IDB WORKING PAPER SERIES No. IDB-WP-561

Challenges in Educational Reform:

An Experiment on Active Learning in Mathematics

Samuel Berlinski
Matías Busso

February 2015

Inter-American Development Bank
Department of Research and Chief Economist

Challenges in Educational Reform:

An Experiment on Active Learning in Mathematics

Samuel Berlinski*
Matías Busso**

* Inter-American Development Bank and IZA
** Inter-American Development Bank

Inter-American Development Bank
2015

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library

Berlinski, Samuel.
Challenges in educational reform: an experiment on active learning in mathematics / Samuel Berlinski, Matías Busso.
p. cm. — (IDB Working Paper Series ; 561)
Includes bibliographic references.
1. Education, Secondary. 2. Mathematics—Study and teaching. 3. Active learning. 4. Curriculum change. I. Busso, Matías. II. Inter-American Development Bank. Department of Research and Chief Economist. III. Title. IV. Series.
IDB-WP-561

http://www.iadb.org

Abstract[1]

This paper reports the results of an experiment with secondary school students designed to improve their ability to reason, argument, and communicate using mathematics. These goals are at the core of many educational reforms. A structured pedagogical intervention was created that fostered a more active role of students in the classroom. The intervention was implemented with high fidelity and was internally valid. Students in the control group learned significantly more than those who received treatment. A framework to interpret this result is provided in which learning is the result of student-teacher interaction. The quality of such interaction deteriorated during the intervention.

JEL classifications: C93, I21, I28, O32
Keywords: Education, Active learning, Curricular reform, Technology, Field experiments

[1] This project is the result of a collaborative effort involving many people. In particular, we would like to thank Horacio Alvarez Marinelli, Floria Arias, Moritz Bilagher, María Eugenia Bujanda, Elsie Campos, Marco Cordero, Ulises Cordero, María Antonieta Diaz, Alvaro Gamboa, Torie Gorges, Mauricio Holtz, Teresa Lara-Meloy, Richard Mayer, Jeremy Roschelle, and Magaly Zúñiga. For helpful comments we also thank Julián Cristia, Luca Flabbi, and seminar participants at CEP (LSE), EDePo (IFS), Fundación Omar Dengo, George Washington University, IDB, Queen Mary - University of London, LACEA (2013), NEUDC (2013), RECODE (2013), Royal Economic Society Conference (2014), Universidad de Chile, Universidad de San Andrés, Universidad Nacional de La Plata, and University of Michigan. Juliana Chen Peraza and Rosa Vidarte provided excellent research assistance. We gratefully acknowledge financial help from the Inter-American Development Bank and Fundación Costa Rica USA. We have no relevant or material financial interests that relate to the research described in this paper. This research was approved by the Consejo Superior de Educación de Costa Rica (CSE-SG-168-2012) and by the Institutional Review Board of Fundación Omar Dengo. This experiment has been registered in the American Economic Association RCT Registry with number AEARCTR-0000337. The views expressed herein are those of the authors and should not be attributed to the Inter-American Development Bank.

1. Introduction

Mathematical competence is a fundamental skill for personal fulfillment, active citizenship, social inclusion, and employability in the modern world. In this paper we report the results of an experiment devised to affect the way mathematics is taught and learned in Costa Rican secondary schools. The objective was to create a scalable intervention that would allow students to achieve mathematical competence. That is to say, students' ability to think, reason, argument, and communicate using mathematics. This concept is prevalent in the design of PISA examinations (OECD, 2009) and in curricular reforms in many countries, including Costa Rica and the United States.

Teaching strategies underpin all learning in the classroom. They determine what is learned and the nature of the interactions between students and teachers. As a recent study from the European Commission highlights (Eurydice, 2011), in order to achieve mathematical competence a common practice pursued by many countries is to give students a more active role in the generation of knowledge. "Moving away from the traditional teacher-dominated way of learning, active learning approaches encourage pupils to participate in their own learning through discussions, project work, practical exercises and other ways to help them reflect upon and explain their mathematics learning" (Eurydice, 2011: 56).

We created a pedagogical intervention designed to give students a more active role in the learning of geometry—one of three units of the seventh grade curriculum, or about three months of teaching. A key aspect of this change relies on providing students with guided opportunities to explore and discover. In mathematics, a potentially important lever in this process is the use of technology.

We randomly assigned 85 participating schools to treatment and control groups. All students (18,000) and teachers (190) in the seventh grade of these schools participated in the experiment. Treatment schools received the active learning intervention. In addition, in order to assess the role of technology keeping constant the pedagogical approach, we randomized treatment schools to receive no technology, an interactive whiteboard, a computer lab, or a laptop for every child in the classroom.

We commissioned the design of pedagogical material for this project to local experts advised by a team from a leading international education academic organization. In order to support teachers and students and with an eye directed at improving fidelity of implementation as

well, we created a teacher's manual and a student's workbook (one for every modality of the intervention but none for the control). Technology was introduced through a set of applets (created on a software familiar to Costa Rican teachers) designed to help students and teachers explore the key concepts of the unit.

In coordination with our local partners, we ensured that all the new resources were in place at the time of the implementation and suitable technical support was provided to guarantee that all resources were functional during the experiment. Teachers received 40 hours of on-site and distance training with virtual support, achieving a 95 percent participation rate. All the teachers in treatment arms received a laptop computer and a manual. All the students in seventh grade (with the exception of the control group) received a workbook.

In collaboration with local and international experts, we designed a psychometrically valid test of geometry to measure the impact of this intervention. The objective of the test was not only to measure the content knowledge of the students but also their mastery of higher-order geometric practices that require, for example, that students pick, compare, justify or refute conjectures and propositions. Before the start of the experiment we tested all the students in their general knowledge of sixth grade mathematics using a standardized international test prepared and administered under the supervision of UNESCO.

Geometry learning was the target outcome of the experiment. However, with a complex intervention like this one, it is of equal importance to understand how we affected the behavior of students and teachers. Only by understanding these underlying mechanisms can we learn why the target outcome has changed. For this purpose, we collected teacher and student surveys that use scales validated in psychology and educational research to measure class dynamics, teaching practices, attitudes and beliefs. We also collected classroom observations to further attest to the changes reported by teachers and students.

Randomization yielded groups with similar observable characteristics. The experiment was implemented with high fidelity. Materials and equipment were distributed where and when expected. They remained functional throughout the experiment. Teachers and students made use of their respective manuals. Indeed, there were significant changes in class dynamics with more participation from students. Teachers in the treatment arms were open to the innovations we introduced in the classroom.

Surprisingly, we find that the control group learned significantly more than any of the intervention groups. The students using only the active learning approach learned about 17 percent less than the status quo. The loss in the group that also received technology was 25 percent of a standard deviation. We also find that the best students were harmed the most by this intervention. Concurrently, their behavior deteriorated and they were less engaged with learning mathematics. The evidence suggests that teachers went through the motions as prescribed but did not master the innovation in a way that would have allowed students to benefit the most from it.

The results of the experiment are not a fluke. This experiment was an example of a salient educational policy. It was internally valid and performed on a nationally representative sample of schools. The resources were deemed useful for classroom use by the teachers and they bought into the changes we proposed. The main outcome was a psychometrically valid measure of geometry knowledge.

The intervention we study in this paper is shorter in duration than many programs that are evaluated in the education and economics literature.[2] However, duration is only one of many factors that we should consider when designing and interpreting the results of an experiment. One of the key trade-offs in complex social experiments lies in balancing the length of the intervention with our ability to control the experimental conditions. In our case, we designed an evaluation with the objective of isolating whether a particular pedagogical reform (active learning) could be effective at raising student knowledge under ideal conditions (i.e., as close as possible to an efficacy trial). In order to isolate the main ingredient of this reform we attempted to control many more levers than would be possible under the business as usual conditions that characterize the evaluation of most educational reforms (i.e., efficiency trials). For example, the quality of the pedagogical team assembled, the hours devoted to work on the material by the team, the depth of the material, the quality and quantity of the technology resources and support available to the schools, and the length of training would be prohibitively costly to reproduce for most research teams over a longer period of time.

Our paper achieves the objective of testing the impact of our main ingredient (active learning) on student achievement in the short run. There is much value added from this exercise. First, it is unlikely that in the short run a business as usual reform (i.e., akin to an efficiency trial)

[2] Note, however, that the intervention protocol and length make this intervention comparable to other high-quality studies assessed in educational resources such as What Works Clearing House (http://ies.ed.gov/ncee/wwc/) and Best Evidence Encyclopedia (http://www.bestevidence.org/).

will achieve learning gains. Second, any hurdles that presented in the efficacy trial are likely to be observed in an efficiency trial. Therefore, the evaluation and implementation of a curricular reform will require a long-run perspective on processes and outcomes.

This study speaks to a growing literature in economics that emphasizes the necessity of studying and ultimately identifying successful pedagogical approaches. For example, Dobbie and Fryer (2013) peep into the black box of 39 charter schools in New York and correlate data on school practices with credible estimates of school's effectiveness. Fryer (2014) looks at the effect of injecting successful charter school strategies into traditional public schools. Machin and McNally (2008) evaluate the reading and overall English attainment of a national pedagogical strategy designed to raise standards of literacy in primary schools through the introduction of a structured and pre-specified pedagogical strategy. The Measures of Effective Teaching (Kane et al., 2010, and Kane et al., 2012) project, designed to identify successful teachers and teaching strategies, also relies heavily on classroom observations as well as student and teacher surveys of the type we administered in our study.

There are also a number of rigorous evaluations in economics that measure the effects on student learning of providing classroom resources to schools such as flipcharts (Glewwe et al., 2004), textbooks (Glewwe, Kremer and Moulin, 2009), libraries (Borkum, He and Linden, 2012), school computers (Angrist and Lavy, 2002) and student laptops (Cristia et al., 2012). The impacts of these interventions are modest at best.[3] McEwan's (2013) recent meta-analysis of randomized control trials in primary schools of developing countries uncovers effect sizes that range from 0.08 to 0.15 of a standard deviation for these types of interventions. The failure of these resources to leverage student learning is commonly attributed to either the interventions not addressing curricular objectives or student needs, or to teachers receiving limited training on how to use these additional inputs effectively. Our study was set up to address many of these concerns, which makes our findings even more striking.

There is surprisingly little empirical evidence on the effectiveness of competing teaching approaches in mathematics. A recent report of the National Mathematics Advisory Panel on instructional practices in mathematics concludes: "For none of the areas examined did the Task Group find sufficiently strong and comprehensive bodies of research to support all-inclusive

[3] Angrist and Lavy (2002) study a program that funded the introduction of computers and software for computer-assisted instruction in Israeli elementary and middle schools. Their quasi-experimental evaluation suggests that the program may have had a negative impact on mathematics learning for 4th and 8th graders.

policy recommendations of any of the practices addressed" (Gersten, 2008: 6-189). Among the practices evaluated the panel looked at the use of teacher-centered versus student-centered approaches and the use of technology in the classroom.[4]

We proceed as follows. In Section 2 we present a conceptual framework. In Section 3 we discuss the distinctive features of our experiment. In Section 4 we explain the data collection process. In Section 5 we discuss the empirical strategy. In Section 6 we present our research sample and discuss the internal validity of the experiment. In Section 7 we show our results on fidelity of implementation, test scores and class dynamics as well as some robustness checks. Section 8 concludes.

2. Conceptual Framework

Educational authorities around the world are often engaged in setting up curricular reforms with the goal of improving the pertinence and quality of student learning. This process usually involves changing syllabus content as well as the pedagogical approach used in the classroom. As teachers and students are usually rooted in old teaching/learning habits, the implementation of these reforms requires professional development, the design of appropriate classroom material, and the provision of adequate resources for teachers and students. Despite these investments it is the teacher who, ultimately, chooses how to lead her class and therefore directly affects the actual result of reforms. In this section we provide a simple conceptual framework to think about our experiment and its results.

Learning in schools occurs in a classroom environment. The knowledge gained in the classroom by an individual student depends on her effort, the inputs she received from her teacher (teaching inputs) and the effort of her peers (Albornoz, Berlinski and Cabrales, 2010).[5] Students choose effort based on the value they attached to learning, their beliefs (given teaching inputs) about the consequences of their effort on learning (Chassang, Padro i Miquel and Snowberg, 2012), and the cost of providing this effort. This cost depends on her innate ability and on the knowledge and learning skills accumulated over her schooling years.

