large effects on math learning. But, why is the program effective? Is it because students are using the platform intensively? And, in that case, is this intensive use happening mainly at school or also at home? To answer these questions and better understand the mechanisms behind the document effects, 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.

Log data shows that the platform was intensively used: students used it on average for 27 hours. Moreover, 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. The use was heavily concentrated at school, which accounted for 98% of the total platform use.[9] These results are further corroborated by Figures 3 and 4 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 this intervention.

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.

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 5 presents the distribution of platform use by month. Results indicate that platform use was similar throughout the school year (once months with incomplete use are excluded).[10]

5.2. Effects on Non-Academic Outcomes

Introducing game elements to instruction can generate effects on a range of outcomes beyond math and language academic achievement. Consequently, we examine effects on non-academic outcomes using data from our endline student survey and from the questionnaire that was applied together with the national standardized exam.

In particular, we construct indices measuring relevant dimensions. For example, we construct an index for intrinsic motivation using nine items included in the endline student survey that were translated into Spanish from the scale used in the 2015 TIMSS math fourth-

[9] 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.

[10] 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 and in July because of the 2-week winter vacation.

grade examination (IEA, 2014). 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, averaged across items for the same construct and later standardized again for easier interpretation.[11] 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 statistically significant effects on two areas that are well aligned to prior expectations.[12] To start with, the basis of gamification involves 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 instead of in the regular classroom. In contrast, only 59% of students in the control group report preferring doing math sessions in the computer lab. This difference translates to a positive effect of 0.40 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 0.10 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. However, we do not find evidence supporting this expectation. In fact, the effect on intrinsic motivation is positive though not statistically significant. The second one is on math self-concept or the self-perception that students hold on their own abilities to solve math exercises. Because ConectaIdeas produced large increases in math achievement, we could expect positive effects on this area. Yet, we do not find statistically significant effects.

In turn, there are two areas in which we find statistically significant 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. We also document negative statistically significant effects on preferences for teamwork of 0.21 standard deviations. 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

[11] Table A.1 presents detailed information on the items used to construct each non-academic outcome.

[12] In this subsection all results that are noted as statistically significant, refer to the 5 percent level.

students may notice the disadvantages of working in teams (e.g. the weaker link between own performance and final outcomes) when participating repeatedly in the ConectaIdeas team competitions.

In this analysis, we checked effects on six different outcomes. Because we are running multiple hypothesis tests, the probability of finding some statistically significant results is heightened. To tackle this issue, we follow the procedure described by Benjamini et al. (2006) and implemented in Stata by Anderson (2008) to produce q-values, which can be interpreted as analogous to p-values once we account for multiple hypothesis testing.[13] The effects on math anxiety and on preferences towards doing math lessons in the computer lab and for teamwork continue being statistically significant after this adjustment. However, the effect on growth mindset loses statistical significance (the associated q-value is 0.16).

Finally, we explore the heterogeneity of ConectaIdeas effects on non-academic outcomes by estimating treatment effects for subsamples defined by gender, mother's education, and baseline academic achievement. Table 6 presents these results. The most consistent documented effects are the increased preference towards having math lessons in the computer lab, which are statistically significant for all subsamples. The positive effects on growth mindset and math anxiety as well as the negative effect on preferences towards teamwork, that were documented for the sample as a whole, are also observed for each subsample though only in some cases the effects are statistically significant. On the other hand, some positive statistically significant effects on math self-concept and intrinsic motivation are now present for some subsamples. However, caution should be exercised when interpreting these coefficients due to the reduction in power associated with focusing on subsamples and also because of the increased likelihood of finding statistically significant effects when running multiple hypothesis tests.

To provide a more definite assessment on this issue, we examine whether the estimated effects are statistically significant for each outcome and across each dimension. For example, we produce the p-value for the difference in the ConectaIdeas effects on math self-concept between boys and girls. In this analysis, we find only four cases in which the effects were statistically significantly different across subsamples. In particular, we find that the effects on teamwork are more negative for girls compared to boys, the effects on preferences towards math lessons in the

[13] Q-values are the lowest critical value at which a null hypothesis is rejected when controlling for the false discovery rate. The false discovery rate is the expected proportion that rejected null hypotheses are indeed true. To estimate q-values we need to specificy a family of related p-values. In this exercise, we consider that the 6 p-values presented in Table 5 belong to the same family.

computer lab and on growth mindset are more positive for students whose mothers have lower levels of education, and the effects on preferences towards math lessons in the computer lab are larger for students with lower baseline academic achievement. However, once we take into account that we are running multiple hypotheses tests for this analysis (by following the procedure described in Benjamini et al., 2006), we found no statistically significant differences in the effect of ConectaIdeas in any of the outcomes-dimensions considered.[14]

6. Robustness checks

6.1. Estimation of Standard Errors

For the main results on academic achievement, we estimate standard errors clustered at the section level. Because we have 48 clusters, it can be the case that the formulas used to compute standard errors may generate conservative estimates. To tackle this issue, we 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 (2015). 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 achievement, at the section level and running a regression at this level including school-fixed effects (as suggested by Bertrand et al., 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 (Table A.2).

