Pedagogical Change in Mathematics Teaching: Evidence from a Randomized Control Trial[1]

Samuel Berlinski
Inter-American Development Bank

Matias Busso
Inter-American Development Bank

July 2013

Abstract

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. Our multi-treatment experiment also explores a gradient of technologies that are complements to this pedagogical approach and vary widely in cost. We find that the control group learned significantly more than any of our treatment arms. Moreover, technology had a rather negative impact at achieving our objective. We provide evidence that the experiment was internally valid and implemented with high fidelity.

Preliminary –Please, do not cite or circulate

JEL classification: C93, I21, I28, O32

Keywords: Education, Technology, Curricular Reform, Teacher Training, Laptops, Interactive whiteboards, Computer Labs, Mathematics, Field Experiments.

[1] Acknowledgements: 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, Maria Eugenia Bujanda, Elsie Campos, Marco Cordero, Ulises Cordero, Julian Cristia, Maria Antonieta Diaz, Alvaro Gamboa, Torie Gorges, Mauricio Holtz, Teresa Lara-Meloy, Richard Mayer, Jeremy Roschelle, and Magaly Zuñiga. We also thank seminar participants at CEP (LSE), EDePo (IFS), Fundación Omar Dengo, IDB and Queen Mary - University of London for helpful comments. 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. 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. This is to say, the student 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 US.

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, p. 56).

We created a pedagogical intervention designed to give students a more active role in the classroom. A key aspect of this change relies on providing them with guided opportunities to explore and discover. In mathematics, a potentially important lever in this process is the use of technology. Our experiment explores a gradient of technologies that are complements to this pedagogical approach and vary widely in cost.

We designed and implemented an experiment with seventh grade students in Costa Rica that blends a modern curricular approach with technology for teaching geometry (one of three units of the seventh grade program or about three months of teaching). We randomly assigned the 85 participant schools in this experiment to one of five conditions: (1) status-quo (i.e., control); (2) new curriculum design; (3) new curriculum design and an interactive whiteboard; (4) new curriculum design and a computer lab; (5) new curriculum design and a laptop for every child in the classroom. All students (18,000) and teachers (190) in the seventh grade of these schools participated in the experiment.

The technologies we chose for this experiment represent a set of relevant options for middle and high income countries and vary widely in their costs. For example, in our case the cost of having one laptop for every student in the mathematics classroom is more than three times larger than the cost of equipping a classroom with an interactive whiteboard. 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).

Lampert (1990) summarizes the recommendations of the US National Councils 1989 mathematics reform documents as follow[2]: "Mathematics students should be making conjectures, abstracting mathematical properties, explaining their reasoning, validating their assertions, and discussing and questioning their own thinking and the thinking of others" (p. 33). Concrete manipulatives are a natural pedagogical device to facilitate exploration and argumentation in the mathematics classroom. In fact, their use has a long tradition and some evidence of success (National Research Council, 2001). Technology can play an important role in helping students to manipulate figures in order to make conjectures, explain their reasoning and interact with others (Clements, 1999).

We commissioned the design of pedagogical material for this project to local experts supervised by a team from a leading international education academic organization. The team decided that given the objective of achieving mathematical competence, traditional classroom routines had to be changed. In particular, the idea was to include a guided exploration of the topics before a more directed formulation of concepts, rather than the more typical lecture on the concepts, followed by an opportunity to practice on typical problems. In the guided exploration students were asked to informally conjecture about relationships, definitions and concepts. Exactly what this part looked like varied by condition.

In order to support teachers and students in this transition and with an eye directed at improving fidelity of implementation as well, we created a teachers' manual and a students' workbook (one for every modality of intervention but none for the control). Technology was introduced through a set of applets created on an open source dynamic mathematics software designed to help students and teachers explore the key concepts of the unit.

In the one-to-one and computer lab conditions, students directly manipulated the software applications. In the interactive whiteboard condition, students shared a dynamic geometry experience on the whiteboard and the teacher determined whether students had direct access to the technology. In the non-technology condition, students used the guided exploration portion of the class to conjecture with hands-on manipulatives.

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

Collaborating with the 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

[2] Cited in Gersten et al. (2008) page 6-15.

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 the Latin American Laboratory for Assessment of the Quality of Education (LLECE) 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 class 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 during 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 four intervention groups. The students using the new curriculum without technology learned about 17 percent less than the status-quo. Learning was around 36 percent of a standard deviation lower in one laptop per student schools in comparison to control establishments. In the race between the three technologies (i.e., keeping the pedagogical approach constant) the interactive whiteboard is the one that fairs better. 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 get the most of it.

The experiment was a salient and significant 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. Therefore, the results of the experiment are not a fluke. But then, what is the long-run prospect for such an intervention? We speculate that either learning by doing (possibly complemented with more training) will eventually lead to positive treatment effects or that many of these teachers will discover that they might not be well suited well suited for an active learning classroom strategy and give-up the approach. Only time and more research will tell.

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[3]. Fryer (2012) 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 by improving the quality of teaching through more focused literacy work. 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 class observations and student and teachers 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 et al., 2009), libraries (Borkum et al., 2012) and even laptop computers (Cristia et al., 2012). The impacts of these interventions are in the best case scenario modest. 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 of 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 policy recommendations of any of the practices addressed." (Gersten (2008), page 6-189). Among the practices evaluated the panel looked at the use of teacher center versus student center approaches and the use of technology in the classroom.

Finally, a recent authoritative meta-analysis (Cheung and Slavin (2011)) of high-quality studies of educational interventions with a focus on mathematics learning in K-12 that involves the use technology helps put our study in context. The quality of a study in this review is related to the rule assigning units to treatment/control status, likeness of treatment and control groups, attrition, outcome of interest, duration, number of participants units and scalability.[4] To our knowledge,

[3] Angrist et al (2013) perform a similar analysis for Boston charter schools.

[4] For example, Cheung and Slavin (2011) include only studies where: (1) Assignment to treatment units is random or there is matching with appropriate adjustments for any pretest differences; (2) Pretest data is available or there are at least 30 units assigned randomly to treatment and controls with no indication of initial inequalities in basic demographics; (3) The outcome of interest is a quantitative measure of mathematics knowledge. Measures made by the experimenter are accepted if they are comprehensive measures of mathematics which would be fair to the control group; (4) The study last for at least 12 weeks; (5) Studies had to have at least two teachers in each treatment group; (6) Programs had to be replicable in realistic school settings.

our randomized control trial is the largest and most comprehensive high-quality study on this topic performed in a developing country.[5]

We proceed as follows: in Section 2 we present the distinctive features of our research design, in Section 3 we explain the data collection process, in Section 4 we discuss the empirical strategy, in Section 5 we present our research sample and discuss the internal validity of the experiment, in Section 6 we show our results on fidelity of implementation, test scores and class dynamics as well as some robustness checks. Section 7 discusses the possible mechanisms behind our results and concludes.