[4] Cheung and Slavin's (2011) meta-analysis of 75 high-quality studies of educational interventions with a focus on mathematics learning in K-12 that involves the use technology almost exclusively finds papers that look at the use of technology as a supplement to teaching in the form of drills and practice. Similarly to McEwan (2013), it reports gains of around 10 percent of a standard deviation for this use of technology. Cheung and Slavin (2011) find no high-quality studies pertaining to the use of interactive whiteboards.

[5] Other aspects beyond the classroom such as family environment, school environment, etc., are also relevant. However, we abstract from them in this discussion.

Classroom teaching is a process by which the teacher decides simultaneously an allocation of time and resources between tasks for every student (i.e., teaching inputs) given her expectation about student effort. The allocation chosen by the teacher depends on her objective function, her beliefs about the consequences of these allocations, and the cost of providing inputs. Incentives in schools are usually low-powered and therefore intrinsically motivated teachers can be assumed to maximize a weighted average of the knowledge gained by their students during the school year.

The cost of providing inputs for teachers depends on their teaching skills (innate and accumulated) and the resources available to them. This is to say, how time and resources are allocated in the classroom between tasks requires effort in preparation and classroom delivery. The latter involves not only the allocation of time and activities in the classroom but also the pertinence of teacher reactions to student participation. Different allocations of time and resources may achieve different learning outcomes for students.

Given the status quo and using a revealed preferences logic, changes in teaching inputs (as occur in educational reforms) require that teachers either modify their beliefs about the effect of the new allocation of inputs on educational outcomes or that the costs of adopting these changes be reduced. In the usual package of educational reform, teachers receive professional development targeted towards both reducing the cost of adoption and affecting the beliefs about the productivity of the new allocation of inputs.

Teachers' decision to comply with the reform (or take-up) is part of a dynamic process marred with uncertainty. It is dynamic in the sense that changing classroom inputs requires investments today (i.e., there are adjustments costs), but the benefits of these efforts may not necessarily be reflected in today's classroom achievements. For example, teachers and students have accumulated skills in using classroom inputs in a particular way, and the switch may result in the loss of some accumulated teaching/learning skills (Helpman and Rangel, 1999).[6] Learning-by-doing by teachers and students may undo this loss in the long run. Therefore, achieving the expected productivity gain from an innovation may require time and practice with the new set of inputs.

[6] There is some previous evidence that innovations that today are considered successful drivers of development (e.g., the micro-computer and the dynamo) did not lead to an automatic increase in productivity (see David, 1990). Helpman and Rangel (1999) provide an interesting explanation for this phenomenon. If productivity with a given technology increases with use, then the switch to another technology may lead to falls in output in the short run if the accumulated skills are not completely transferable.

The process is also uncertain because the effect of the new allocation of inputs is unknown ex-ante. For example, teachers and students may have types that make them more or less productive with different allocations of teaching inputs. If teachers are imperfectly informed about these types, they can learn about them through experimentation as in Karlan, Knight and Udry's (2012) analysis of entrepreneurial experimentation.

Our intervention encourages a new allocation of time to different tasks in the classroom accompanied by a new set of classroom and teaching materials. We support teachers in this transition by providing training commensurate to the length and depth of the experiment. Since our main objective is to affect classroom learning during our experiment, the main outcome is a test designed to measure learning of content and skills that should have been learned in the classroom in that period. As our conceptual framework highlights, for the experiment to achieve its objective it requires that teachers actually adopt the new approach and that students act accordingly. Therefore, we also measure the extent of teachers' compliance with new allocation of time and students and teachers attitudes towards these innovations.

The skills that teachers and students bring to the classroom may cause systematic differences in the effect the experiment has on learning. These differences may be driven by either differential take-up or by differences in the effectiveness of the new approach for different groups. For example, more teaching experience may be positively associated with the type of classroom management skills that are required for the new approach. However, those teachers may also be less enthusiastic or productive with the new approach because of their lack of familiarity with modern teaching methods. Also, they may have less time to reap the benefits of their investment. Our empirical analysis looks at heterogeneity of take-up for teachers and on learning and attitudes for students and teachers. For the above reasons, however, we are agnostic about the heterogeneous effects of the intervention.

3. Research Design

3.1 Context

Costa Rica is a relatively small middle-income developing country. In 2011, it had a population of 4,726,575 people, a GDP per capita of $10,085 (2005 PPP USD), and was ranked 69th place by the United Nations Human Development Index. The country boasts a long tradition of

publicly provided education, which is free and has been compulsory since 1870, with an adult literacy rate of 96 percent.

The educational system is divided into four levels: preschool education (ages 4-6); primary education (6 years); secondary education-middle school (seventh, eighth and ninth grade) and high school (tenth, eleventh and twelve grade). Currently, education is free and compulsory from the last year of preschool to the end of middle school.

The academic year runs from February to December. During 2012, the school year in middle school was 196 days long. In the seventh grade, students in public schools have six mathematics lessons a week of 40 minutes each (many of those arranged in contiguous blocks). The school year is divided into three terms, and the mathematics curriculum in the seventh grade covers Integers, Geometry, and Rational Numbers. Unlike the mathematics lessons in primary schools, mathematics is taught by a specialized teacher. In their annual teaching plan, teachers assign one term to each topic.

The teaching of mathematics in the Costa Rican context is not very different from those of other secondary schools in low and middle income countries around the world. The mathematics class is characterized by lecture-style teaching where the teacher writes down a definition or procedure in the blackboard using a particular example. Students take notes, ask questions and practice what the teacher explained. The theorems and procedures are taken as given truths and the objective is to practice until the students achieve mastery of their use. Teachers rely on a commercial textbook of their choice for teaching and provide students with a list of examples to practice.

Costa Rica has a long tradition of introducing technology into schools (Zúñiga, 2003). The Educational Informatics Program created in 1988 (a joint effort between the Ministry of Education and Fundación Omar Dengo) is a national informatics program that serves students in preschool, primary school, and middle school (since 2001). Technology is introduced into the school through computer laboratories with the objective (among others) of promoting the development of logical thinking by using computers to solve problems and working in teams. The program is not intended as a complement for teaching core subjects.

3.2 Experimental Design

In order to reduce implementation costs, we invited to participate in the experiment all schools that complied with the following two criteria: a minimum of 2 and a maximum of 12 classes of seventh grade math in 2011 and that were located in urban and semi-rural areas easily accessible by roads from the capital. This rule delivered an initial list of 100 eligible schools which were all invited to participate by the Ministry of Public Education towards the end of the 2011 academic year. From that initial list, 85 schools signed agreement letters and were included in the experiment (i.e., in the treatment randomization).[7] No school dropped out of the experiment after randomization; that is, our results include data from all 85 schools. All seventh grade math teachers and students in these schools participated in the experiment.

Schools were assigned to one of the following five conditions: control (20 schools), active learning (20 schools), active learning plus an interactive whiteboard (15 schools), active learning plus a computer lab (15 schools), and active learning plus one computer per student (15 schools). The intervention was much more costly to implement in schools that received some technology. Therefore, from a statistical power consideration, it is optimal to have larger sample sizes in less costly intervention groups (Nam, 1973, and Liu, 2003).[8]

Randomization was done by the research team on a computer in January of 2012. We assigned schools to their experimental status using a block randomization procedure based on schools according to seventh grade enrollment.[9,10] We notified the government and the schools of

[7] The Superior Council of Education of Costa Rica approved the experiment by resolution CSE-SG-168-2012.We obtained an IRB from Fundación Omar Dengo.

[8] A sample size of 40 schools with 60 students per school allows us to detect a minimum effect size of 0.25 standard deviation of the main outcome variable (a geometry test) with a statistical power of 0.8. The power calculation was done using the optimal design software version 3.0. It assumes that the treatment is delivered equiproportionally at the school level, and the outcome is measured at the student level. It also assumes a statistical significance of 0.05 and a conditional intra-class correlation (ICC) of 0.06. This ICC is consistent with our data and econometric model and it is also similar to that found in other studies (Hedges and Hedberg, 2007).

The relative cost of delivering the intervention and collecting data in the cheaper intervention group (active learning) was half of the cost budgeted for the more expensive intervention groups (those that received technology). Therefore following Liu (2003), the sample size in the cheaper intervention group should be around 33 percent larger.

[9] We stratified on enrollment not only to increase the precision of our estimates but, more importantly, because of costs considerations: ordering schools by enrollment was helpful to minimize the cost of buying technology.

[10] First, we ordered schools according to seventh grade enrollment in 2011. Second, we defined 15 bins: 10 bins of five schools and five bins of seven schools and we randomly ordered these 15 bins. Third, we populated the bins with schools. For example, if the first bin was of size five, we put the five largest schools into that bin; if the second bin was of size seven, we put the next seven largest schools into that bin, and so on. Fourth, we randomly assigned schools to treatment arms within each bin. In bins of size five every school had the same probability of receiving

the lottery results on February 2, 2012 (at the beginning of the school year). All the schools in the experiment, including those in the control group, were asked to teach Geometry during the second trimester (May 17 - September 5) to accommodate teacher training, the deployment of technology, and the printing of classroom materials.

In Table 1, we compare the mean characteristics of the 85 schools that participated in the experiment with the characteristics of schools in Costa Rica. The first column shows statistics for the 85 schools in our sample. Column 2 shows statistics for schools in Costa Rica that satisfy the experiment eligibility criteria (i.e., size and urban/semi-rural location). Column 3 shows statistics for the whole country. On average, schools in the experiment tend to be slightly larger and are growing a bit more slowly than schools in the rest of the country. They have similar infrastructure (as measured by access to library, computers, number of classrooms, and restrooms) to other schools in urban and semi-rural parts of the country as well as similar demographic characteristics.

3.3 Intervention: Active Learning[11]

The intervention aimed for students to achieve mathematical competence as defined by PISA (OECD, 2009) and the mathematics curricula of many countries including Costa Rica. Mathematical competence is understood by the student's ability to think, reason, argument and communicate using mathematics. This requires that students pose and solve mathematical problems and model mathematical situations using appropriate representations, symbols, tools and technology.

Following the advice of local experts it was considered that generating mathematical competence required moving away from a traditional teaching style to a more active learning approach. In terms of our conceptual framework this would require a new allocation of teaching inputs. In order to reduce the cost for teachers of adopting this pedagogical approach, we commissioned the design of pedagogical material by local experts from Fundación Omar Dengo and Universidad de Costa Rica, advised by a team of international experts from the Center for Technology in Learning (CTL) at SRI International.

---

each treatment whereas in bins of size seven a school had a probability of 1/7 of receiving a treatment that involved technology and 2/7 probability of going into a control or the active learning treatment arm.

[11] This section draws from a report prepared for this project by Arias and Zúñiga (2012a).

As part of the design of the experiment, we produced a teacher's manual and a student's workbook. The teacher's manual was elaborated with a structure that would not demand significant time for study by teachers. The sessions covered all the materials in the seventh grade Geometry curriculum of Costa Rica.[12] Each session had the same structure to help teachers with the use and understanding of the material. The main body of the session was divided into three activities: exploration, formalization, and practice of concepts.

The first of these activities is the largest departure from the traditional classroom model. It relies on a cycle that starts by presenting a situation to students. The teacher then guides the students to predict how that situation will unfold, to compare those predictions with a set of mathematical realities, and finally to explain the differences between their predictions and these facts. The ultimate goal is to construct and understand the underlying geometric concept that governs the situation. In contrast to the traditional lecture style, the students have a very active role in this process and the teachers a less controlling one.

After exploration the teacher is responsible for formalizing the knowledge previously established. However, this does not mean that we expect the teacher to recite or copy mathematical results on the board (a strategy usually pursued in a traditional lecture style). In fact, we introduced a cycle in which the teacher clarifies concepts, formalize them, and finally verify that the students comprehend what is being taught. During this process the students still have active participation.

Finally, the practice part is the more akin to the usual geometry class. It includes some applications of the concepts that were studied and the conclusion of the session. Unlike in the typical class, the manual does not offer a long list of exercises. The idea is that the mathematical work started during exploration, and this just provides an opportunity to consolidate what has been learned.

In order to standardize and facilitate the work of the students and to provide support to the teachers in implementing this pedagogical approach, we created a student workbook as well. The student workbook has hands-on paper-based activities and is identical in knowledge content to the teacher's manual. The main difference is that the teacher's manual has advice on how to

[12] The teacher's manual for the technology interventions included three more sessions at the beginning to introduce teachers and students to the use of technology in the classroom.

proceed or motivate the students at different points in the lesson; this is not included in the student workbook.

The materials considered the use of other class resources that facilitate these three activities: software in the technology arms and images or paper in the treatment arm with no technology. For instance, in the exploration case, students were presented with a situation (e.g., three different triangles printed on paper or on the screen in which the angles were highlighted with a different color). They were next asked to make a conjecture about the sum of internal angles. In the technology group, they form the conjecture by manipulating the shape of the triangles with the mouse and to add up the value of the internal angles. In the no technology group, students form the conjecture by cutting and pasting the angles of the four triangles.