6.2. Spillovers on the Control Group

It is possible that the introduction of the program may have affected the behavior of teachers and students in the control sections. For example, teachers in control sections may have exerted more effort to compensate for potential program effects or they could have become discouraged for not receiving additional support. In either case, the difference in academic achievement between

[14] In this exercise, we consider that all p-values of the differential effects across dimensions for the non-academic outcomes belong to the same family. Note that there are six non-academic outcomes for each of the three dimensions for a total of 18 p-values.

students in treatment sections and those in control sections do not reflect the causal effects of ConectaIdeas. Though spillover effects within schools can play an important 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 difference-in-difference 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 sections A and B in comparison schools. These comparison schools included those located in Santiago, classified in the bottom two socieconomic categories 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 better comparison group, we estimated the propensity score of being a school that participated in the experiment in 2017 using students' age, gender, kindergarten attendance, and mother's education. Table A.3. presents the estimated spillover effects on control sections. The results suggest that the program did not generate measurable effects on either students' math and language achievement in control sections. These results are present both for the baseline specification and also when using propensity-score reweighting. Consequently, these findings suggest that the difference in academic achievement between students in treatment sections and those in control sections reflect the causal effects of ConectaIdeas.

6.3. Effects Measured using Academic Tests Applied as Part of the Study

In addition to the effects estimated using our primary outcome measure (the national standardized exam), we also measured effects on math and language using tests developed and administered by testing companies contracted as part of the study. These midline and endline tests, were applied as a backup in case we did not gain access to the national standardized exams data.

Table A.4 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 4 months after the program started and the national standardized exam was applied after 7 months of program exposure. In contrast, the results from the endline exam applied as part of the study show smaller effects of 0.13 standard deviations in math achievement.

There are a number of potential reasons why we document lower effects on the endline study test compared with the midline study test and the national standardized exam that are related to how these tests were developed. For the study 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 study endline test, the testing company used a standard exam that typically applies in schools interested in documenting how much their students are learning. The schools that purchase this service 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 from those covered in the schools participating in this study (mainly public, low-performing schools). Hence, it may be the case that important skills that were emphasized in the intervention (and that were covered in the national standardized exam) were not adequately covered in the endline exam applied by the testing company.[15]

6.4. Non-Experimental Estimates

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. To do so, we provide additional non-experimental evidence regarding the effects of ConectaIdeas by using school-level longitudinal data from the fourth-grade national standardized exams in the period prior to the 2017 experimental

[15] In line with this explanation, there is a strong overlap between the learning objectives that students practiced in the platform and that were assessed in the midline study test but 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 for 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).

evaluation.[16] In particular, we exploit the implementation of ConectaIdeas in 11 schools in the district of Lo Prado, in Santiago, from 2011 until 2016 in a difference-in-difference framework.[17]

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.[18] 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.

We construct a comparison group by focusing on schools located in Santiago, classified in the bottom three socioeconomic categories, that participated consistently in the national standardized exam during the period, and that had fourth-grade enrollment above 8 students in all years. Next, using school-level characteristics from the pre-program period, we estimate the propensity score of receiving the program as a function of the average math and language scores and the proportion of students that attended kindergarten. Finally, we use the predicted propensity score to generate school-weights and use propensity score re-weighting in our estimations.[19]

Table 7 presents summary statistics for average school pre-program characteristics for treatment and comparison schools. At first glance, these two groups of schools look quite different (column 3). For instance, treatment schools underperform comparison schools in both math and language. In column (4) we restrict the sample to those schools for which there is an 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 treatment and comparison samples shrink considerably. Finally, in column (5) we show differences between treatment and comparison schools after applying propensity score re-weighting.

We estimate the following model to assess the effect of ConectaIdeas on achievement:

[16] We build on work in Araya (2018), who evaluates ConectaIdeas using a before and after approach.

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

[18] 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.

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

$$y_{ist} = \alpha + \beta Treatment_{st} + \tau_t + \phi_s + X_{ist} + \varepsilon_{ist}$$

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, a 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.