2. Research Design

2.1 Context

Costa Rica is a relatively small middle-income developing country. In 2011, 4,726,575 people lived in Costa Rica, GDP per capita (2005 PPP $) was $ 10,085 and the United Nations Human Development Index ranked it in the 69th place. 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 in four levels: preschool education (ages 4-6); primary education (6 years); secondary education which also has two cycles, middle school (seventh, eighth and ninth grade) and high school (10th and 11th grade). Currently, education is free and compulsory from the last year of preschool to the end of middle school.

During 2012, the school year in middle school was 196 days long. In the seventh grade, students have six[6] mathematics lessons a week of 40 minutes each (many of those arranged in contiguous blocks). The school year is divided in 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

[5] Cheung and Slavin (2011) identified 75 studies for K-12 published or unpublished in English from 1980 to 2010. Of those, only 25 were RCTs. The latter had an average of 9 schools (only 2 studies have more than 30 schools) and 780 students (the largest study has 3,136 students) participating in these experiments. None of the studies reported were performed in a developing country. There are no high-quality studies reported in Cheung and Slavin (2011) pertaining to the use of interactive whiteboards. The study almost exclusively finds papers that look at the use of technology as a supplement to teaching in the form of drills and practice. It reports modest gains of around 10 percent of a standard deviation for this use of technology.

[6] One of these lessons is assigned for revision, recovery of lost time and to support students that are lagging behind.

procedure in the blackboard using a particular example. Students take notes, ask questions and practice what the teacher explained. To support the teacher in this endeavor he/she follows a commercial textbook of his/her choice and provides the students a long list of examples to practice.

The theorems and procedures are taken as given truths and the objective is to practice until the students achieve mastery of their use. The problems and procedures are not necessarily associated neither to the solution of a hypothetical, real life, or scientific problem.

The interaction between students and teacher can be characterized as follows: The teacher is responsible for explaining and clarifying doubts or difficulties the students have, he/she controls the beginning of the class, the pertinence and formalization of the students' interventions, and validates students' reasoning, explanations or interventions. Students listen to their teachers interventions, participate when they don't understand, and contribute with fragmented phrases to statements made by their teachers.

Costa Rica has a long tradition of introducing technology in schools (Zuñiga, 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 services students in preschool, primary school and middle school (since 2001). Technology is introduced in the school through computer laboratories[7] with the objective (among others) of promoting the development of logic thinking by using computers to solve problems and work in teams. The program is not intended as a complement for teaching core subjects.

2.2 Experimental design

We started with a sample of 100 secondary schools with a minimum of 2 classes and a maximum of 12 classes of seventh grade math in 2011. In order to minimize implementation costs, we choose schools that were located in urban and semi-rural areas and were easily accessible by roads. We invited schools to participate in the experiment during November 2011 and only the 85 schools that signed agreements to participate were included in the experiment.[8]

In order to assign schools to their experimental status we performed a blocked randomization. First, we ordered schools according to seventh grade enrollment[9] in 2011 and constructed (at random) ten bins of five schools and five bins of seven schools. In every bin schools were assigned to either one of the following 5 conditions: control (20 schools), new curriculum (20 schools), new curriculum plus an interactive whiteboard (15 schools), new curriculum plus a

[7] Although some small rural schools receive equipment for classroom use.
[8]The Superior Council of Education of Costa Rica approved the experiment by resolution CSE-SG-168-2012.We obtained an IRB from Fundacion Omar Dengo.
[9]Ordering schools by enrollment was useful to reduce the cost of buying technology.

computer lab (15 schools) and new curriculum plus one computer per student (15 schools). All seventh grade teachers and students in the school were assigned to their corresponding experimental status.

We notified the government and the schools of the lottery results on February 2 (at the beginning of the school year). We organized meetings with the schools to inform them how the experiment would be implemented. During these meeting, principals received a manual with the following information: description of the project, main objectives, grade and topic to be covered, teacher training and equipment to be assigned, responsibility over the equipment during and after the project, verification of school infrastructure, logistics for the distribution of equipment, data collection, relevant information for parents and teachers, a calendar for the experiment implementation, and contact details.

In Table 1 we compare the average characteristics of the 85 schools that participated in the experiment with the characteristics of schools in Costa Rica. On average schools in the experiment tend to be slightly larger and are growing a bit more slowly. They have similar infrastructure (as measured by access to library, computers, number of classrooms and restrooms) than schools in urban and semi-rural parts of the country as well as similar demographic characteristics.

2.3 The design and production of the pedagogical material[10]

We commissioned the design of pedagogical material for this intervention to local experts from Fundacion 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.

The objective of the material is 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.

A lecture-teaching style where students are passive receivers of information does not encourage activities that are important for the construction of mathematical knowledge such as: exploration, verification, and communication of mathematical facts. These activities are important to develop the type of inductive and deductive reasoning that allows them to solve both scientific and everyday situations.

[10] This sections draws from a report prepared for this project by Arias and Zuñiga (2012a).

As part of the design of the experiment, we commissioned the production of a teacher manual as a way of accompanying teachers, reducing take-up costs, and improving the fidelity of implementation of the experiment. The manual was elaborated with a structure that would not demand significant time to study by the teachers. The material was gathered in 19 thematic sessions with themes that would be easily recognizable by teachers.[11] The sessions covered all the materials in the seventh grade Geometry curriculum of Costa Rica.

Each session had the same structure to help teachers with the use and understanding of the material. The sessions were classified in seven sections: (1) topic description, (2) specific objectives, (3) materials, (4) session length, (5) presentation, (6) body of the session, and (7) end of session.[12]

The body of the session was subdivided in 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 predict-check-explain cycle to construct a geometric concept. Ultimately, the understanding and explanation demand the use of a specific property. In contrast to the traditional lecture style, the students have a very active role in this process and the teachers a less controlling one.

During the formalization part the teacher is responsible for institutionalizing the knowledge established in the exploration section. However, this does not mean that we expect the teacher to recite or copy mathematical results on the board. In fact, we introduced a clarify-formalize-validate/verify cycle with the students actively participating from this formulation process.