Beyond the validation received by experienced mathematic teachers in Costa Rica and the support from international experts of CTL at SRI, the material was also reviewed by those in charge of training the teachers and ultimately received the feedback from the teachers who participated in the training (about 45 days before starting the experiment). One manual was printed for each of the four conditions. All students received a student's workbook with their name tag before the beginning of the experiment.

3.4 Intervention: Technology

The class is structured around a single pedagogical design independent of technology. This is to say, learning is driven by the mathematical actions that are required and by the student-teacher role in producing these activities rather than the technology. This is a key concept in our research design because, by providing a common pedagogical approach for the experiment, we are able to disentangle the role of the pedagogical setting from the contribution of technology.[13]

Of course, all this raises the question of how we planned the use of technology in this experiment. The use of technology in the mathematics classroom (like other manipulatives) contributes to learning because it allows time to be devoted to activities that are harder to do using only a blackboard and a textbook, such as grouping and classifying objects, establishing relations, visualizing generalizations, and discovering properties. However, the introduction of technological resources in the classroom can be disruptive. It changes classroom routines for

[13] It also considerably simplifies the production of a large amount of pedagogical material.

both students and teachers, which can demand the establishment of new rules of engagement. It may also require substantive new knowledge from teachers.

Keeping these hurdles in mind, the study considered a relatively simple approach to the use of technology. First, we chose the software program GeoGebra, which was familiar to teachers in Costa Rica.[14] Second, rather than requiring teachers and students to program in GeoGebra, we developed a set of applets in which the students had a series of elements (or buttons) that could be used to manipulate geometrical objects. These manipulations were planned so as to put the students in the best possible position to visualize, explore, conjecture, and construct mathematical arguments.

The GeoGeobra applets were the same in the three technology options; what varied was the time of exposure that the children would naturally have and their opportunities to drive the exploration. For example, students who had individual laptops could use these to check predictions individually, while the teacher could perform the check phase in an interactive whiteboard setting.

The intervention required deployment and installation of equipment in the technology schools. This task was undertaken by a local NGO. Schools were inspected by an engineer after the lottery results were announced. In coordination with the principal, classrooms that fulfilled security and structural conditions were chosen for installation of the equipment.[15] In schools assigned to the one-to-one status, classrooms were equipped with one laptop per student, one desktop computer, one router, two laptop carts to store and charge the laptops (while not in use), and one LCD projector. The initial investment cost (including hardware and set-up costs) per student calculated over a life span of 5 years and using a discount rate of 10 percent was 111 dollars. Adding recurrent costs to compute the total cost of ownership took this figure to 130 dollars.[16]

In schools assigned to the interactive whiteboard status, classrooms were equipped with one interactive whiteboard, one desktop, and one router. The interactive whiteboard uses pressure-sensing technology, which means that a finger or any other writing object can be used

[14] GeoGebra (http://www.geogebra.org/) is free and open source multi-platform dynamic mathematics software for all levels of education that joins geometry, algebra, tables, graphing, statistics and calculus in an easy-to-use package.
[15] Minor adjustments were made in some schools as needed (e.g., wires and sockets were changed, walls were fortified to support the whiteboard). These changes were minor and very unlikely to affect the learning environment.
[16] To put these numbers in perspective, buying laptops for every student in primary schools in Costa Rica would be equivalent to a 20 percent permanent increase in the per-student educational expenditure (Berlinski et al., 2011).

to move, click and operate the computer. The initial investment cost per student was 34.9 dollars, and when recurrent costs are included the total cost of ownership over a five-year life span was 40.8 dollars.

Schools assigned to the computer lab status had one laptop for every two students available at least for 2 hours a week. Schools either used their existing computer laboratories or, if they had no laboratory installed, we created a mobile laboratory for them. The initial investment cost per student was 73.4 dollars, and when recurrent costs are included the total cost of ownership over a five-year life span was 88.3 dollars.

3.5 Teacher Training[17]

Teacher training for this experiment provided first-hand practice with the active learning approach and also aimed at affecting teachers' beliefs about the benefits of the new allocation of time and resources to different activities in the class. In particular, we emphasized the importance of devoting class time for student exploration and the role of the teacher as a guide/mediator for the students in this process. Training was not academic or theoretical in nature, but rather geared towards fulfilling the needs of teachers and students.

The training we offered to teachers in the experiment focused on the following areas:

- Giving teachers an immersion into the approach we use to develop mathematical competence.
- Familiarizing teachers with how to use the teachers' manual, the students' workbook and the GeoGebra applets. We placed particular emphasis in practicing didactic strategies and how to involve the use of applets in these activities.
- Familiarizing teachers with the technology they have been assigned to if any (i.e., interactive whiteboard or students laptops).

At the beginning of the training program, each teacher received their laptop, corresponding teacher's manual and a CD with the presentations for each training session, the GeoGebra applets, a manual for the use of the LCD projector, and other complementary material.

[17] This section draws from a report prepared for this project by Arias and Zúñiga (2012b).

The full training session had 40 hours, divided into 4 weeks[18] with 10 hours each week. From these 10 hours, 5 hours were allocated to on-site training and 5 hours to distance work supported by a virtual classroom. Training was organized by modality (i.e., active learning, one-to-one, interactive-whiteboard, and computer lab) and delivered at two regional sites.

A total of 130 teachers participated in the full training program. Among those who participated, 115 received a certificate that provided professional points in the Civil Service Career system. Successful completion of the training was based on attendance, classwork, and completion of a final assignment. There were 16 teachers who did not attend the full training session at some point in time. These teachers were offered a recovery training session; 9 of them attended and 7 were absent.

3.6 Design of the Assessment[19]

A central part of the design of any experiment is to determine the outcome measure.[20] We developed an assessment to measure learning and to potentially distinguish the gains from the different conditions. The process started at the end of the curriculum development phase.

First, we determined which content and skills should be included in the assessment. Following an in-depth analysis of the curriculum, a group of experts determined the topics to be covered in the assessment. In addition, we outlined the two types of conceptual skills that we expect students to develop during the unit: basic and higher-order skills.

Basic skills typically covered by the seventh grade syllabus require that students use parts of definitions, classify figures according to given properties, locate parts of figures (points, segments, angles), and make simple calculations to find a missing side or angle. Higher-order geometric skills require that students pick, compare, justify or refute conjectures and propositions; deduce a third observation from two givens; formulate or justify a geometric argument; and generalize from one or more cases. Table A1 details the differences between these skills.

We developed the actual test using the following procedure: i) we created a pool of items for both basic and higher-order content areas; ii) experts reviewed the items; iii) we carried out

[18] Training occurred between the April 9 and May 4.
[19] This sections draws from a report prepared for this project by Lara-Meloy et al. (2012).
[20] We followed the model of assessment development described in Shechtman et al. (2010).

think-aloud sessions with students on the higher-order items;[21] iv) we piloted the test and used Item Response Theory (IRT) to select items for the final assessment;[22] and v) a panel of experts reviewed the final assessment form.

This process resulted in a 32-item geometry test that we administered on pen and paper during the experiment. For grading simplicity we decided that all items had to be multiple-choice. The test has sound psychometric properties. The scale reliability coefficient (Cronbach's alpha) is 0.71 in the control group data. The test also shows a correlation of 0.52 across schools with the (end of primary school) general SAT that we describe below.

4. Data

We collected student and teacher data before the intervention started, between late April and early May of 2012. During the intervention we also collected teacher logs and class observations. From mid-August to mid-September, we collected another round of information from teachers and students. We additionally gathered administrative information provided by the schools.

The intervention affected nearly 18,000 students and 190 teachers in 85 schools. We collected data on all teachers and all schools. However, because of cost considerations, it was not feasible to collect data on all students. Each teacher was, on average, in charge of three sections (classrooms). Therefore, we decided to randomly select one section per teacher and collect data on that section only. This also had the advantage of allowing us to relate the results of class observations, teacher surveys with the information (surveys and tests) of students in those same classes.

We also administered a student survey before the beginning of the geometry unit to compare the distribution of characteristics of students in different treatment arms and determine whether the randomization had created comparable groups. The student survey was collected in the classroom and contained information on students' characteristics, their family and socio-economic background, and experience with computers.

[21] During the think-aloud sessions students solved the test exercises, explaining their reasoning behind their chosen answer to a trained observer.

[22] We modeled each item using a two-parameter logistic curve conditional on a single overall student test score. One parameter manipulates the location of the curve (difficulty parameter), and the second parameter affects the slope of the curve (discrimination parameter) when the probability of answering the question correctly is half. We discarded items with very high or very low estimates of the difficulty parameter and items with very low discrimination parameters. This corresponds to items that either did not correlate well with the total test scores or could be solved by intelligent guessing.

At baseline we administered a standardized achievement test[23] used in a regional study of educational attainment in Latin America in 2008. The test was prepared and administered under the supervision of the Latin American Laboratory for Assessment of the Quality of Education (LLECE) of UNESCO.

Towards the end of the geometry unit, we administered the geometry test discussed in Section 3.6; this is the main outcome of interest.[24] Additionally, after completing the test the students filled a student questionnaire designed to measure treatment compliance, fidelity of implementation, class dynamics, and students' attitudes towards mathematics.

Before and after the intervention took place, we also collected teachers' surveys. At baseline we asked for information regarding background characteristics of the teachers and general information about their mathematics class. This information is used to verify comparability of treatment and control groups. In a second survey, after the intervention, we incorporated a set of questions to measure fidelity of implementation and changes in class dynamics and teaching practices.

Throughout the questionnaires we asked students and teachers a series of questions that we later used to form several scales. Each scale was pre-specified and had been previously used and validated in other studies: the 2011 surveys designed by the University of Chicago Consortium on Chicago School Research, the Manual for the Patterns of Adaptive Learning Scales (PALS) developed by the University of Michigan, and several scales developed by CTL at SRI International.

Tables A2 and A3 list the scales (Column 1) and the questions (Column 2) used to build each scale. Column 1 also presents the source of each scale and the Cronbach's alpha reliability coefficient that each scale had in our sample. The questions pertaining to each scale were randomly mixed in the student and teacher survey instruments. Students and teachers could give categorical answers of the type "strongly agree," "agree," etc. We aggregated those answers into scales using a maximum likelihood principal components estimator where only one latent factor was retained.[25] The models were estimated on the control sample only. Column 1 in Tables A2 and A3 present the eigenvalue of each latent factor and Column 3 shows the loading associated

[23] The Second Regional Comparative and Explanatory Study (SERCE).
[24] The endline test was administered between August 20 and September 10, 2012.
[25] Results were almost identical when building the scales via a polytomous IRT model.

with each variable. After the prediction was computed to produce each scale, we standardized them using the mean and standard deviation of the control group.

As is common practice in the educational literature, we collected two additional pieces of information from the teachers. First, we asked them to fill a short teacher log every month on a pre-specified date where they reported concrete features of their instructional practices, such as topics covered, pedagogical strategies used, any events that affected the normal delivery of lectures (e.g., technical problems with the equipment), and actual use of equipment, textbooks, and other class material.

One potential limitation of logs is that teachers may reflect intended rather than actual behavior. Kennedy (1999) argues that logs are effective for collecting information about topics and tasks, particularly in mathematics, but are less well suited for capturing class dynamics. She adds that class observations, on the other hand, "can document the intellectual complexity of the work students are doing in class. By observing the kinds of intellectual demands that are placed on students in the classroom, we might be able to infer the kinds of intellectual work in which they are likely to show improvement" (p. 347). Therefore, we aimed at conducting one classroom observation per teacher by an external observer. In order to homogenize observation and recording criteria, we created a protocol and an instrument to perform the class observations, which were done by properly trained psychology students.

Finally, we collected a rich dataset of administrative information about the school. This information included data on school location; computer equipment and infrastructure; size in terms of students, classrooms and teachers; repetition rate, etc. The information for 2011 was used mainly to stratify the sample when we randomized treatment. The information for 2012 was used to assess balance of characteristics and as a sampling universe to build our sample of students to be surveyed and tested.