Columns (1) and (2) in Table 8 present difference-in-differences estimates using the entire sample, while columns (3) and (4) restricts the sample to the common support and employs propensity score re-weighting. Results indicate that ConectaIdeas improved math achievement in 0.19 to 0.22 standard deviations and that there were no effects on language scores. These results are similar to those obtained in our experimental design though slightly smaller. Overall, these results provide additional evidence regarding the effectiveness of ConectaIdeas to boost math learning.

7. 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 increased learning in math by 0.27 standard deviations as measured by the Chilean national standardized exam. We do not find any significant spillovers to language test scores. The program also affected non-academic outcomes. On the positive side, the program increased students' preference towards using computers in math instruction and promoted the idea among students that study effort can raise intelligence . 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 considering extrapolating the results to other contexts or scaling it up. The program targeted schools with students from poorer backgrounds and low average performance in Santiago, Chile. Consequently, effects might be different in other contexts where students are achieving higher levels of learning. Also, to implement all the features described in this paper it is necessary that schools count with a computer lab with reliable internet connection. This condition will not be met in many schools around the developing world, especially in those located in rural areas. However, it is possible to design and implement a version of the program that will not include group competitions, which will not require internet access. Finally, the program as evaluated here involved the implementation by external coordinators. To facilitate the scalability and reduce costs, it is relevant to test a version of the program that can be implemented by regular teachers. These teachers should received training and pedagogical support to facilitate the development of the necessary skills regarding how to use the platform for learning. Moreover, it could be fruitful to explore some potential complementary actions, such as providing a small payment to teachers, to compensate them for the extra effort required to adopt these new practices to their work

Beyond these important issues regarding extrapolation and scalability, the substantial positive academic effects documented here suggests that using gamification in education may be a promising strategy to increase student achievement. This is especially relevant considering that Chile, and many other countries in the developing and developed world, are seeking effective strategies to improve learning levels and reduce learning gaps. Moreover, using gamification approaches could take advantage the substantial investments that many countries have made to increase access to computers and internet in schools as well as rising access to internet-connected devices in households (Arias Ortiz and Cristia, 2014).

However, more research is needed to better understand how the different gamification strategies included in ConectaIdeas (and other innovative strategies) can affect student engagement and learning. Moreover, considering the effects documented here regarding increased anxiety and reduced preferences from teamwork, it is important to further explore the robustness of these results and potential ways to eliminate these unintended effects.

References

Alsawaier, R. 2018. “The Effect of Gamification on Motivation and Engagement.” *The International Journal of Information and Learning Technology* 35: 56–79.

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

Arias Ortiz, 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, DC.

Banerjee, A. et al. 2007. “Remedying Education: Evidence From Two Randomized Experiments in India.” *The Quarterly Journal of Economics* 122:1235–1264.

Barata, G. et al. 2013. “Engaging Engineering Students with Gamification.” *International Conference on Games and Virtual Worlds for Serious Applications* 1–8. IEEE.

Bassi, M., Meghir, C., and Reynoso, A. 2016. “Education Quality and Teaching Practices.” NBER Working Paper 22719. Cambridge, United States: National Bureau of Economic Research.

Bellei, C. 2009. “Does Lengthening the School Day Increase Students’ Academic Achievement? Results from a Natural Experiment in Chile.” *Economics of Education Review* 28: 629–640.

Bettinger, E. 2012. “Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores.” *Review of Economics and Statistics* 94: 686–698.

Bertrand, M., Duflo, E., and Mullainathan, S. 2004. “How Much Should We Trust Differences-in-Differences Estimates?” *The Quarterly Journal of Economics* 119: 249–275.

Buchanan, L. 2018. “The Hottest Education Startup in the U.S. Is a $700 Million Company Built by a Guatemalan Engineer in Pittsburgh.” *Inc.*, December 18, 2018. https://www.inc.com/leigh-buchanan/duolingo-700-million-language-learning-startup-pittsburgh-2018-surge-cities.html

Busso, M. et al. Eds. 2017. “Learning Better: Public Policy for Skills Development.” Inter-American Development Bank, Washington, DC.

Cameron, A., and Miller, D. 2015. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources* 50: 317–372.

Cialdini, R. et al. 2006. “Managing Social Norms for Persuasive Impact.” *Social Influence* 1: 3–15.

Claro, S., Paunesku, D., and Dweck, C. 2016. “Growth Mindset Tempers the Effects of Poverty on Academic Achievement.” *Proceedings of the National Academy of Sciences* 113: 8664–8668.