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 with the exploration part and this just provides an opportunity to consolidate what has been learned.

In order to standardized the work of the students and to provide support to the teachers in implementing this pedagogical approach, we decided to create a student workbook as well. The student workbook has hands-on paper based activities and is identical in knowledge content to the teacher manual. The main difference is that the teacher manual has advice on how to proceed or motivate the students at different points in the class and this is not included in the student workbook.

[11] The teacher 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.

[12] Taking into consideration the usual disruptions to the school calendar (e.g., strikes, school functions, and teacher absences), the duration of each session assumed that the real teaching time would be four lessons per week rather than six.

Even though the experiment introduces four different interventions (curriculum, one-to-one, computer lab, and interactive whiteboard), 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 by 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 begs 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 the classroom routines for student and teachers which can demand the establishment of new rules of engagement. It may also require substantive new knowledge from the teachers.

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

In the teacher/student manual, technology is introduced in the exploration and formalization phases of the lesson. In the exploration phase, characterized by the predict-check-explain cycle, the GeoGebra applets are key to the check stage where students may explore with technology more freely. In the formalization phase, characterized by the clarify-formalize-validate cycle, the applets are used to systematically guide students through important cases or variations.

The GeoGeobra applets are the same in the three technology options; what varies is the time of exposure that the children will naturally have and their opportunities to drive the exploration. For example, when students have individual laptops they can use these to check predictions individually, while the teacher can perform the check phase in an interactive whiteboard setting.

The manual for the curriculum intervention proposes activities that use manipulatives such as images or paper rather than technology.

[13] It also simplifies considerably the production of large amount of the pedagogical material.
[14] GeoGebra (http://www.geogebra.org/) is a 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.

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 that participated from the training (about 45 days before starting the experiment). One manual was printed for each of the four conditions. All students received a student workbook with their name tag before the beginning of the experiment.[15]

2.4 The Deployment of equipment

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.[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, this means that you can use your finger or any other writing object to move, click and operate the computer. [17]

Schools assigned to the computer lab status had one laptop every two students available at least for 2 hours a week. Schools either used their existing computer laboratories or if they have no laboratory installed we created a mobile lab.[18]

The schools were inspected by an engineer after the lottery results were announced. In coordination with the principal, classrooms that fulfilled security and structural condition were chosen for the installation of the equipment. Minor adjustments (e.g., wires and sockets were changed, walls were fortified to support the whiteboard) were required in some schools and applied.

All teachers in the four treatment categories received a laptop computer that should be returned to the school at the end of the school year.

---

[15] The manual for the computer lab condition was a combination of the one-to-one and the curriculum manual as we could not predict which sessions would occur in the lab and which would occur in the classroom. The student workbook had 150 pages, 146 pages, 246 pages and 183 pages for the interactive whiteboard, one-to-one, computer lab, and curriculum conditions respectively.

[16] In schools with 6 sections of mathematics or less we equipped one classroom and for schools with seven or more sections we equipped two classrooms. Each classroom had 30 Classmates Magalhães laptops. The computers are particularly suitable for school environments: they are resistant to falls, water spills, and do not have sharp edges. They have battery life of 4 hours, 1GB RAM, and 160 Gb HDD. We installed Ubuntu (a Linux based open software) in all computers as their operating system and GeoGeobra as their main mathematics package.

[17] In schools with 6 sections of mathematics or less we equipped them with one interactive whiteboard and for schools with seven or more sections we equipped them with two. The interactive whiteboard model was an IQ-Board PS.

[18] Our mobile labs were set up with 15 laptops, one laptop cart, one LCD projector, one desktop computer, and one router.

The costs of deploying technology in any of the three forms are quite significant and vary considerably by technology. Table 2 reports three measures: the initial investment cost per student, the recurrent cost per student and the total cost of ownership (TCO) per student. Costs are calculated on the assumption that the program is scaled over the five years of secondary education.

The initial investments are capital costs for acquisitions and installation of items at the outset of the program. They are typically incurred in lump sum form. On the other hand, recurrent expenditures are monthly ongoing costs incurred over the lifetime of the project to effectively deploy the program. The total cost of one-to-one is three times higher than the interactive-board and the computer laboratory is twice as high.

2.5 Teacher training[19]

The development of mathematical competence requires a learning environment where the student plays a commanding role. In comparison to lecture-style teaching, the emphasis of classroom activity switches from transmitting information and performing numerous drills to creating opportunities for exploration, construction and mathematical argumentation.

Teaching to achieve mathematical competence requires a change in the traditional role of the Costa Rican mathematics teacher. Therefore, teacher training for this experiment aimed at affecting teachers' beliefs about how mathematics is taught and learned. 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.

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

- 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.
- Putting teachers randomized into a technology group in contact with their corresponding technology.

In line with the supporting material we designed for the experiment, we assigned in training a central role to the pedagogical use of technology rather than to the mastery of the technological instrument per se.[20] In fact, in a recent representative survey in Costa Rica (FOD-MEP 2010),

[19] This sections draws from a report prepared for this project by Arias and Zuñiga (2012b).

[20] We performed a diagnostic survey with 88 teachers in schools that receive technology in the planning phases of training. From them, 66.7% have received some training on the use of computers in teaching, 67% on the use of

teachers identified the didactic use of technology in the classroom as a priority of professional development rather than the use of internet, software and multimedia.

At the beginning of training every teacher received their laptop, the corresponding teacher 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.

The full-training session had 40 hours[21], divided in 4 weeks[22] 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.[23] Training was organized by modality (i.e., curriculum, one-to-one, interactive-whiteboard, and computer lab) and delivered in two regional sites.

A total of 130 teachers participated from the full training program. From those that 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 passing a final assignment. There were 16 teachers that did not attend the full training session at any point in time. These teachers were offered a recovery training session; 9 of them attended and 7 were absent.

2.6 The design of the assessment[24]

A central part of the design of any experiment is to determine the outcome measure. 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. We followed the model of assessment development described in Shechtman et al. (2010).

First, we determined which content and skills should be included in the assessment. Following an in depth analyses 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.

GeoGebra and 49.2% on the elaboration of applications for GeoGebra. The results of the diagnostics were important to reinforce the idea that given that many teachers had basic knowledge in the use of technology, training should emphasize the pedagogical approach.