5. Empirical Strategy

We estimate by ordinary least squares a set of models of the following form:

$$Y_{isj} = \alpha_0 + \sum_{k=1}^{4} \alpha_k T_{sj}^k + \beta X_{isj} + \delta_{sj} + \varepsilon_{isj}, \qquad (1)$$

In model (1), Y_{isj} is an outcome of interest (e.g., student performance on the geometry test or teachers' openness to innovation) of individual *i*, in strata *s* and in school *j*. T_{sj}^k is a dummy

variable equal to one if the school j was assigned to treatment k={1,2,3,4}={active learning, interactive whiteboard, computer lab, one-to-one}. For brevity, in the main tables of the paper we pool together the three technology arms into one group (letting k={1,2}) and present the results with four treatment dummies in the appendix. We condition on X_{isj}, a vector of student, teacher and school control variables. We also condition on strata fixed effect, δ_{sj}, and we include a random specification error, ε_{isj}. The parameter of interest α_k is the average treatment effect[26] (e.g., the average effect on student performance in the geometry test of being in a one-to-one school versus the status quo).

The vector X_{isj} includes several variables from student (pre-treatment SAT score, dummies of gender, dummies of age, mother's education, and number of books at home), teachers (gender, age, and years of experience) and schools (number of students in seventh grade math, number of classrooms in seventh grade math, dummies of province, and a dummy variable that is equal to one if the school had a computer lab already available before treatment). Our main results control on observable characteristics since this might lead to more precise estimates. We relegate to an appendix results without control variables X_{isj}.

All models take into account the potential correlation between students' and teachers' performance and behavior by clustering the standard errors at the school level (i.e., the unit of randomization). However, the standard error estimates are typically not sensitive to the level of clustering.

6. Research Sample and Internal Validity

Our research sample covers all 85 schools and all 190 teachers that participated in the study. All students were eligible except for those with some kind of known physical or cognitive disability.[27] We surveyed one complete class of eligible students per teacher.[28]

[26] There is perfect compliance of the schools with the randomization status. Therefore, there is no practical distinction between the average treatment effect and intention to treat effect.

[27] These are 1 percent of the students in our schools, who are usually allowed special dispensations during exams which we could not accommodate. These students were identified before the intervention started.

[28] Of the 190 teachers, 46 are in the control group, 44 in the curriculum group, 36 in the interactive whiteboard group, 28 in the group that received computers labs, and 36 in the one-to-one group. The distribution of students in these classes were as follows: 1,108 in control, 1,196 in active learning, 970 in interactive whiteboard, 740 in labs and 868 in one-to-one. As we show in Table 3 and in Appendix Table B3 the number of students per school and the number of classes/teachers per school are balanced across treatment arms.

Participation rates in our measurement activities were high as we describe in Table 2. Row 1 shows that the student post-treatment survey and geometry test had a non-response rate of 9.1 percent. Columns 2 and 3 show the regression coefficients and standard errors of a model described by equation (1), without controls, where the dependent variable is a dummy equal to one if the student was missing on the geometry test day and zero otherwise. The base category is the control group. There are no significant differences between treatment and control. Column 4 presents the p-value of a joint Wald test of the null hypothesis that all coefficients are equal to zero, which we cannot reject.

The administration of the test was contracted to a polling company that surveyed schools and classes during a four-week period. We followed the protocol used by UNESCO for the administration of the test. The teacher did not know the test ahead of time and was not in the classroom during the test. The exam was administered by trained external invigilators who were instructed not to answer students' questions. We agreed with the survey company a schedule that would balance the days on which schools in different treatment arms would be visited. The geometry test was administered on average about 6 days before the end of the second term. By design, it can be seen from row 2 columns 2 and 3 that there are very little differences in dates between schools in each treatment arm.[29]

The pre-treatment SAT test had a non-response rate of 21.1 percent with a larger non-response rate in the schools that received a computer lab.[30] Nevertheless, we cannot reject the null that all treatment arms had similar non-response rates at baseline.[31] In order to preserve sample size when conditioning on SAT in our econometric models we impute missing SAT using mean class characteristics and adding a categorical variable to denote this imputation.[32]

[29] In each visit, the survey team would provide the teacher with numbered copies of the exam so that he/she would administer the test to the absent students within the following week. About 5.6 percent of the tests were administered by the teacher rather than by the survey team, with no statistically significant difference between students in different treatment status.

[30] Differences in response rates between the baseline SAT and the endline geometry test can be explained by differences in the strategies used by UNESCO and the contracted polling firm to collect data. The polling firm scheduled visits to school so that they coincided with the normal math class schedule. UNESCO, on the other hand, announced to the school a slot where they would be visit and administered the test to the students that were present.

[31] When we use three dummies for technology rather than one (Appendix Table B2), we find that Missing SAT and Non-eligible students are statistically significant at the 10 percent level for students assigned to the laboratory condition.

[32] Our main results are robust to dropping observations with an imputed SAT in the sense that the point estimates do not change. However, the standard errors are slightly larger.

We collected teacher data using three instruments: surveys, class observations, and logs. The non-response rate of teachers' surveys was very low in both waves; with no significant differences by treatment status. We set out to collect data from each class through a class observation. Due to logistical and budget constraints we only managed to visit 80 percent of the classrooms. Those that received technology were 6 percent more likely to be administered a classroom observation than control classes, but this difference was non-statistically significant.[33]

We also collected teacher logs at the end of June, July, and August. The first round of teacher logs was completed by 89 percent of the teachers. The non-response rate increased with time, reaching 16 percent in July and almost 24 percent in August. Unfortunately, non-response rates seem to be correlated with treatment. For that reason we decided to limit the use of these logs in our main analysis.[34]

Table 3 shows pre-treatment sample means and differences in those means across treatment groups. Overall these differences are small and not statistically significant. Half of the students in the research sample are female and on average they are approximately 13 years old. Approximately 42 percent of the students have mothers with primary education and 41 percent with secondary education. On average they report having 3 books at home, and 74 percent have access to computers at home, which suggests familiarity with technology. For all these variables, the Wald test of the null hypothesis that all coefficients are equal to zero cannot be rejected.[35]

The pre-treatment SAT score is perhaps the most important variable since it provides an indication of math knowledge acquired by these students before starting seventh grade. On average these students correctly answered 46.6 percent of the questions on the exam.[36] The differences among treatment arms are negligible.

The characteristics of teachers and schools are also very similar across treatment groups. On average about half of the teachers are male, and they have about 11 years of experience. Schools have on average 220 students and 2.2 math teachers in seventh grade math. The majority of them have some kind of computer lab and internet access. The repetition rate is 8.7 percent,

[33] When we use three dummies for technology rather than one (Appendix Table B2), we find that one-to-one classes were 12 percent more likely to be observed and this coefficient is statistically significant at the ten percent level.
[34] We only use this information in Figure 4.
[35] When we use three dummies for technology rather than one (Appendix Table B3), three out of the 60 coefficients are statistically significant. In particular, students in the interactive whiteboard are younger and more likely to be female than those in the control group. Also, teachers in the interactive whiteboard condition are marginally more likely to be male. However, the magnitude of these differences tends to be very small.
[36] In the nationally representative sample of sixth grade students in 2008 the response rate was 49 percent.

and 44.7 percent are suburban schools. We cannot reject the null that all of these variables are the same across intervention groups.

At the beginning of the school year, we announced to schools their treatment status. In Table 4 we show that by the time of the announcement 83.7 percent of the teachers had already been assigned to their classes, with no apparent differences in the different intervention groups.

We asked all teachers of the schools participating in the experiment to teach seventh grade geometry during the second term of the school year (the term suggested by the Ministry of Education). Only 3 classes (in the control group) out of 190 did not comply with this request and about 12.6 percent taught the introductory four units of geometry during the first term and then stopped at our request. Again, the majority of these classes are in the control group. If knowledge depreciates over time then this deviation should bias the results against the control group schools.

7. Results

We present the results of the paper in this section. We proceed in four steps. First, we show how the intervention changed teaching inputs. That is, we look at the change in the use of resources in class, the time allocated by the teacher to different activities, and the resulting change in student class participation. Second, we present the effects of the intervention on student learning and assess robustness of the test results. Third, we investigate treatment effect heterogeneity. We conclude this section by analyzing teacher attitudes, aspects of student-teacher interactions and student effort that might explain the test results.

Throughout this section we present OLS estimates of equation (1), which include controls for randomization strata, student, teachers and school characteristics. We present results only controlling for strata in Appendix Tables C. Not surprisingly, given that the variables are balanced, point estimates do not change.[37]

[37] As in Section 6, in Appendix Tables B we also present results using the three technology dummies.

7.1 Teaching Inputs

The take-up of the treatment was high. In other words, the usage of resources, the allocation of time and the resulting class dynamics, with more student participation, show high compliance with the treatment.[38]

We use the endline student surveys to create indicators of classroom material and technology use. The results for these outcomes are presented in the top panel of Table 5 where we estimate equation (1) using as a dependent variable whether students had access and report to have used the materials and technology we provided for their class or not.[39] All estimates for the use of classroom materials are positive and large. Indeed, we cannot reject the null hypothesis that all classes in the treatment arms used the materials. Did the teachers use the available technology in class? Again, we cannot reject the null hypothesis that classes assigned to technology used it.[40]

Using classroom observations we measure whether teachers and students were observed using different materials and tools in class. In the bottom panel of Table 5, we present results showing that students' workbooks and teachers' manuals were used in all treated classrooms and the technology arms used the prescribed software (GeoGebra). Interestingly, students did not appear to have used the Internet in the treated classrooms more than in the control schools. Finally, all treatment groups used the traditional blackboard less than the control classrooms.

The intervention promoted a new allocation of time in class with more time devoted to exploration rather than practice and a more active classroom experience for the student. We analyze these outcomes in Table 6. First, we analyze the time devoted to different classroom activities. Second, we constructed scales of active engagement in class and classroom activity. Third, we analyze measures of the math practices teachers used in the classroom.

In the first eight rows of Table 6, we use classroom observations to show the way time was allocated in the classroom. The observer recorded the duration in minutes of the three main

[38] As discussed in Section 3.5, the vast majority of teachers in the treatment arms participated and passed the training. The percent of teachers that were trained in each treatment arm is as follows: zero percent in the control group, 91 percent in the active learning group, 97 percent in the interactive board group, 100 percent of teachers in computer labs schools, and 94 percent in one-to-one.

[39] Each variable is a dummy equal to one if at least half of the students in the class reported to have had access and used class materials, interactive whiteboards, laptops or some technology. In Appendix Table D we report measures built using teacher data. We found very similar results using teacher-level data and class observation data.

[40] Reassuringly, interactive whiteboards were used only in interactive whiteboards schools and computers only in schools that should have received computers (Table B5).

moments of the class: exploration of new concepts, formalization, and practice. Similarly, s/he also recorded the amount of time allocated to different strategies used to teach: plenary lecture, class discussion, group work, in pairs or individually. Using this information, we constructed a set of variables that measure the proportion of total class time allocated to each moment and to each teaching strategy. In the treatment classrooms there was more time allocated to discussion and less to the plenary lecture and individual work. As our treatment intended, more time was devoted to exploration and formalization and less to practice.

In the ninth and tenth rows of Table 6, we use student questionnaires to study whether the intervention fostered an environment where students were more actively engaged. Looking at the questions that pertain to these scales (Appendix Tables A2-A3), we infer that students in the treatment group explained concepts to the class more often, prepared more exercises for others to solve, and frequently discussed possible solutions or arguments with other students.

Furthermore, in last two rows of Table 6, we show that students in all treatment groups were stimulated in ways consistent with the objective of achieving mathematical competence. In particular, the class observer recorded whether students made, explained and validated mathematical conjectures, explained relations between concepts, manipulated propositions, or discovered mathematical rules from observing and analyzing patterns. The first scale looks at students prescribed learning practices while the second looks at whether or not teachers purposefully fostered those practices. We see positive point estimates for all groups with larger magnitudes and statistical significance in the technology arm.

7.2. Student Learning

We interpret the changes in teaching inputs and in class dynamics as an indication that teachers, familiar with the intervention, took the option of using the offered materials and equipment. Unfortunately, this did not translate into gains in learning.

Table 7 presents the main results of the paper. The dependent variable is the score in the geometry test (computed using the IRT parameters of the control group) and then standardized using the mean and standard deviation of the control group.[41] Therefore the coefficients can be interpreted as the treatment effects in terms of that standard deviation.

[41] Results are basically the same if instead of constructing a test score using IRT we use the percent of correct answers as a dependent variable.

The treatment caused a significant loss in geometry knowledge. The average treatment effect of the active learning treatment alone is a reduction in test scores of 17.1 percent of a standard deviation. The effect of active learning with technology is a loss of 24.7 percent of standard deviation. Both coefficients are statistically significant and a Wald test cannot reject the null hypothesis that they are equal.