DellaVigna, S., and Pope, D. 2018. “Predicting Experimental Results: Who Knows What?” *Journal of Political Economy* 126: 2410–2456.

Denny, P. 2013. “The Effect of Virtual Achievements on Student Engagement.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 763–772. ACM.

Domínguez, A. et al. 2013. “Gamifying Learning Experiences: Practical Implications and Outcomes.” *Computers & Education* 63: 380–392.

Dweck, C. 2006. “Mindset: the New Psychology of Success.” Random House Incorporated.

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

Fryer, R. 2011. “Teacher Incentives and Student Achievement: Evidence from New York City Public Schools.” NBER Working Paper 16850. Cambridge, United States: National Bureau of Economic Research.

Hanus, M., and Fox, J. 2015. “Assessing the Effects of Gamification in the Classroom: A Longitudinal Study on Intrinsic Motivation, Social Comparison, Satisfaction, Effort, and Academic Performance.” *Computers & Education* 80: 152–161.

Hill, C. et al. 2008. “Empirical Benchmarks for Interpreting Effect Sizes in Research.” *Child Development Perspectives* 2: 172–177.

Ibragimov, R., and Müller, U. 2010. “t-Statistic Based Correlation and Heterogeneity Robust Inference.” *Journal of Business & Economic Statistics* 28: 453–468.

IEA. 2014. “Trends in International Mathematics and Science Study 2015.” Student Questionnaire, Grade 4.

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: 208–231.

Lai, F., et al. 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.

Li, T. et al. 2014. “Encouraging Classroom Peer Interactions: Evidence from Chinese Migrant Schools.” *Journal of Public Economics* 111: 29–45.

Lister, M. 2015. “Gamification: The Effect on Student Motivation and Performance at the Post-Secondary Level.” *Issues and Trends in Educational Technology* 3: 1–22.

Malamud, O. et al. 2019. “Do Children Benefit from Internet Access? Experimental Evidence from Peru.” *Journal of Development Economics* 138: 41–56.

Markopoulos, A. et al. 2015. “Gamification in Engineering Education and Professional Training.” *International Journal of Mechanical Engineering Education* 43: 118–131.

McDaniel, R., Lindgren, R., and Friskics, J. 2012. “Using Badges for Shaping Interactions in Online Learning Environments.” *International Professional Communication Conference* 1–4. IEEE.

Mekler, E. et al. 2017. “Towards Understanding the Effects of Individual Gamification Elements on Intrinsic Motivation and Performance.” *Computers in Human Behavior* 71: 525–534.

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: 300–323.

Muralidharan, K., Singh, A., and Ganimian, A. 2019. “Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India.” *American Economic Review* 109: 1426–1460.

OECD. 2017. “PISA 2015 Results (Volume V): Collaborative Problem Solving.” PISA, OECD Publishing, Paris.

Rutherford, T. et al. 2014. “A Randomized Trial of an Elementary School Mathematics Software Intervention: Spatial-Temporal Math.” *Journal of Research on Educational Effectiveness* 7: 358–383.

Weiner, B. 1990. “History of Motivational Research in Education.” *Journal of Educational Psychology* 82: 616–622.

Wijekumar, K. et al. 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 (0.05)*	*1,089*
Language	-0.05	0.00	-0.04 (0.07)	*1,057*
Student Characteristics				
Female	0.48	0.47	0.02 (0.02)	*1,089*
Age	9.74	9.76	-0.02 (0.03)	*1055*
Attended Kindergarten	0.98	0.99	-0.01 (0.01)	*788*
Mother with secondary education	0.47	0.53	-0.06 (0.02)**	*837*
Father at home	0.53	0.54	-0.01 (0.02)	*873*
Indigenous mother	0.16	0.14	0.03 (0.02)	*737*
Has internet	0.81	0.82	-0.02 (0.02)	*840*

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 Non-Academic Outcomes

	Treatment (1)	Control (2)	Raw Difference (3)	*N* (4)
Prefers Lab for Math	0.42	0.00	0.40 (0.06)***	*787*
Growth Mindset	0.06	0.00	0.10 (0.05)*	*790*
Math Intrinsic Motivation	0.09	0.00	0.10 (0.08)	*797*
Math Self-Concept	0.06	0.00	0.10 (0.07)	*706*
Math Anxiety	0.15	0.00	0.13 (0.05)**	*883*
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: Heterogeneous Effects on Non-Academic Outcomes