[21] The results of the diagnostic survey highlighted the need of training teachers in general aspects of the use of UBUNTU and the use of interactive whiteboards, given that 72.4% replied that they did not know the operating system and 86.4% have never worked with an interactive-whiteboard before. We organized two additional sessions of 4 hours each to familiarize teachers with the operating system and the use of the interactive whiteboards (for those randomized to that condition).

[22] Training occurred between the 9th of April and the 4th of May.

[23] Previous to the start of training, a virtual classroom was designed and installed to support the distance work of teachers. This is a moddle based site which contains forums and spaces to share resources, questions, and issues with the individual and group assignments.

[24] This sections draws from a report prepared for this project by Lara-Meloy, et al (2012).

Basic skills typically covered by 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: (1) we created a pool of items for both basic and higher-order content areas;[25] (2) experts reviewed the items; (3) we performed think-alouds with students of the higher-order items;[26] (4) we piloted the test and used Item Response Theory (IRT) to select items for the final assessment; and (5) a panel of experts reviewed the final assessment form.

For grading simplicity we decided that all items had to be multiple-choice. Basic skills items were drawn from existing Costa Rican tests and teacher materials. A group of experts and teachers generated a pool of higher-order skills items.[27]

Two test forms of thirty-five items each were created from seventy candidate items; the two forms shared no items in common.[28] We piloted these tests in eighth grade classrooms at two schools not participating in the project.[29] Within each class, students were randomly assigned to one form or the other. A total of 400 students were assigned to the test. On the day of the exam 184 and 181 took each exam.

For the purposes of selecting items for the final instrument we subjected the test forms to IRT analysis. 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 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.[30]

[25] We built a blue print listing how the basic and higher-order skills would be evaluated.
[26] During the think-alouds students solved the test exercises explaining their reasoning behind their chosen answer to a trained observer.
[27] We developed four templates that were used by the experts in producing test items.
[28] We classified questions by difficulty level: low, medium, and high. We created bins of 5 questions for each form and randomly assigned equal proportion of low, medium and high difficulty questions to each bin. We randomized all the choices for the multiple choice questions.
[29] We took the eighth grade rather than the seventh because most concepts in the exam are not taught until the seventh grade.
[30] See Appendix I for more details

The end result of this process is a 32 items geometry test that we administered during the experiment. 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) general SAT that we describe below.

3. 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, that is after the geometry unit had finished, we collected another round of information from teachers and students. We also gathered administrative information provided by the schools.

The intervention affected nearly 18,000 students, 190 teachers in 85 schools. We collected data on all teachers and all schools. However, because of cost considerations it was unfeasible to collect data on all students. On average each teacher was in charge of 3 sections (classrooms). Therefore, we decided to randomly select one section per teacher and collect data on that section only.

We 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.

At baseline we administered a standardized achievement test[31] 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.

At the end of the geometry unit we administered the geometry test discussed in Section 2.6; this is the main outcome of interest. Additionally, after completing the test the students filled a student questionnaire designed to measure treatment compliance, fidelity of implementation, class dynamics, and student's 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

[31] The Second Regional Comparative and Explanatory Study (SERCE).

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.[32] 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 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 class 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

[32] Results were almost identical when building the scales via a polytomous IRT model.

terms of students, classrooms and teachers; repetition rate, etc. The information of 2011 was used mainly to stratify the sample when we randomized treatment. The information of 2012 was used to assess balance of characteristics and as a sampling universe to build our sample of students to be surveyed and tested.

4. Empirical strategy

We want to measure the causal effect of assignment to one of the four possible treatment conditions. Random assignment of schools to treatment/control status allows us to identify the average treatment effect by simply comparing the means of each of the four treatment groups with respect to the control group.

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

$$Y_{isj} = \alpha_0 + \sum_{k=1}^{4} \alpha_k T_{sj}^k + \delta_{sj} + \varepsilon_{isj}. \quad (1)$$

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}={new pedagogical approach, interactive whiteboard, computer lab, one-to-one}. In some specification we pool together the three technology arms into one group (letting k={1,2}). We 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[33] (e.g., the average effect on student performance in the geometry test of being in a one-to-one school versus the status quo).

Identification of the causal parameters of interest with randomization is an asymptotic property and our samples are finite. Thus, we also present results that control on observable characteristics (we also use equation (1) to show that these characteristics are balanced at the start of the experiment) with the added bonus that this might lead to smaller standard errors (i.e., more precise estimates). In other words, we estimate by ordinary least squares a set of models of the following form:

$$Y_{isj} = \alpha_0 + \sum_{k=1}^{4} \alpha_k T_{sj}^k + \beta X_{isj} + \delta_{sj} + \varepsilon_{isj}, \quad (2)$$

where X_{isj} is a vector of student, teacher and school control variables. Student controls include pre-treatment SAT score, dummies of gender, dummies of age, mother's education and number of books at home. Teacher control variables include gender, age, and years of experience. School control variables include number of students in seventh grade math, number of classrooms in

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

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.

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.

5. Research sample and internal validity

Our research sample covers all 85 schools and all 190 teachers that participated in the study. Table 3 shows its sample size. On average schools had 2.2 to 2.4 seventh grade math classes per school; except in the computer lab treatment arm, where there were on average 1.8 classes in each school. We surveyed one complete class of eligible students per teacher. All students were eligible except for those with some kind of known physical or cognitive disability. On average each class had 25 students and about 1 percent of the students presented some disability.

Non-response rates were low. The first column of Table 4 presents the average and the standard deviation of several quality measures of our sample. Row 1 shows that the student post-treatment survey and geometry test had a non-response rate of 9.1 percent. Columns 2-5 show the regression coefficients and standard errors of a model described by equation (1) 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 arms. Column 6 presents the p-value of a joint Wald test of the null hypothesis that all coefficients are equal which we cannot reject.

The administration of the test was contracted to a polling company that surveyed schools and classes during a 4 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 14 days after the end of the second term. By design, it can be seen from row 2 columns 2-5 that there are very little differences in dates between schools on each treatment arm.[34]

[34] 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.

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. Nevertheless, we cannot reject the null that all treatment arms had similar non-response rates at baseline.[35] Since the SAT is a control variable in our main econometric models, in order to avoid losing too many observations we imputed the missing SAT using mean class characteristics and added a dummy in our models if the student had an imputed SAT. Our main results are robust to dropping observations with an imputed SAT in the sense that the point estimates do not change.[36] However, the standard errors are a bit 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 classes. The one-to-one classes were 12 percent more likely to be observed than the control classes. The remaining differences between treatment and control are smaller and non-statistically significant.