In rows 2 and 3 of Table 7, we separate the score between basic and higher-order skills items. Recall that the basic skills items are designed to measure foundational geometry abilities or basic concepts whereas the items related to higher-order skills are designed to measure higher-order geometric practices; which *a priori* are easier to acquire using the active learning approach. We find no differences in results when comparing the performance on basic and higher-order items to the overall performance. We speculate that in order for students in the treatment group to outperform students in the control group on the higher-order items they should have done at least similarly on the basic items, but they did not.

We next provide evidence that the results are robust. The geometry unit is divided into five sections: introduction, measurement and classification of angles, relations between angles, triangles and quadrilaterals. In Panel A of Figure 1, we remove, one at a time, all the items that belong to each of the five sections. If a given treatment group found the material on a section particularly difficult then we would find some reversal in the relative rank of the treatment effects when that section is not considered in the score. That is not the case.

The items on the test also vary by difficulty. In Panel B we classify the 32 test items into eight groups of four items each according to the number of percent correct answers in the sample. We remove one group of items at a time and re-standardize the score. The relative performance on the test is the same when any given difficulty group is discarded.

We also check, in Panel C, whether the results are driven by particular schools in the sample. To do this and still preserve the validity of the experiment, we make use of the stratification of our research sample. We remove one stratum at a time. Results are very stable, which suggest that no individual school or strata drives the treatment effect and that there is not much heterogeneity in the treatment effect with school size.[42]

[42] Panels A and B in Appendix Figure 1 presents a similar exercise as the one done for Figure 1 but instead of taking out one section at a time (Panel A) we estimate the impact only on items of sections 1, 2…, 5. Treatment effects are slightly higher (and noisier) than the average for earlier sections. Qualitatively, however, the results as well as the ranking of treatment effects across treatment groups are basically the same. A similar conclusion is reached when

In Figure 2, we present the coefficients of estimating equation (1) with the three technology dummies rather than the aggregating them into one dummy as in Table 7 (with the corresponding results in Appendix Table B7). Combining the one-to-one technology with the active learning approach led to an additional loss of 18.4 percent of a standard deviation, taking the total loss in this treatment arm to 35.5 percent. Results for computer lab are very similar to the results of the active learning approach. The usage of interactive whiteboards slightly ameliorates the negative impact of the change in pedagogy. Students in this group learned 15.5 percent of a standard deviation less than those in the control group.

In Appendix Table A4, we show one sided p-values of pair-wise comparisons between different treatments. In each case, the null hypothesis is that the treatment effects are equal and the alternative is that one treatment effect is smaller than another. Basically, the one-to-one treatment effect is smaller than any of the other treatment effects at standard levels of significance. However, we cannot reject that interactive whiteboards, computer labs, and the active learning approach without technology have the same deleterious effect. The results are not surprising as the interactive whiteboard is *a priori* less disruptive and closer to teacher and students classroom habits than computers. Although the availability of computers on a one-to-one basis was intended to give teachers and students more opportunities for meaningful hands-on exploration, it also introduced another layer of innovation, which may have further hindered the learning process.

7.3. Treatment Effect Heterogeneity

As we explained in our conceptual framework, the skills of teachers and students may affect both the input decision of the teachers and the gradient of the teaching inputs on student learning. We speculate that the higher the level of skills, the higher the opportunity cost of switching/complying but also the higher the productive effect of the new allocation of inputs. Thus, it is hard to predict *a priori* whether higher skills will result in more/less adoption or in worse/better results for students.

---

estimating the treatment effects only on a subset of items of a given difficulty (akin to Panel B of Figure 1). In the case of strata, because each stratum has only 5 to 7 schools/clusters, we estimated a local polynomial regression of the outcome on the strata dummies. We found that the treatment effect is always negative and has a u-shape with middle-sized schools performing relatively worse.

We use two measures of teachers' skills: years of teaching experience and teacher's implicit quality. We built the latter as follows. First, within treatment arms we compute the percentile rank of each student in the baseline SAT and in the geometry test. Second, we take the difference between the geometry percentile rank and the SAT rank and average it across teachers. Teachers that were able to improve further the students' percentile rank are considered better.[43] We measure students' skills by their pre-treatment SAT. To be parsimonious in our regression analysis we use the median in our sample of experience, quality, and skills to divide the sample into high and low groups of the underlying variable.

We start by looking at the adoption decision. We consider the outcomes of Table 5 and 6 and investigate whether there are heterogeneous effects on teachers' experience and quality. For the set of outcomes in each table, we estimate a set of regression models using equation (1) individually for each group and then test the null joint hypothesis that all coefficients in those models are equal between groups (e.g., high experience versus low experience teachers) for each treatment arm.[44] We cannot reject the null that the adoption was homogeneous between teachers with high and low experience or between teachers' of high and low quality.[45]

In Figure 3, we show a local polynomial regression of the geometry test-scores (controlling for strata fixed effects) on the three mediating variables. In each graph we show three lines. The dashed line is for the control group, the solid line is for those students in the active learning condition only, and the long-dashed line is for those students in the active and learning and technology groups. At the bottom of the graph, we overlap a histogram of the mediating variable, and the vertical line marks the median of the mediating variable distribution.

In panel A, we explore the relationship between treatment effects and teacher experience. Looking at the control group, test-scores first increase with experience up to about 7 years then fall monotonically until 21 years and start rising again afterwards.[46] Technology follows a similar pattern to the control group, while the active learning group looks flat over the whole range of experience. The treatment effect is negative at low levels of experience but the magnitude

[43] Reassuringly, this measure of teacher quality is orthogonal to treatment and is able to explain a large proportion of the total variance in geometry test scores.
[44] We implement this test using seemingly unrelated regressions and taking into account the covariance between groups and outcomes.
[45] To save space, we do not show these results, which are available from the authors upon request.
[46] The lack of monotonic relationship is not particularly surprising as several studies have found difficult to pin down any sort of systematic relationship between student test-scores and teacher experience (e.g., Aaronson et al., 2007, and Harris and Sass, 2011).

shrinks with experience up to a point where both treatment effects become positive. In Table 7 rows (A) and (B), we show that on average, classes led by teachers with experience below the median tended to perform worse than in classes with more experienced teachers.[47]

In panel B, we measure the relationship between treatment and teacher quality. First, we can see that the data supports our interpretation of our measure of teacher quality as we see that student geometry scores in the control group rise monotonically with the quality of the teacher. Although this relationship falls towards the end of the quality spectrum, there is little mass at that point. The treatment groups follow a similar pattern, but the gradient of quality is smaller. In Table 7 rows (C) and (D), we do not find significant differences in treatment effects for teachers of low and high quality.

In panel C, we look at the pre-treatment SAT. Most of the mass is towards the middle of the support of the distribution. Performance in the geometry test increases with pre-treatment SAT. The line for the control group is always above the lines for the treatment arms. The line for the control group increases faster with pre-treatment SAT than for the treatment arms. Therefore, there is a larger loss for students with higher knowledge of math at baseline.

We confirm this result in the second panel of Table 7 where we show separate estimates for students below (row E) and above (row F) the median of the ability distribution. The main differences in treatment effects are observed in the active learning group. In the technology group that difference is smaller. A possible explanation is that the traditional lecture teaching style was geared towards the more able students. The intervention changed class dynamics, assigning relatively more emphasis to tasks that benefited lower-achieving students.[48]

7.4. What Went Wrong?

A loss in student learning has to be the result of either lower effort from students or teachers, or a worsened interaction between the teachers, the students and the subject. The role of the teacher in the classroom is to facilitate the interaction between the students and the subject matter. A failure in learning is tantamount to a failure in this process. Do we have any evidence of that occurrence? First, in the absence of a direct measure of teacher effort, we study teachers'

[47] However, we cannot reject that the coefficients for each group by treatment arms are jointly equal.
[48] The p-value of the joint Chow test is 0.1. We also found that there are no differences in the treatment effects between males and females. Results are available from the authors upon request.

attitudes towards innovations and the quality of the interactions they have with their students. Second, we analyze how student effort was affected by the intervention.

In Table 8, we look at the attitudes of teachers in the new environment. We start by analyzing three scales. Access to new ideas aims at measuring how much professional development and feedback or discussions about new teaching strategies each teacher had recently. Innovation measures whether teachers in the school are willing to innovate in their daily teaching practices. Reflective dialogue captures how much discussion exists among teachers of the school regarding the curriculum and general goals. We find positive effects but only at the margin of statistical significance, even after aggregating these variables in a single scale (see, for instance, Kling, Liebman and Katz, 2007).[49]

We also analyze measures of the quality of the interactions between teachers and students. The scale, built using class observations, recorded whether teachers maintained order in the class, offered students clear instructions, and properly answered students' questions. A second related scale, teaching efficacy (built using surveys), measures whether teachers exposed to the new curriculum felt less in control of the class and of the learning experience of their students. We find overall a negative impact of the treatment on these two scales, particularly when we combined them into one measure.[50]

As further evidence of the deterioration of the teacher-student interaction we use teacher logs to track how classes progressed through the syllabus during the trimester.[51] In Figure 4, we show for every unit of the syllabus the proportion of teachers that have completed them at three different points in the calendar: June, July and August. As can be seen, the control group progressed significantly faster than the treatment arms, with no discernible difference between treatment groups. Although we have shown in the robustness analysis that this slower progression cannot explain the negative test results,[52] it does raise the point that there might have

[49] We estimate equation (1) by a set of seemingly unrelated regressions for all the outcomes and use the covariance matrix to compute the standard error of the average (combined) treatment effect.
[50] We found no evidence that there was heterogeneity in these treatment effects by the experience and quality of the teachers. Results are available from the authors upon request.
[51] A priori, the intervention could have speeded or delayed the completion of the syllabus. On the one hand, the pedagogical approach was new and could have slowed down the class. On the other hand, we provided structured printed material and training which should have reduced the burden of class preparation.
[52] We designed the test so as to have a heavier load of questions in the middle of the syllabus to guard us from the possibility that the treatment could slow down the delivery of material.

been significant adjustment costs and teachers in the treatment arms may have rushed over some of the material to catch up.

We next analyze how the intervention affected student effort. We do not observe student effort directly so, instead, we look at scales for bad behavior, academic press, avoidance of novelty, academic engagement, and preference for math. In the first five rows of Table 9, we estimate equation (1) on each separate scale and confirm that students' behavior deteriorated, they were less willing to experience with new learning strategies, they were more disengaged from the class, they were less pressed to exert effort, and they liked math less. Then we estimate the average treatment effect on all five outcomes combined. The estimates are overall negative but insignificant.

In Figure 5, we show local polynomial regressions for these five outcomes scales (controlling for strata fixed effects) on student pre-treatment SAT. The index of bad behavior (e.g., "sometimes I bother my teacher during class"), falls monotonically with SAT for the control group. In fact, it does fall at a faster pace in the control than in any of the treatment arms. Similarly, academic press (e.g., "the teacher expects everyone to work hard") increases monotonically with SAT for the control group but is flat or concave for the treatment arms. The behavior of the other outcomes is similar and highlights that students with higher pre-treatment SAT were disproportionally disengaged from the class.[53]

The similar heterogeneity patterns in learning outcomes and behavioral responses highlighted in this section are reassuring. So, why did the better students fail to engage? One possible interpretation is that they were better equipped to learn under the old regime. Therefore, they exerted more effort, felt more engaged, behaved better, and ultimately learned more. A second explanation is that the intervention provided new opportunities for students to get distracted. Indeed, if we separate the technologies we find that the strongest negative results are for the high ability students in the one-to-one schools.[54] A third possible explanation is that status quo teaching was geared toward high-ability students and that a structured curriculum affected teachers' ability to pursue this strategy.

[53] We separated the sample according to the pre-treatment SAT and found that high ability students were less engaged than students in the control group. Using a Chow test we reject the null that the coefficients are equal for high and low ability students.

[54] Results are available upon request.

8. Conclusion

In this paper we report the results of an experiment with seventh grade Costa Rican children designed to improve their ability to think, reason, argument, and communicate using mathematics. We created a structured pedagogical intervention that allowed students the opportunity for a more active role in the classroom. The intervention blends a modern curricular approach with technology for teaching geometry.

We randomly assigned 85 participating schools to treatment and control groups. All students (18,000) and teachers (190) in the seventh grade of these schools participated in the experiment. Treatment schools received the active learning intervention. In addition, in order to assess the role of technology keeping constant the pedagogical approach, we randomized treatment schools to receive no technology, an interactive whiteboard, a computer lab, or a laptop for every child in the classroom.

We find that the control group learned significantly more than any of the four intervention groups. The students using only the active learning approach learned about 17 percent less than the status quo. The loss in the group that also received technology was 25 percent of a standard deviation. We find that the best students were harmed the most by this intervention. Concurrently, their behavior deteriorated and they were less engaged with learning mathematics. The evidence suggests that teachers went through the motions as prescribed but did not master the innovation in a way that would have allowed students to benefit the most from it.