	Gender		Education		Baseline Score	
	Boys	Girls	Yes	No	Low	High
	(1)	(2)	(3)	(4)	(5)	(6)
Prefers Lab for Math	0.49	0.34	0.29	0.61	0.53	0.29
	(0.07)***	(0.07)***	(0.09)***	(0.09)***	(0.09)***	(0.08)***
Growth Mindset	0.19	0.05	0.01	0.27	0.06	0.17
	(0.05)***	(0.08)	(0.06)	(0.09)***	(0.07)	(0.07)**
Math Intrinsic Motivation	0.06	0.20	0.15	0.12	-0.01	0.20
	(0.09)	(0.12)*	(0.12)	(0.11)	(0.13)	(0.10)**
Math Self-Concept	0.04	0.25	0.18	0.18	0.08	0.22
	(0.08)	(0.13)*	(0.10)*	(0.08)**	(0.12)	(0.07)***
Math Anxiety	0.08	0.18	0.08	0.11	0.10	0.15
	(0.06)	(0.07)**	(0.09)	(0.07)	(0.07)	(0.08)*
Teamwork	-0.10	-0.33	-0.25	-0.08	-0.10	-0.31
	(0.08)	(0.07)***	(0.10)**	(0.07)	(0.07)	(0.08)***

Notes: This table presents estimated effects of Conecta Ideas on indices representing students' perceptions for differente subsamples 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. 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 7: 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*

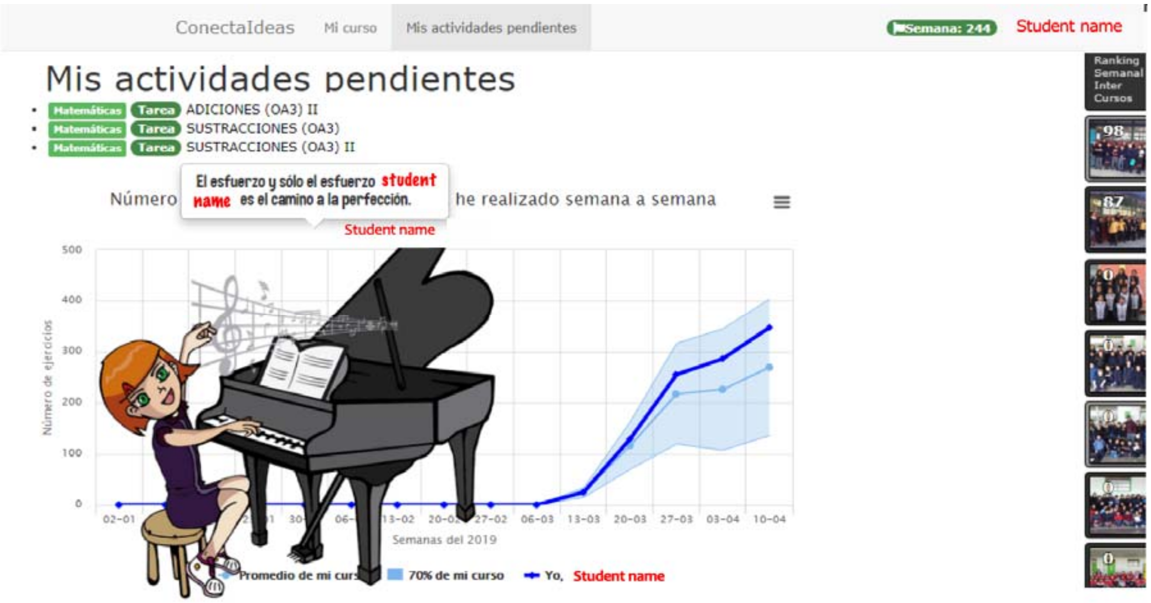
Notes: This table presents means and estimated differences between the treatment and comparison groups used for the non-experimental analysis. Data from the fourth-grade national standardized exam for 2005 to 2010 are used. All test scores have been normalized subtracting the mean and dividing by the 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 8: 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 presents 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