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 only to descriptive statistics.

Table 5 shows pre-treatment sample means and differences in those means across treatment groups. Overall these differences are small and usually not statistically significant. Half of the students in the research sample are female and on average they are approximately 13 years old. Three out of the sixty coefficient in Table 5 are statistical significant.[37] The students in the interactive whiteboard are younger and more likely to be female than those in the control group. However, these differences are very small. In the control group 50.4 percent are male and the average age is 12.96 while in the interactive whiteboard 46.7 percent are male and the average age is 12.86.

The rest of the student's background characteristics are similar across treatment groups. Approximately 42 percent of the students have mothers with primary education and 41 percent with secondary education. On average they report to have 3 books at home and 74 percent have

[35] 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.

[36] When dropping out observations with imputed SAT, the pre-treatment characteristics of the four treatments and control groups remain statistically equivalent. Moreover, the main treatment effects do not change. Results are available from the authors upon request.

[37] Appendix Figure 2 shows the distribution of p-values across hypothesis and outcomes.

access to computers at home which suggests familiarity with technology. For all these variables, we cannot reject that they are statistically the same across treatment groups.

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 responded correctly 46.6 percent of the questions in the exam[38]. The differences between 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. The majority of them have some kind of computer lab and internet access. Repetition rate is 8.7 percent and 44.7 percent are suburban schools. We cannot reject the null that all of these variables are the same for all five treatment groups.

At the beginning of the school year we announced to schools their treatment status. In the first panel of Table 6 we show that by the time of the announcement 83.6 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 4 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.

6. Results

We present in this section the results of the paper. We proceed in three steps. First, we show that the teachers used the materials and equipment as intended. Second, that the intervention produced the desired changes in terms of class dynamics. Third, we estimate the average effects of the program on geometry knowledge using the results of the test. We assess robustness of the results and also investigate treatment heterogeneity on tests scores. Finally, we look at class mediation.

Throughout this section we present OLS estimates of equation (2) which include controls for randomization strata, student, teachers and school characteristics. We present results only controlling for strata (i.e., equation (1)) in Appendix Tables B. Not surprisingly, given that the variables are balanced, point estimates do not change.

[38] In the nationally representative sample of sixth grade students in 2008 the response rate was 49 percent.

6.1. Treatment take-up

The take up of the treatment was high. As it was discussed in Section 2.5 the vast majority of teachers in the treatment arms took part and passed the training.[39] During training, before the intervention started, teachers became familiar with the equipment and material and had the opportunity to make several suggestions which were incorporated in the final version used in class.

We use the endline student surveys to create indicators of class material and technology use. The results for these outcomes are presented in the top panel of Table 7 where we estimate equation (2) 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.[40]

All estimates for the use of class 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 that were assigned to technology did use it. Reassuringly, interactive whiteboards were used only in interactive whiteboards schools and computers only in schools that should have received computers.

We use class observations to measure whether teachers and students were observed using different materials and tools in class. In the bottom panel of Table 7 we present results that show that students' workbooks and teachers' manuals were being used in almost all of the treated classroom. The three technology arms used the prescribed software (Geogebra), especially in the interactive whiteboard and the one-to-one groups. As expected, because schools using computer labs were supposed to use the lab only once or twice a week, usage of Geogebra in those schools was lower. Interestingly, students did not appear to have used internet in the classroom more than in the control schools. Finally, all treatment groups used the traditional blackboard less than the control classrooms.

6.2. Class Dynamics

High treatment take-up translated into changes in class dynamics. The pedagogical approach prescribed for the intervention had the objective of promoting an environment where there was

[39] The percent of teachers that were trained in each treatment arm is as follows: zero percent in the control group, 91 percent in the curriculum group, 97 percent in the interactive board group, 100 percent of teachers in computer labs schools and 94 percent in one-to-one.

[40] In Table 8 we report measures built using student data. Each variable is a dummy equal to one if at least half of the student in the class reported to have had access and used class materials, interactive whiteboards, laptops or some technology. We found very similar results using teacher level data and class observation data.

more student participation and more time devoted to exploration rather than practice. We analyze this by looking at what happened inside the classroom. First, using the perspective of the students, we constructed scales of active learning and classroom activity. Second, using data from the classroom observations, we analyze the time devoted to different classroom activities. Third, we build and analyze measures of the practices teachers used in the classroom. We present these results in Table 8.

Looking at the first two rows of Table 8, we show that the intervention generated a more active learning environment and an increase in classroom activity, especially in the group of schools that received technology. This implies that the students in the treatment group report explaining concepts to the class more often, preparing more exercises for others to solve, and frequently discussing possible solutions or arguments with other students.

The next eight rows of Table 8 use the classroom observations to highlight how the changes perceived by the students actually translated into differences in the use of time in the classrooms. The observer recorded the duration in minutes of the three main moments of the class: exploration of new concepts, formalization and practice. Similarly, he/she also recorded the amount of time allocated to different strategies used to teach: plenary lecture, class discussion, work in groups, 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 classes there is more time devoted to discussion and less to the teacher lecture. As the pedagogical changes in our new curricula prescribed more time is devoted to exploration and formalization and less to practice. Students are less likely to work individually.

In last two rows 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 make, explain and validate mathematical conjectures, explain relations between concepts, manipulate propositions, or discover 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 foster those practices. We see positive point estimates for all groups with larger magnitudes and statistical significance in the technology arms.

6.3. Student Learning

We interpret high take up and changes 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 translated into gains in learning.

Table 9 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.

All treatment groups that changed the pedagogy to a student-center approach learned less geometry than the control group. The average treatment effect of the new curricular approach is a reduction in test scores of 17.1 percent of a standard deviation. The effect of technology is mostly negative. Combining the one-to-one technology with the new curriculum 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 new curriculum suggesting that it does not help much. 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 lab and new curriculum without technology have the same deleterious effect.

In rows 2 and 3 of Table 9, 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 new curricular 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 the 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, however; the relative performance of each group is maintained when each section is removed.

The items on the test also vary by difficulty. In Panel B we classify the thirty-two test items in eight groups of four items each according to the number of percent correct answers in the

[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.

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 strata 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 on the treatment effect with school size.[42]

6.4. Treatment Effect Heterogeneity

Did any group experience gains in test-score results? No. In Figure 2 we show a local polynomial regression of the geometry test-scores (controlling for strata fixed effects) on 3 mediating variables: student pre-treatment SAT, teacher experience and teacher quality. In each graph we show 3 lines, the dashed line is for the control group, the solid line is for those students in the curriculum condition, and the 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.