We can rule out several nuisance interpretations and explanations of these findings. First, classroom material was designed by a team of recognized local experts advised by a group of international specialists for a non-negligible portion of the seventh grade curriculum. Moreover, it was aligned with current curricular reforms in the country. Therefore, the experiment sheds light on a salient and significant educational policy.

Second, the clustered randomized design ensures neither schools, nor teachers, nor students could have selected into the treatment. Furthermore, the fact that all teachers and all students participated in the experiment rules out other sources of possible biases. Indeed we showed that the experiment had perfect compliance, was internally valid, and was implemented in a large representative sample of schools. That is, this is not a result of a small experiment on a bizarre sample.

Third, there were very high levels of teacher participation in training where the material was tried and validated by teachers before the intervention started. We interpret high take-up rates of the materials/equipment and the changes in class dynamics as suggestive that the resources were deemed useful for classroom use and that teachers bought into the changes we proposed.

Fourth, we use a psychometrically valid test which was designed to measure not only knowledge of basic concepts but also higher order skills (that we expected the intervention would foster). We find that the treatment groups performed worse than the control in both learning dimensions. We also present evidence that the results are robust to redefining the test by leaving out certain syllabus sections or items of different difficulty level. It is also reassuring that the heterogeneity observed in learning is consistent with the heterogeneity in student behavior, effort, and engagement.

To conclude, these results are not at odds with the possibility that with more training, fine-tuned materials, and the benefits of learning by doing active learning with blended technology may lead to significant improvements in mathematical competence. As we have shown, however, educational reform may entail sizeable costs in the short run. This implies that policy makers should monitor carefully the performance of the educational system during reforms and consider compensatory programs.

References

Aaronson, D., L. Barrow and W. Sander. 2007. “Teachers and Student Achievement in the Chicago Public High Schools.” *Journal of Labor Economics* 25: 95–135.

Albornoz, F., S. Berlinski and A. Cabrales. 2010. “Incentives, Resources and the Organization of the School System.” CEPR Discussion Paper 7964. London, United Kingdom: Centre for Economic Policy Research.

Angrist, J., and V. Lavy. 2002. “New Evidence on Classroom Computers and Pupil Learning.” *Economic Journal* 112: 735-765.

Arias, F., and M. Zuñiga. 2012a. “Desarrollo de Curriculum y Materiales para el Aula del Proyecto de Enseñanza de Geometría en el Séptimo Grado en Costa Rica.” San Jose, Costa Rica: Fundación Omar Dengo.

Arias, F., and M. Zuñiga. 2012b. “La Capacitación de Profesores en el Proyecto de Enseñanza de Geometría en el Séptimo Grado en Costa Rica.” San Jose, Costa Rica: Fundación Omar Dengo.

Berlinski, S. et al. 2011. “Computers in Schools: Why Governments Should Do Their Homework.” In: A. Chong, editor. *Development Connections: Unveiling the impact of New Information Technologies*. New York, United States: Palgrave Macmillan.

Borkum, E., F. He and L. Linden. 2012. “School Libraries and Language Skills in Indian Primary Schools: A Randomized Evaluation of the Akshara Library Program.” NBER Working Paper 18183. Cambridge, United States: National Bureau of Economic Research.

Chassang, S., G. Padro i Miquel and E. Snowberg. 2012. “Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments.” *American Economic Review* 102: 1279–1309

Cheung, A., and R. Slavin. 2011. *The Effectiveness of Educational Technology Applications for Enhancing Mathematics Achievement in K-12 Classrooms: A Meta-analysis*. Baltimore, United States: Johns Hopkins University, Center for Research and Reform in Education.

Cristia, J. et al. 2012. “Technology and Child Development: Evidence from the One Laptop per Child Program.” Working Paper IDB-WP 304. Washington, DC, United States: Inter-American Development Bank.

David, P. 1990. “The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox.” *American Economic Review* 80: 355-361.

Dobbie, W., and R. Fryer. 2013. “Getting beneath the Veil of Effective Schools: Evidence from New York City.” *American Economic Journal: Applied Economics* 5(4): 28-60.

Eurydice. 2011. *Mathematics Education in Europe: Common Challenges and National Policies*. Brussels, Belgium: European Commission, Education, Audiovisual and Culture Executive Agency. Available at: http://eacea.ec.europa.eu/education/eurydice.

Fryer, R. 2014. “Injecting Successful Charter School Strategies into Traditional Public Schools: Evidence from Field Experiments.” Forthcoming in *Quarterly Journal of Economics*.

Gersten, R. et al. 2008. “Report of the Task Group on Instructional Practices.” In: *Foundations for Success: Report of the National Mathematics Advisory Panel*. Washington, DC, United States: United States Department of Education. Available at: http://www2.ed.gov/about/bdscomm/list/mathpanel/report/instructional-practices.pdf

Glewwe, P., M. Kremer and S. Moulin. 2009. “Many Children Left Behind? Textbooks and Test Scores in Kenya.” *American Economic Journal: Applied Economics* 1: 112-135.

Glewwe, P. et al. 2004. “Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya.” *Journal of Development Economics* 74: 251-268.

Harris, D., and T. Sass. 2011. “Teacher Training, Teacher Quality and Student Achievement.” *Journal of Public Economics* 95: 798–812

Hedges, L., and E.C. Hedberg. 2007. “Intraclass Correlation Values for Planning Group-Randomized Trials in Education.” *Educational Evaluation and Policy Analysis* 29: 60-87.

Helpman, E., and A. Rangel. 1999. “Adjusting to a New Technology: Experience and Training.” *Journal of Economic Growth* 4: 359–383.

Kane, T. et al. 2010. “Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project.” MET Project Research Paper. Seattle, United States: Bill & Melinda Gates Foundation. Available at: http://www.dartmouth.edu/~dstaiger/Papers/2010/Learning About Teaching MET Project 2010.pdf.

Kane, T. et al. 2012. “Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.” MET Project Research Paper. Seattle, United States: Bill & Melinda Gates Foundation. Available at: http://www.dartmouth.edu/~dstaiger/Papers/2012/MET_Gathering Feedback_Research Paper.pdf.

Karlan, D., R. Knight and C. Udry. 2012. "Hoping to Win, Expected to Lose: Theory and Lessons on Micro Enterprise Development." NBER Working Paper 18325. Cambridge, United States: National Bureau of Economic Research.

Kennedy, M.M. 1999. "Approximations to Indicators of Student Outcomes." *Educational Evaluation and Policy Analysis* 21: 345-363.

Kling, J.R., J.B. Liebman and L.F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica* 75: 83–119.

Lara-Meloy, T. et al. 2012. "Design and Development of the Student Assessment Instruments for Active Learning in Seventh Grade Geometry in Costa Rica." Mimeographed document.

Liu, X. 2003. "Statistical Power and Optimum Sample Allocation Ratio for Treatment and Control Having Unequal Costs per Unit of Randomization." *Journal of Educational and Behavioral Statistics* 28(3): 231-248.

Machin, S., and S. McNally. 2008. "The Literacy Hour." *Journal of Public Economics* 9: 1441–1462.

McEwan, P. 2013. "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments." Forthcoming in *Review of Educational Research*. Available online at: http://academics.wellesley.edu/Economics/mcewan/PDF/meta.pdf

Nam, J. 1973. "Optimum Sample Sizes for the Comparison of the Control and Treatment." *Biometrics* 29(1): 101-108.

Organisation for Economic Co-operation and Development (OECD). 2009. *Learning Mathematics for Life: A Perspective from PISA*." Paris, France: OECD. Available at: http://browse.oecdbookshop.org/oecd/pdfs/free/9809111E.PDF

Shechtman, N. et al. 2010. "Design and Development of the Student and Teacher Mathematical Assessments." Technical Report 05. Menlo Park, United States: SRI International. Available at: http://math.sri.com/publications/Simcalc_TechReport_05.pdf

Zuñiga, M. 2003. "Aprendizaje Mediado por Tecnologías Digitales: La Experiencia de Costa Rica." In: *Educación y Nuevas Tecnologías: Experiencias en América Latina*. Buenos Aires, Argentina: Instituto International de Planeamiento de la Educación / United Nations Economic, Social and Cultural Organization. Available at: http://unesdoc.unesco.org/images/0014/001423/142319s.pdf

Tables and Figures

Table 1: Background and Sample Comparison
(mean characteristics)

	Schools in Sample	Costa Rica (restricted)	Costa Rica
	[1]	[2]	[3]
Enrollment			
Students per school (2011)	736	601	408
Log (change students per school) 2007-2011	3%	4%	5%
Students in 7th grade (2011)	228	188	124
Log (change in 7th grade) 2007-2011	1%	2%	1%
Demographics (2011)			
Percent of female students (school)	50%	51%	50%
Percent of female students (7th grade)	48%	48%	47%
Average age (school)	14.5	14.6	14.8
Average age (7th grade)	13.1	13.1	13.1
Infrastructure (2011)			
Percent of schools with library	74%	73%	56%
Percent of schools with restrooms	62%	65%	61%
Average number of classrooms	20.0	17.1	12.4
Average number of computers	32.0	29.6	22.8
Number of Schools (2011)	85	397	773

Note: Column [1] shows means for schools in the sample which are located in Alajuela, Cartago, Desamparados, Heredia, Occidente, Puriscal, San Jose (Central, Norte) and San Ramon. Column [2] is restricted to schools in the country that satisfy the experiment elegibility criteria. In column [3] we show the average for Costa Rica.

Table 2: Non-Response Rates

	Sample mean and [s.d.]	OLS coefficients and		p-value joint test coeffs equal to zero	Sample Size
		Active Learning	Active Learning &Technology		
	[1]	[2]	[3]	[4]	[5]
Student Level Variables					
Missing on Geo test day	0.091	-0.017	0.008	0.461	4625
	[0.288]	[0.024]	[0.018]		
Geo test date (# days after end of geo unit)	6	1.813	-0.323	0.425	4157
	[6.489]	[1.971]	[1.815]		
Missing SAT (among eligible students)	0.211	-0.027	-0.098	0.215	4157
	[0.408]	[0.091]	[0.070]		
Student with disability (did not take geo test)	0.011	-0.010	-0.017	0.233	4881
	[0.103]	[0.012]	[0.011]		
Teacher Level Variables					
Missing teacher survey (baseline)	0.005	-0.025	-0.021	0.396	190
	[0.073]	[0.019]	[0.015]		
Missing teacher survey (endline)	0.032	0.003	-0.012	0.865	190
	[0.175]	[0.035]	[0.031]		
Missing class observation	0.195	0.027	-0.059	0.444	190
	[0.397]	[0.095]	[0.072]		
Missing teacher log June	0.111	-0.022	-0.163	0.082	190
	[0.314]	[0.127]	[0.092]*		
Missing teacher log July	0.163	-0.147	-0.192	0.022	190
	[0.370]	[0.082]*	[0.068]***		
Missing teacher log August	0.237	-0.102	-0.175	0.062	190
	[0.426]	[0.101]	[0.073]**		

Note: Each row shows statistics for a different variable Y_{isj} of individual (student, teacher or school) i, in strata s and in school j. Column [1] shows the sample average and the standard deviation in square brackets. Columns [2]-[3] show the regression coefficients and the standard errors in square brackets corresponding to equation (1), a regression model that only includes controls for strata. Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. Column [4] shows the p-value of a test of all coefficients jointly equal to zero. Column [5] shows the sample size.