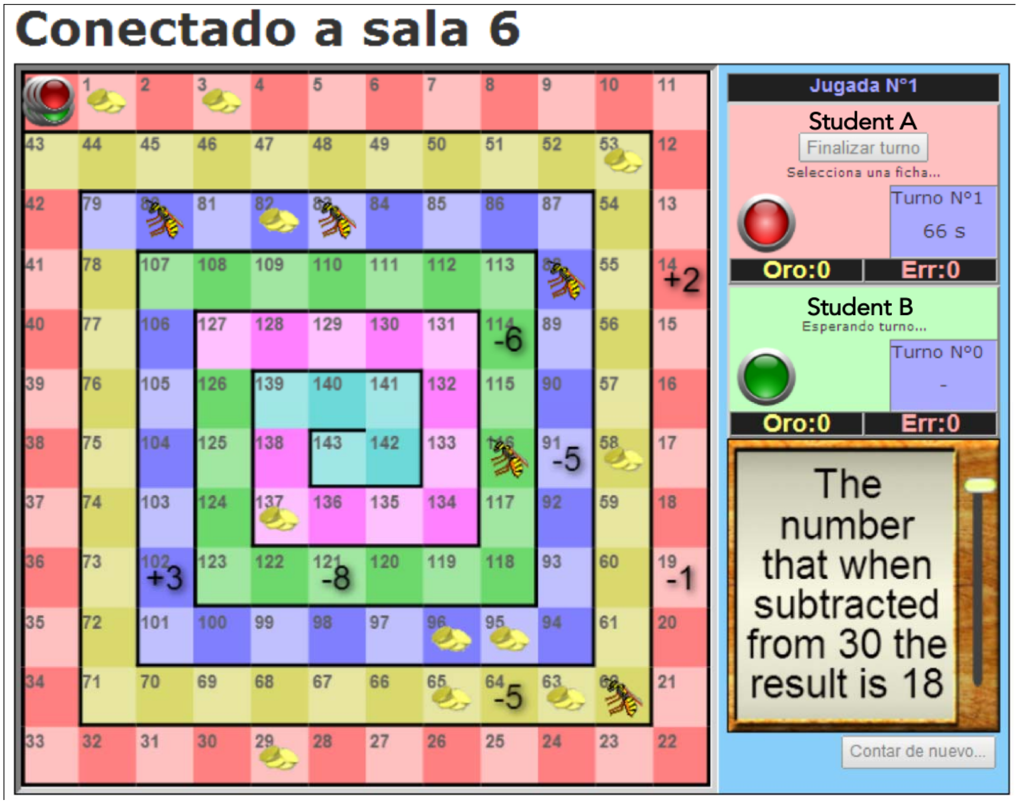


Mis actividades pendientes

- 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 the Tournament Game



Conectado a sala 6

	1	2	3	4	5	6	7	8	9	10	11
43	44	45	46	47	48	49	50	51	52	53	12
42	79	80	81	82	83	84	85	86	87	54	13
41	78	107	108	109	110	111	112	113	88	55	14 +2
40	77	106	127	128	129	130	131	114 -6	89	56	15
39	76	105	126	139	140	141	132	115	90	57	16
38	75	104	125	138	143	142	133	116	91 -5	58	17
37	74	103	124	137	136	135	134	117	92	59	18
36	73	102 +3	123	122	121 -8	120	119	118	93	60	19 -1
35	72	101	100	99	98	97	96	95	94	61	20
34	71	70	69	68	67	66	65	64 -5	63	62	21
33	32	31	30	29	28	27	26	25	24	23	22

Jugada N°1

Student A

Finalizar turno

Selecciona una ficha...

Turno N°1

66 s

Oro:0 Err:0

Student B

Esperando turno...