In panel A 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 9 where we show separate estimates for students below (row A) and above (row B) the median. The main differences in treatment are between the curriculum and the one-to-one arms where the loss for the high ability students is more than twice the loss of the low ability students. 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 student. In rows (C) and (D) we show that there are no differences in the treatment effects between males and females.

[42] Panels A and B in Appendix Figure 3 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 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 strata 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.

In panel B of Figure 2 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.[43] Technology follows a similar pattern than the control group while curriculum looks flat over the whole range of experience. The treatment effect is negative at low levels of experience but the magnitude shrinks with experience up to a point where both treatment effects become positive. In Table 9 rows (E) and (F), we divide the sample according to whether teachers have up 11 years of experience (the median in our sample) or more. On average, classes led by teachers with experience below the median tended to perform significantly worse than in classes with more experienced teachers.

Finally, we measure the relationship between treatment and teacher quality. We built a measure of teacher quality 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 change the students on their percentile rank by more are considered better.[44] The data supports this interpretation. In the third panel of Figure 2, 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 9 rows (G) and (H), we divide the sample according to whether teachers are below or above the median quality. We do not find significant differences in treatment effects for teachers of low and high quality.

6.5. Class Mediation

The role of the teacher in the classroom is to facilitate or mediate the interaction between the students and the subject matter. A failure in learning is tantamount to a failure in mediation. Do we have any evidence of that occurrence? First, we show that classrooms in treatment conditions move through the syllabus at a slower pace. Second, we look at measures of student behavior and attitudes. Third, we study teacher's mediation.

We use teacher logs to track the progress in the completion of the geometry syllabus.[45] In Figure 3, 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 it can be seen the control group

[43] 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)).

[44] 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.

[45] 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.

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,[46] it does raise the point that there might have been significant adjustment costs.

The new pedagogical strategy assigned new roles to students and teachers in the classroom. Although we have shown that classes were more participative they may have also led to more disruption or students may have felt uncomfortable with their new roles. In particular, our heterogeneity analysis showed that the most abled students suffer significantly more from this intervention.

In Figure 4 we show local polynomial regressions for five outcomes scales (controlling for strata fixed effects) on student pre-treatment SAT. In particular, we look at scales for bad behavior, academic press, avoidance of novelty, academic engagement and preference for math. If we start by looking at bad behavior (e.g., "sometimes I bother my teacher during class") the index 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.

In the first five rows of Table 10 we estimate equation (2) on each separate scale and confirm that students' behavior deteriorates, they are less willing to experience with new learning strategies, they are more disengaged from the class, they are less pressed to exert effort, and they like math less. Then, we estimate the average treatment effect on all five outcomes combined.[47] The estimates are overall negative but insignificant. In the last two rows we separate the sample according to the pre-treatment SAT and find that high ability students are less engaged than students in the control group.

The similar heterogeneity patterns in learning outcomes and behavioral responses highlighted in this section are reassuring. So, why did the better students failed? 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 the strongest negative results in Table 10 are for the high ability students in the one-to-one schools.

In Table 11 we look at the experience of teachers with the new environment. We start by analyzing three scales. Access to new ideas aims at measuring how much professional

[46] 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.

[47] We estimate equation (2) 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 (see, for instance, Kling, Liebman and Katz, 2007).

development and feedback or discussions about new teaching strategies each teacher recently had. 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 but statistically insignificant effects. In the second panel we compute the average treatment effect on the combination of these three outcomes and this slightly improves the precision of the estimates.

We also analyze measures of teaching mediation and efficacy. The former, built using class observations, records whether teachers maintain the order of the class, offer student clear instructions, and properly answer students' questions. The latter, built using surveys, measures whether teachers exposed to the new curriculum feel less in control of the class and the learning experience of their students.

We find overall a negative impact of the treatment on these two scales. In the second panel we combined them into one measure and find negative and statistically significant treatment effects for all treatment arms. In the last panel we separate the effects by the quality of the teacher and show that, except in the case of the interactive whiteboard, it is the low quality teachers that tend to perform worse in terms of class mediation.

7. Concluding Remarks

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 (one of three units of the seventh grade program or about three months of teaching). We randomly assigned the 85 participant schools in this experiment to one of five conditions: (1) status-quo (i.e., control); (2) new curriculum design; (3) new curriculum design and an interactive whiteboard; (4) new curriculum design and a computer lab; (5) new curriculum design and a laptop for every child in the classroom. All students (18,000) and teachers (190) in the seventh grade of these schools participated in the experiment.

We find that the control group learned significantly more than any of the four intervention groups. The students using the new curriculum without technology learned about 17 percent of a standard deviation less than the status-quo. Learning was around 36 percent lower in the one laptop per student schools compared to control establishments. In the race between the three technologies (i.e., keeping the pedagogical approach constant) the interactive whiteboard is the one that fairs slightly better. 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 also suggests that teachers went through the motions as prescribed but did not master the innovation in a way that would have allowed students to get the most of 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 supervised 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 rule out other sources of possible biases. Indeed we showed that the experiment had perfect compliance, was internally valid and implemented on 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 the basic concepts but also higher order skills (that we expected the intervention would foster). We found that the treatment groups performed worse than the control in both learning dimensions. We also presented 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.

The question that remains to be answered is whether these results will persist in the long run. We can offer two possible conjectures. One possibility is that the pedagogical change produced a loss of specific human capital in the short run. The introduction of the micro-computer and the dynamo (see David, 1990) did not lead to an automatic increase in productivity. In fact, there was much surprise with the sluggish growth of productivity in the western world. 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.

We can apply their basic model to our setting. Our intervention induced a reduction in the cost of switching the pedagogical paradigm. Suppose that teachers maximize an inter-temporal utility function that is positively related with student learning and negatively affected by the cost of changing pedagogy. Then, they might decide to switch to another pedagogical approach even in the presence of a short run fall in student learning (a consequence of their loss of specific human capital) if this can be compensated by the long run gains. If we observe teachers switching, we might infer by a revealed preference argument, that in the long run learning would increase.

A second explanation is that teachers do not know if they can work well with a more active student role in the classroom. Our intervention reduced the cost of experimentation leading to losses for those teachers that are not capable of mediating active learning experiences. Assume that there is a distribution of types in terms of teachers' abilities to lead a more active classroom and teachers are imperfectly informed about their type. Analogously to Karlan's et al (2012) analysis of entrepreneurial experimentation, teachers can learn about their type through experimentation but they face cost constraints to experiment. Our intervention lowered this cost leading teachers to experiment with more active classes even when some of them were not equipped to deal with such demands.