Table 3: Differences in Pre-Treatment Means

	Sample mean and [s.d.]	OLS coefficients and [s.e.]		p-value joint test coeffs equal to	Sample Size
		Active Learning	Active Learning &Technology		
	[1]	[2]	[3]	[4]	[5]
Student Level Variables					
Percent Male	0.489	-0.029	-0.018	0.300	4157
	[0.500]	[0.019]	[0.018]		
Age (years)	12.970	0.072	-0.022	0.227	4127
	[0.878]	[0.061]	[0.041]		
Mother's Education (Primary)	0.419	0.046	0.010	0.495	4106
	[0.493]	[0.044]	[0.043]		
Mother's Education (Secondary)	0.406	0.003	0.008	0.942	4106
	[0.491]	[0.025]	[0.025]		
Number of books at home	3.161	-0.085	-0.052	0.578	3560
	[1.565]	[0.083]	[0.094]		
Have a PC/laptop at home	0.735	-0.033	-0.004	0.576	3543
	[0.442]	[0.036]	[0.031]		
SAT (% Correct)	0.466	-0.019	-0.008	0.496	3278
	[0.145]	[0.017]	[0.017]		
Teacher Level Variables					
Percent Male	0.486	0.029	0.146	0.243	185
	[0.501]	[0.127]	[0.102]		
Age (years)	36.668	0.853	0.104	0.795	184
	[7.772]	[1.385]	[1.122]		
Experience (years)	11.652	0.500	0.400	0.895	184
	[6.543]	[1.251]	[0.950]		
School Level Variables					
Students 7th Grade	219.694	-0.650	-1.065	0.993	85
	[114.174]	[16.949]	[8.923]		
Classes 7th Grade	6.847	-0.000	-0.194	0.710	85
	[3.053]	[0.380]	[0.259]		
Math teachers 7th grade	2.235	0.100	0.044	0.948	85
	[1.221]	[0.355]	[0.331]		
Computer Lab	0.741	0.000	-0.017	0.986	85
	[0.441]	[0.148]	[0.124]		
Internet in School	0.729	0.150	-0.010	0.296	85
	[0.447]	[0.136]	[0.129]		
7th Grade Repetition	0.087	-0.018	-0.011	0.644	85
	[0.062]	[0.020]	[0.016]		
Not Urban	0.447	-0.050	-0.068	0.852	85
	[0.500]	[0.148]	[0.121]		

Note: Each row shows statistics for a different variable Yisj of individual (student, teacher or school) i, in strata s and in school j. Column [1] shows the sample average and the standard deviation in square brackets. Columns [2]-[3] show the regression coefficients and the standard errors in square brackets corresponding to equation (1), a regression model that only includes controls for strata. Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. Column [4] shows the p-value of a test of all coefficients jointly equal to zero. Column [5] shows the sample size.

Table 4: Gaming

	Sample mean and [s.d.]	OLS coefficients and		p-value joint test coeffs equal to	Sample Size
		Active Learning	Active Learning &Technology		
	[1]	[2]	[3]	[4]	[5]
Learned teaching assignment before lottery	0.837	-0.106	-0.054	0.398	190
	[0.370]	[0.079]	[0.064]		
Class learned geometry 1st Term	0.016	-0.020	-0.041	0.386	190
	[0.125]	[0.050]	[0.044]		
Class learned 4 geo units in 1st Term	0.126	0.066	-0.080	0.135	190
	[0.333]	[0.122]	[0.094]		

Note: Each row shows statistics for a different variable Yisj of individual (student, teacher or school) i, in strata s and in school j. Column [1] shows the sample average and the standard deviation in square brackets. Columns [2]-[3] show the regression coefficients and the standard errors in square brackets corresponding to equation (1), a regression model that only includes controls for strata. Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. Column [4] shows the p-value of a test of all coefficients jointly equal to zero. Column [5] shows the sample size.

Table 5: Use of Classroom Resources

	OLS coefficients and [s.e.]		p-value of	
	Active Learning [1]	Active Learning &Technology [2]	Wald that [1]=[2] [3]	Sample Size [4]
Access/ reported use (by students):				
Class materials	0.764	0.789	0.587	190
	[0.066]***	[0.054]***		
Some technology in class	-0.046	0.897	0.000	190
	[0.054]	[0.047]***		
Observed use:				
Class uses student's workbook	0.811	0.989	0.003	153
	[0.060]***	[0.030]***		
Class uses teacher's manual	0.855	0.966	0.064	153
	[0.055]***	[0.036]***		
Class uses Geogebra software	-0.010	0.766	0.000	153
	[0.054]	[0.059]***		
Class uses internet	0.004	0.034	0.180	153
	[0.014]	[0.022]		
Class uses regular blackboard	-0.267	-0.391	0.175	135
	[0.109]**	[0.100]***		

Note: Each row shows statistics for a different variable Yisj of individual (student, teacher or school) i, in strata s and in school j. Columns [1]-[2] show the regression coefficients and the standard errors in square brackets corresponding to equation (2), a regression model which includes strata controls, individual controls (gender, age, mom education, books, SAT), teacher controls (gender, age, experience) and school controls (# students in 7th grade, # classrooms in 7th grade, Lab in school, region dummies). Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. Column [3] shows the p-value of a Wald test that coefficients [1] = [2]. Column [4] shows the sample size.

Table 6: Time Allocation and Class Dynamics

	OLS coefficients and [s.e.]		p-value of Wald that [1]=[2]	Sample Size
	Active Learning [1]	Active Learning &Technology [2]	[3]	[4]
Share of class time devoted to:				
Exploration	0.310 [0.080]***	0.452 [0.065]***	0.029	153
Formalization	-0.102 [0.041]**	-0.063 [0.043]	0.127	153
Practice	-0.208 [0.094]**	-0.389 [0.076]***	0.013	153
Class plenary lecture	-0.064 [0.037]*	-0.055 [0.033]*	0.754	153
Class discussion	0.117 [0.058]**	0.168 [0.055]***	0.308	153
Work in groups	0.010 [0.043]	-0.054 [0.035]	0.070	153
Work in pairs	0.010 [0.032]	0.004 [0.027]	0.870	153
Work individually	-0.073 [0.059]	-0.062 [0.060]	0.912	153
Active engagement in class	0.028 [0.047]	0.079 [0.034]**	0.210	4052
Classroom activity	0.121 [0.044]***	0.166 [0.038]***	0.292	4157
Math prescribed learning practices (Student)	0.300 [0.253]	0.602 [0.207]***	0.121	153
Math prescribed teaching practices (Teacher)	0.362 [0.234]	0.513 [0.201]**	0.414	153

Note: Each row shows statistics for a different variable Yisj of individual (student, teacher or school) i, in strata s and in school j. Columns [1]-[2] show the regression coefficients and the standard errors in square brackets corresponding to equation (2), a regression model which includes strata controls, individual controls (gender, age, mom education, books, SAT), teacher controls (gender, age, experience) and school controls (# students in 7th grade, # classrooms in 7th grade, Lab in school, region dummies). Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. Column [3] shows the p-value of a Wald test that coefficients [1] = [2]. Column [4] shows the sample size.

Table 7: Geometry Test Results

		OLS coefficients and [s.e.]		p-value of Wald that [1]=[2]	Sample Size
		Active Learning [1]	Active Learning &Technology [2]	[3]	[4]
Geometry score		-0.171 [0.080]**	-0.247 [0.081]***	0.272	4157
Geometry score (Basic skills)		-0.142 [0.079]*	-0.209 [0.080]***	0.286	4157
Geometry score (Higher-order skills)		-0.126 [0.054]**	-0.204 [0.055]***	0.122	4157
Dependent Variable: Geometry Score					
Teacher:	(A) Low experience	-0.317 [0.136]**	-0.312 [0.126]**	0.929	2182
	(B) High experience	0.001 [0.110]	-0.152 [0.067]**	0.172	1862
	(C) Low quality	-0.147 [0.066]**	-0.216 [0.063]***	0.148	1929
	(D) High quality	-0.139 [0.123]	-0.246 [0.079]***	0.159	1867
Student:	(E) Low skilled	-0.041 [0.080]	-0.144 [0.066]**	0.125	1658
	(F) High skilled	-0.248 [0.122]**	-0.257 [0.113]**	0.981	1620

Note: Each row shows statistics for a different variable Yisj of individual (student, teacher or school) i, in strata s and in school j. Columns [1]-[2] show the regression coefficients and the standard errors in square brackets corresponding to equation (2), a regression model which includes strata controls, individual controls (gender, age, mom education, books, SAT), teacher controls (gender, age, experience) and school controls (# students in 7th grade, # classrooms in 7th grade, Lab in school, region dummies). Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. Column [3] shows the p-value of a Wald test that coefficients [1] = [2]. Column [4] shows the sample size.
Samples of low/high ability students, low/high ability teachers, low/high quality teachers are described in Section 7.

Table 8: Teachers Attitudes and Quality of Interactions

	OLS coefficients and [s.e.]		p-value of Wald that [1]=[2]	Sample Size
	Active Learning [1]	Active Learning &Technology [2]	[3]	[4]
(A) Access to new ideas	0.187 [0.262]	0.374 [0.199]*	0.356	184
(B) Innovation	0.232 [0.220]	0.076 [0.171]	0.423	184
(C) Reflective dialogue	0.302 [0.212]	0.417 [0.197]**	0.453	185
(D) Quality of teacher-students interactions	-0.840 [0.384]**	-0.651 [0.256]**	0.672	153
(E) Teaching efficacy	-0.198 [0.178]	-0.213 [0.162]	0.946	187
Teacher Innovation Scale (A+B+C)	0.241 [0.165]	0.289 [0.142]**	0.139	184
Teacher Mediation Scale (D+E)	-0.519 [0.208]**	-0.432 [0.154]***	0.013	153

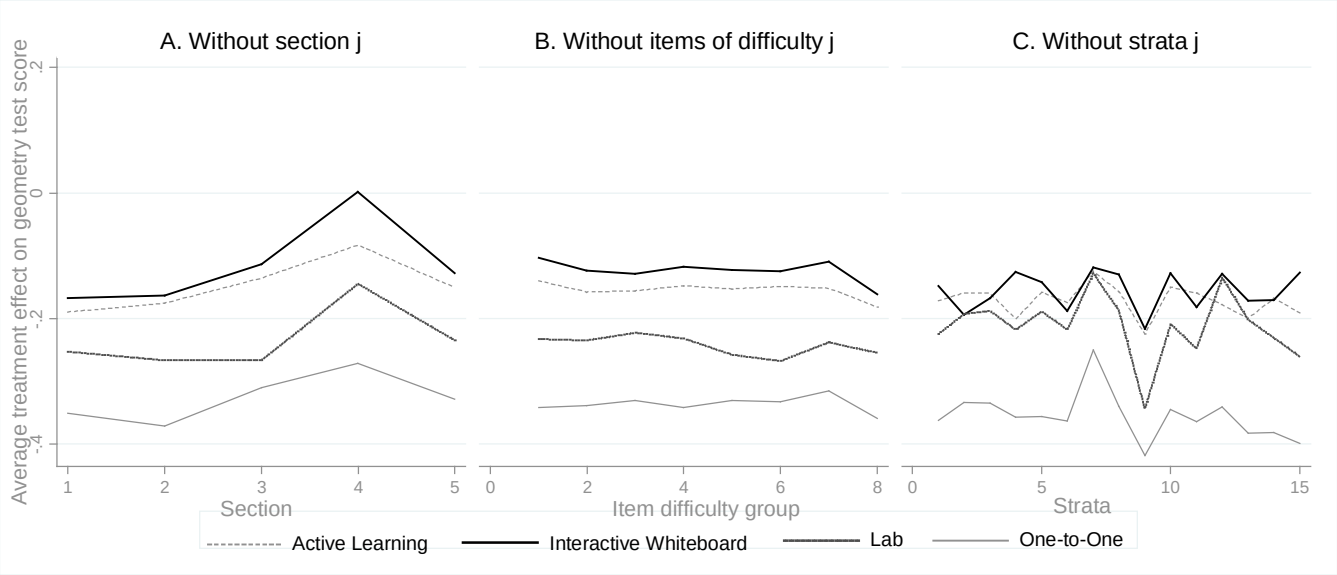
Note: Each row shows statistics for a different variable Yisj of individual (student, teacher or school) i, in strata s and in school j. Columns [1]-[2] show the regression coefficients and the standard errors in square brackets corresponding to equation (2), a regression model which includes strata controls, individual controls (gender, age, mom education, books, SAT), teacher controls (gender, age, experience) and school controls (# students in 7th grade, # classrooms in 7th grade, Lab in school, region dummies). Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. Column [3] shows the p-value of a Wald test that coefficients [1] = [2]. Column [4] shows the sample size.