Turno N°0

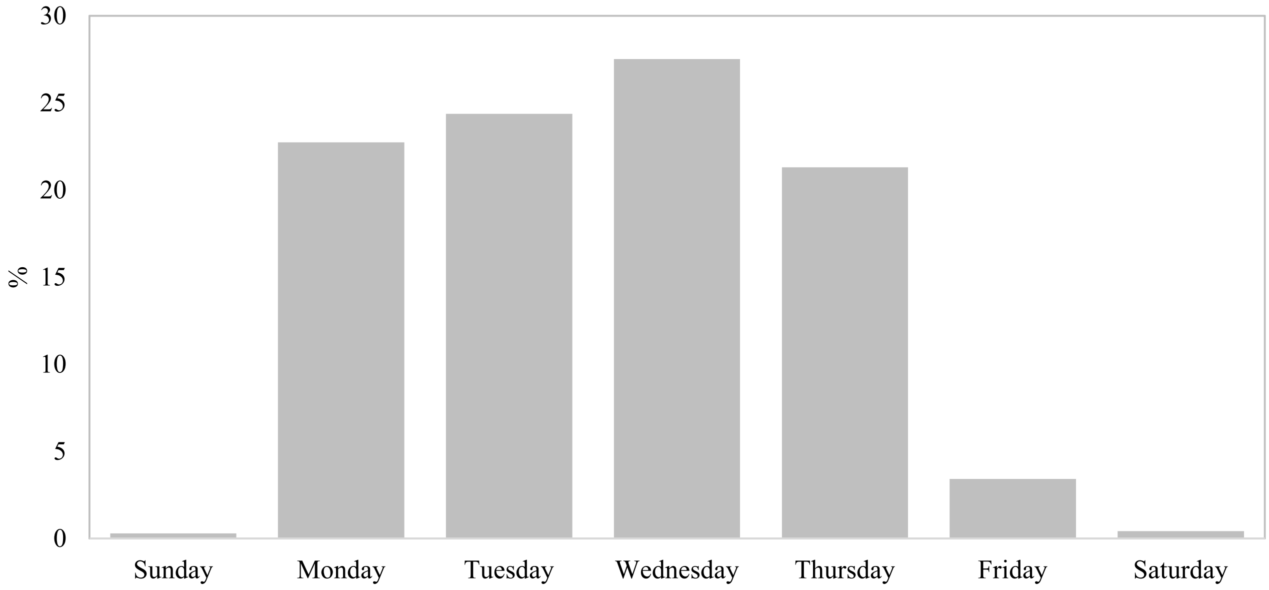
-

Oro:0 Err:0

The number that when subtracted from 30 the result is 18

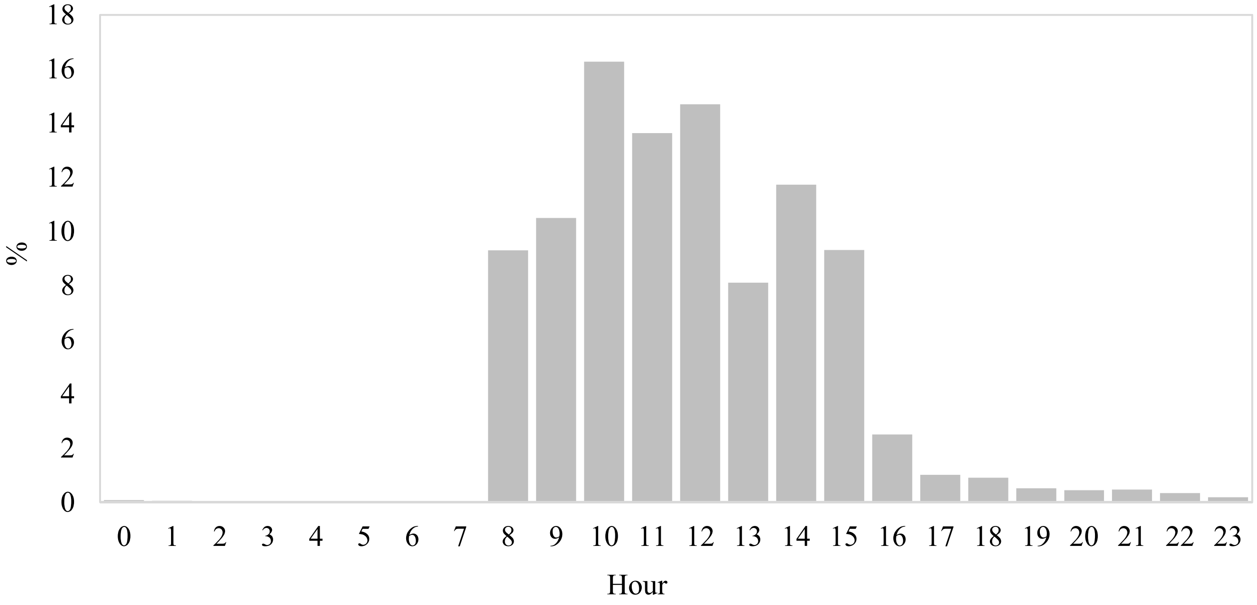
Contar de nuevo...