In both interpretations high take-up is expected but the long-run consequences are different. In the first case, learning by doing will lead to positive treatment effects. In the second case, teachers will learn that they are not well suited for this strategy and give up this approach.

The policy advice is that the road for educational reform is costly in the short run. In the best case scenario, if we pursue these changes despite the short run costs, outcomes may improve with more training, fine-tuned materials, and the benefits of learning by doing. In the worst case scenario, the current cohort of teachers may never be able to deliver the desired results. There is a middle ground. The current crop of teachers self-selected into the profession with a lecture-style teaching approach. With the appropriate incentives they may leave the profession and make room for candidates better suited to work with this new pedagogical approach. In any case, governments should always weigh in the cost of reform on the current cohort of students.

References

Angrist, J., P. Pathak, and C. Walters, 2013, Explaining Charter School Effectiveness, forthcoming *American Economic Journal: Applied Economics*.

Arias, F., and M. Zuñiga, 2012a, Desarrollo de curriculum y materiales para el aula del proyecto GeoMate.

Arias, F., and M. Zuñiga, 2012b, La capacitación de profesores en el proyecto GeoMate.

Aaronson, D., L. Barrow, and W. Sander, 2007, Teachers and student achievement in the Chicago public high schools, *Journal of Labor Economics* 25, pp. 95–135.

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 Papers, 18183.

Berlinski, S., M. Busso, J. Cristia, and E. Severin, 2011, Computers in Schools: Why Governments Should Do their Homework, in *Development Connections: Unveiling the impact of New Information Technologies*, Alberto Chong (ed), New York: Palgrave Macmillan.

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

Clements, D., 1999, 'Concrete' Manipulatives, Concrete Ideas, Contemporary Issues in Early Childhood 1, pp. 45-60.

Cristia, J., P. Ibarraran, S. Cueto, A. Santiago, E. Severín, 2012, Technology and child development: Evidence from the One Laptop per Child Program, IDB Working Paper Series, 304.

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

Dobbie, W. and R. Fryer, 2013, Getting Beneath the Veil of Effective Schools: Evidence from New York City, forthcoming *American Economic Journal: Applied Economics*.

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

Fundación Omar Dengo (FOD) - Ministerio de Educación Pública (MEP), 2010, Diagnóstico Nacional de los niveles de Acceso, Uso y Apropiación de las tecnologías digitales en los docentes del sector público y subvencionado.

Fryer, R., 2012, Injecting Successful Charter School Strategies into Traditional Public Schools: Early Results from an Experiment in Houston, mimeo Harvard University.

Gersten, R., J. Ferrini-Mundy, C. Benbow, D. Clements, T. Loveless, V. Williams, I. Arispe, M. Banfield, 2008, Report of the Task Group on Instructional Practices, in: Foundations for Success: Report of the National Mathematics Advisory Panel. 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, pp. 112-135.

Glewwe, P., M. Kremer, S. Moulin, and E. Zitzewitz, 2004, Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya, *Journal of Development Economics* 74, pp. 251-268.

Harris, D., and T. Sass, 2011, Teacher training, teacher quality and student achievement, *Journal of Public Economics* 95, pp. 798–812

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

Kane, T. et al, 2010, Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project, MET Project Research Paper, Bill & Melinda Gates Foundation, 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, Bill & Melinda Gates Foundation, 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, mimeo.

Kennedy, M. M., 1999, Approximations to indicators of student outcomes. *Educational Evaluation and Policy Analysis, 21*, 345-363.

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

Lampert, M., 1990, When the problem is not the question and the solution is not the answer: Mathematical knowing and teaching, *American Educational Research Journal* 35, pp. 281-310.

Lara-Meloy Arias, T., S. Berlinski, M. Busso, L. Gallagher, J. Roschelle, and M. Zuñiga, 2012, Design and development of the student assessment instruments for GeoMate.

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

National Research Council, 2001, Adding it up: Helping children learn mathematics. J. Kilpatrick, J. Swafford, and B. Findell (Eds.). Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

OECD, 2009, Learning Mathematics for Life: A Perspective from PISA. Available at: http://www.oecd.org/pisa/pisaproducts/pisa2003/44203966.pdf

Shechtman N., G. Haertel, J. Roschelle, J. Knudsen, C. Singleton, 2010, Design and Development of the Student and Teacher Mathematical Assessments, SRI International

Zuñiga, M., 2003, Aprendizaje mediado por tecnologías digitales: La experiencia de Costa Rica, in UNESCO (2003), Educación y Nuevas Tecnologías: Experiencias en América Latina, pp. 99-114.

Appendix: Item Response Theory

In Figure A1, we plot the item characteristic curve for question 3 in exam 1 and question 34 in exam 2, the student scores are split into deciles representing the 10th, 20th, etc. percentiles of ability scores (θ), and the proportion of students answering the item correctly in each decile is plotted with a red dot. Ideally, these dots will track the item characteristic curve relatively well, indicating a good model fit.

The shape of the curve is determined by two parameters that roughly correspond to its center on the horizontal axis and the "steepness" of the curve. The difficulty parameter represents the location of the curve on the horizontal scale – technically it is the θ-coordinate where the curve crosses the line Y = 0.5. It represents the place on the ability axis where a student would have a 50% probability of answer this item correctly. Items with difficulty parameters in the high positive range would correspond to curves shifted to the right and therefore more difficult items (since a student would have to be located relatively high on the ability axis in order to have a 50% chance of answering the item).

A second parameter controls the "steepness" of the curve (one might think of it as the slope of the tangent line where the curve crosses Y = 0.5). Items with relatively steep curves are considered discriminating (and thus the parameter is called the discrimination parameter). Ideally, students with abilities slightly lower than that indicated by the item's difficulty would be highly unlikely to answer the item correctly, while those with abilities slightly higher than the item's difficulty would be very likely to answer the item correctly. The item is useful for discriminating between higher and lower ability students at that particular point on the ability spectrum.

Consider an item with a very shallow discrimination parameter, as question 34 in exam 2 in Figure A1. A student at the very low end of the ability distribution (θ=-3) has slightly more than a 30% chance of answering the item correctly, while a student at the high end of the distribution (θ=3) has just under a 40% chance of answering the item correctly. In other words, knowing whether a student answered this item correctly or not gives us very little information regarding their likely location on the ability scale.