Table 9: Student Effort

	OLS coefficients and [s.e.]		p-value of Wald that [1]=[2]	Sample Size
	Active Learning	Active Learning &Technology		
	[1]	[2]	[3]	[4]
(A) Bad behavior	0.089	0.071	0.619	4030
	[0.056]	[0.054]		
(B) Avoid novelty	0.072	0.085	0.781	3943
	[0.053]	[0.048]*		
(C) Academic engagement	-0.040	0.015	0.353	3973
	[0.075]	[0.066]		
(D) Academic press	-0.011	-0.033	0.597	3917
	[0.048]	[0.039]		
(E) Preference for math	-0.140	-0.055	0.162	3970
	[0.077]*	[0.059]		
Student Combined Scale (-A-B+C+D+E)	-0.070	-0.046	0.081	3970
	[0.041]*	[0.038]		

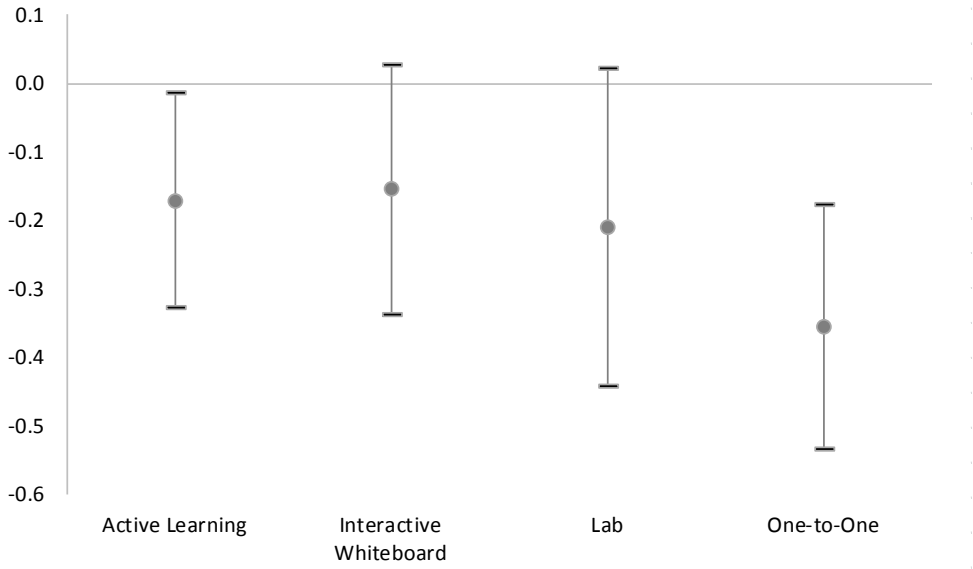
Note: Each row shows statistics for a different variable Yisj of individual (student, teacher or school) i, in strata s and in school j. Columns [1]-[2] show the regression coefficients and the standard errors in square brackets corresponding to equation (2), a regression model which includes strata controls, individual controls (gender, age, mom education, books, SAT), teacher controls (gender, age, experience) and school controls (# students in 7th grade, # classrooms in 7th grade, Lab in school, region dummies). Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. Column [3] shows the p-value of a Wald test that coefficients [1] = [2]. Column [4] shows the sample size.

Figure 1: Robustness of results on scores



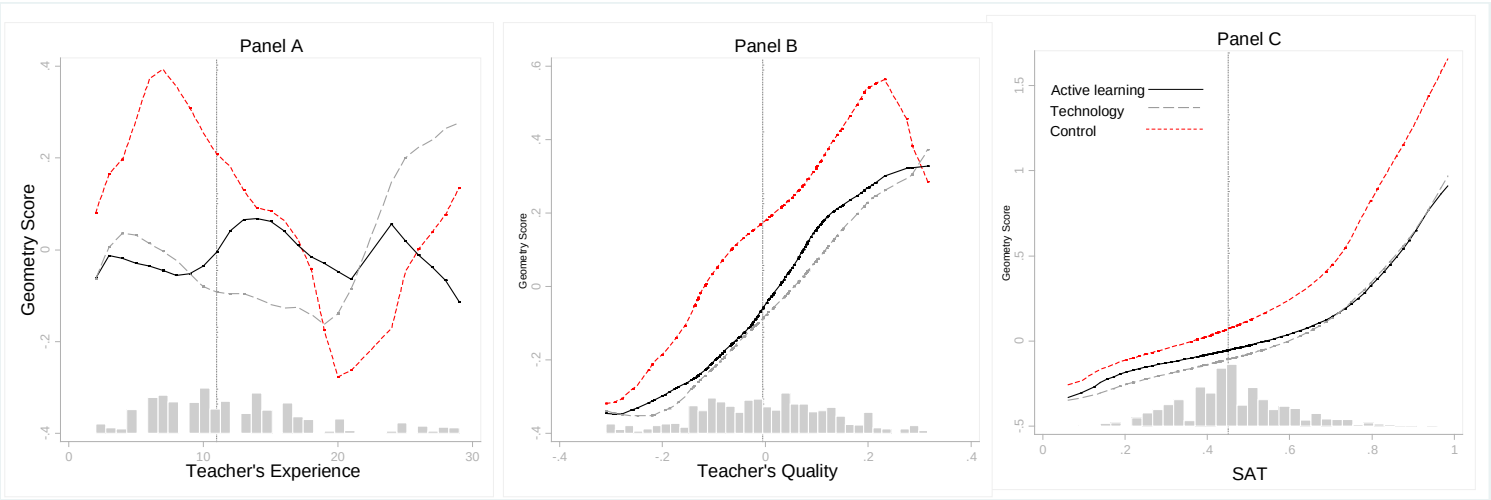
Note: The y-axis shows the treatment effect of a standardized geometry test score on treatment dummies estimated following equation (2). Panel A shows estimates obtained by removing items that belong to one (syllabus) section at a time. Panel B shows estimates obtained by removing items of one difficulty group at a time. Panel C shows estimates obtained by removing schools in one strata at a time.

Figure 2: Geometry Test Result by Technology



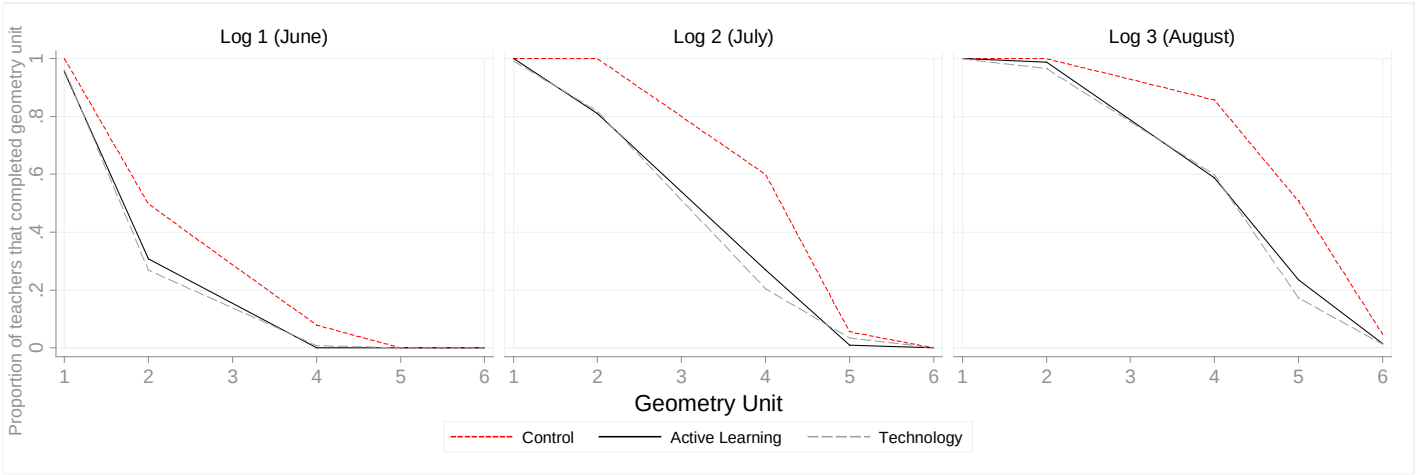
Note: The middle dot is the OLS coefficient estimated following equation (2), the vertical bars are 95% confidence interval.

Figure 3: Treatment Effect Heterogeneity (Geometry Score)



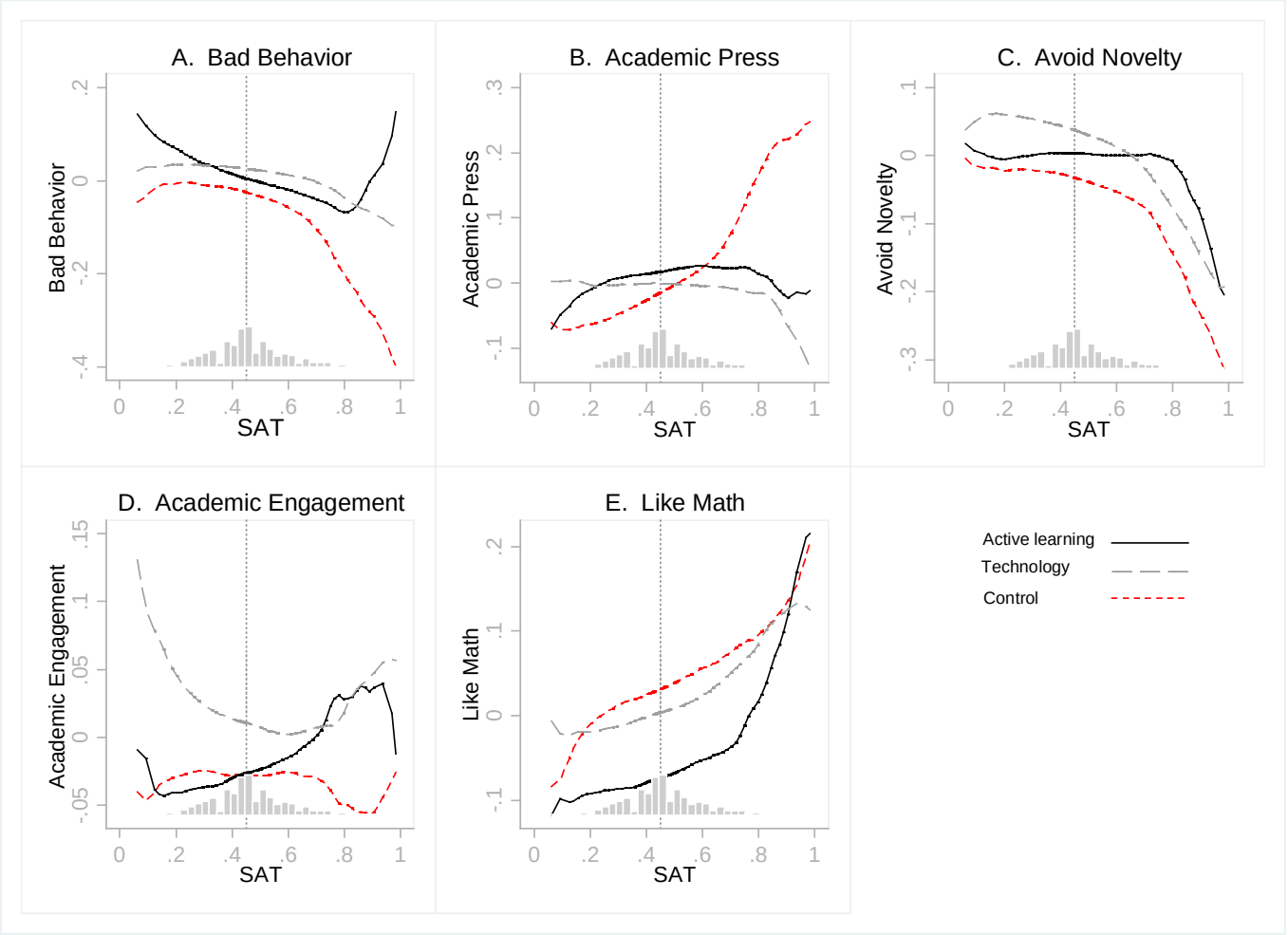
Note: Each line presents a local polynomial regression of the geometry test-scores (y-axis) --controlling for strata fixed effects-- on a mediating variable (x-axis): teacher experience (panel A), teacher quality (panel B), and student pre-treatment SAT (panel C). The red dashed line is for the control group, the black solid line is for those students in the active learning condition, and the grey long-dashed line is for those students in the three technology groups. At the bottom of the graph we overlap a histogram of the mediating variable and the vertical line marks the median of the mediating variable distribution (to save space we do not plot the frequency on any axis). The local polynomial regressions were estimated using an Epanechnikov with a bandwidth of 0.15 (panel A), 2 (panel B) and 0.10 (panel C).

Figure 4: Geometry Unit Progression



Note: The y-axis shows the proportion of teachers that completed a given geometry unit (x-axis). Each panel shows this for a different teacher log and point in the calendar (June, July and August).

Figure 5: Treatment Effect Heterogeneity (Students Scales)



Note: Each line presents a local polynomial regression of a student scale (y-axis) --controlling for strata fixed effects-- on student pre-treatment SAT (x-axis). The red dashed line is for the control group, the black solid line is for those students in the active learning condition, and the grey long-dashed line is for those students in the three technology groups. At the bottom of the graph we overlap a histogram of the mediating variable and the vertical line marks the median of the mediating variable distribution (to save space we do not plot the frequency on any axis). The local polynomial regressions were estimated using an Epanechnikov with a bandwidth of 0.15.