Figure 3: Platform Use by Day of the Week



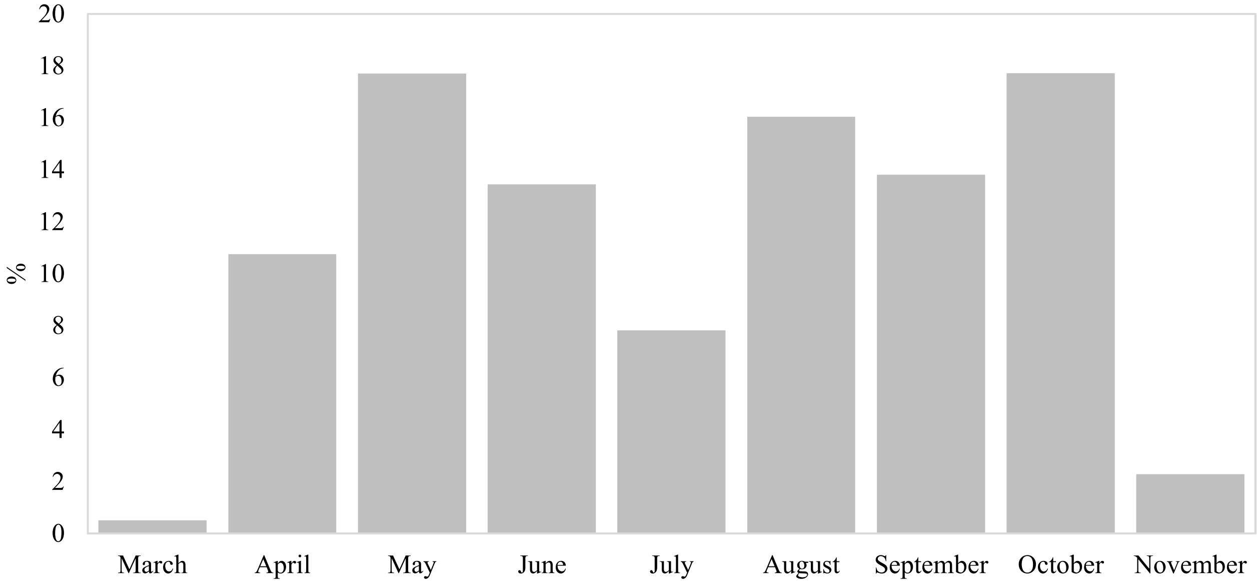
Notes: This figure presents the distribution of platform use by day of the week. Statistics correspond to the period from March 28, 2017 to November 6, 2017.

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: Construction of Non-Academic Outcomes

Outcome	Item	Item Source
Prefers lab for math	I would rather have math classes in the lab than in the classroom	Created for this study
Growth mindset	Intelligence is something that cannot be changed*	Claro et al. (2016)
	You can learn new things, but you cannot change your intelligence*	Claro et al. (2016)
	My parents say that I am capable of learning	SIMCE
Math intrinsic motivation	I enjoy learning math	TIMSS
	I wish I did not have to study math*	TIMSS
	Math is boring*	TIMSS
	I learn many interesting things in math	TIMSS
	I like math	TIMSS
	I like any schoolwork that involves numbers	TIMSS
	I like to solve math problems	TIMSS
	I look forward to math lessons	TIMSS
	Math is my favourite subject	TIMSS
Math self-concept	I usually do well in math	TIMSS
	Math is harder for me than for many of my classmates*	TIMSS
	I am just not good at math*	TIMSS
	Usually I do well in math tests	SIMCE
	Math is easy for me	SIMCE
	My teacher tells me I am good at math	SIMCE
Math anxiety	I am afraid that math questions are too hard for me	SIMCE
	I am worried of having bad grades in math	SIMCE
	I get nervous before math tests	SIMCE
	I get nervous when I don't understand a math homework	SIMCE
Teamwork	I prefer working as part of a team to working alone	PISA
	I find that teams make better decisions than individuals	PISA
	I find that teamwork raises my own efficiency	PISA
	I enjoy co-operating with peers	PISA

Notes: This table presents the items used to construct the non-academic outcomes and the source for each item. All non-academic outcomes are typically constructed as indices that combine different items. Items whose source is TIMSS, PISA, Claro et al. (2016), or marked as "Created for this study" were included in the endline questionnaire that we applied to students participating in the experimental evaluation. In contrast, items whose sources is SIMCE were included in the questionnaire that was applied as part of the 2017 Chilean national standardized exam for fourth grade. The source TIMSS includes items from questions 16 and 18 of the student questionnaire for fourth graders of the 2015 Trends in International Mathematics and Science Study (IEA, 2014). The source PISA includes items from the index "valuing teamwork" developed for the 2015 report on "Collaborative Problem Solving" of the Programme for International Student Assessment (OECD, 2017). The source Claro et al. (2016) includes items from this study which aimed to measure growth mindset in Chilean students. The source "Created for this study" refers to items that we developed without using an external reference. All questions are 4-point Likert scales with the sole exception being the item "I would rather have math classes in the lab than in the classroom" that is a yes/no question. Items that are marked with an asterisk (*) need to be reversed to compute the index for an outcome. For example, responses to the item "Math is harder for me than for many of my classmates" are reversed so that higher values correspond to higher math self-concept.

Table A.2: 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.3: 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.4: 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.