After pilot testing, a summary document was prepared of each item's statistical characteristics, as well as the relevant sections from the think-aloud protocols. Each item was examined in turn, with the goal of selecting a final set of items out of the pool of seventy that were piloted. In many cases poorly discriminating items were discarded; in others it was determined that minor modifications could correct the obvious flaws in the item based on information from the think-aloud protocol.

Table 1: Backround and sample comparison
(mean characteristics)

	Schools in Sample [1]	Costa Rica (restricted) [2]	Costa Rica [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)			
% of female students (school)	50%	51%	50%
% 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)			
% of schools with library	74%	73%	56%
% 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

Notes: 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: Annualized Unit Costs (USD)

	Interactive Whiteboard	Computer Lab	One-to-One
Equipment Deployment			
Initial Investment	34.9	73.4	111.3
Recurrent Costs	6.0	14.9	19.1
Total	40.8	88.3	130.4

Note: Equipment deployment includes hardware and set up costs. For one-to-one and laboratory hardware includes laptops, desktops, carts, routers, projectors, and lockers. For smart-boards hardware includes smart-boards, laptops for teachers, desktops, routers, projector, and lockers. Infrastructure adjustments necessary for set up are included in the set up costs. The assumed lifespan of the hardware (laptop, desktop, smartboard, etc) is 5 years. The recurrent repair costs are assumed to be 5% of the initial costs.

Table 3: Sample Size

	Control	Curriculum	Interactive Whiteboard	Computer Lab	One-to-One	Total
Schools	20	20	15	15	15	85
Classes 7th Grade	44	46	36	28	36	190
Teachers in 7th Grade Math	44	46	36	28	36	190
Students	1084	1182	965	738	861	4830

Note: Each column shows the expected sample size (count) of school, classes, teachers and student in each sample.

Table 4: Non-Response Rates

	Average and S.D. All	Difference w.r.t. Control (coeff and s.e.)				p-value	Sample Size
		Curriculum	Interactive Whiteboard	Computer Lab	One-to-One		
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Student Level Variables							
Missing on Geo test day	0.091	-0.017	-0.008	0.029	0.009	0.201	4625
	[0.288]	[0.024]	[0.021]	[0.022]	[0.026]		
Geo test date (# days after end of geo unit)	14	1.813	-0.963	2.367	-1.675	0.094	4157
	[6.489]	[1.971]	[1.956]	[1.902]	[2.419]		
Missing SAT (among eligible students)	0.211	-0.027	-0.084	-0.138	-0.083	0.572	4157
	[0.408]	[0.091]	[0.070]	[0.084]*	[0.079]		
Student with disability (did not take geo test)	0.011	-0.010	-0.018	-0.020	-0.014	0.585	4881
	[0.103]	[0.012]	[0.012]	[0.012]*	[0.011]		
Teacher Level Variables							
Missing teacher survey (baseline)	0.005	-0.025	-0.020	-0.021	-0.020	0.902	190
	[0.073]	[0.019]	[0.015]	[0.016]	[0.015]		
Missing teacher survey (endline)	0.032	0.003	0.016	-0.046	-0.013	0.304	190
	[0.175]	[0.035]	[0.038]	[0.038]	[0.033]		
Missing class observation	0.195	0.027	-0.003	-0.041	-0.127	0.117	190
	[0.397]	[0.095]	[0.088]	[0.102]	[0.072]*		
Missing teacher log June	0.111	-0.022	-0.142	-0.216	-0.145	0.288	190
	[0.314]	[0.127]	[0.100]	[0.102]**	[0.095]		
Missing teacher log July	0.163	-0.147	-0.244	-0.242	-0.105	0.145	190
	[0.370]	[0.082]*	[0.074]***	[0.087]***	[0.081]		
Missing teacher log August	0.237	-0.102	-0.217	-0.162	-0.143	0.709	190
	[0.426]	[0.101]	[0.088]**	[0.107]	[0.093]		

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]-[5] shows the regression coefficients and the standard errors in square brackets corresponding to equation (1), a regression model that only include controls for strata. Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. Column [6] shows the p-value of a test of all coefficients jointly equal to zero. Column [7] shows the sample size.

Table 5: Differences in Pre-Treatment Means

	Average and s.d. All	Difference w.r.t. Control (coeff and s.e.)				p-value	Sample Size
		Curriculum	Interactive Whiteboard	Computer Lab	One-to-One		
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Student-level Variables							
Percent Male	0.489	-0.029	-0.038	0.008	-0.016	0.330	4157
	[0.500]	[0.019]	[0.022]*	[0.026]	[0.025]		
Age (years)	12.970	0.072	-0.093	0.032	0.015	0.004	4127
	[0.878]	[0.061]	[0.043]**	[0.052]	[0.058]		
Mother's Education (Primary)	0.419	0.046	-0.021	0.042	0.019	0.368	4106
	[0.493]	[0.044]	[0.048]	[0.064]	[0.050]		
Mother's Education (Secondary)	0.406	0.003	0.026	-0.045	0.030	0.180	4106
	[0.491]	[0.025]	[0.027]	[0.037]	[0.031]		
Number of Books at home	3.161	-0.085	-0.024	0.038	-0.151	0.541	3560
	[1.565]	[0.083]	[0.100]	[0.124]	[0.128]		
Have a PC/laptop at home	0.735	-0.033	0.020	-0.033	-0.009	0.342	3543
	[0.442]	[0.036]	[0.033]	[0.045]	[0.039]		
SAT (% Correct)	0.466	-0.019	0.003	-0.007	-0.022	0.394	3278
	[0.145]	[0.017]	[0.021]	[0.017]	[0.017]		
Teacher Level Variables							
Percent Male	0.486	0.029	0.181	0.201	0.076	0.426	185
	[0.501]	[0.127]	[0.110]*	[0.150]	[0.122]		
Age (years)	36.668	0.853	0.799	-1.359	0.490	0.165	184
	[7.772]	[1.385]	[1.248]	[1.234]	[1.452]		
Experience (years)	11.652	0.500	1.414	0.154	-0.389	0.428	184
	[6.543]	[1.251]	[1.070]	[1.293]	[1.138]		
School-Level Variables							
Students 7th Grade	219.694	-0.650	-2.643	-5.310	4.757	0.916	85
	[114.174]	[16.949]	[11.334]	[12.440]	[12.564]		
Classes 7th Grade	6.847	-0.000	-0.306	-0.372	0.094	0.616	85
	[3.053]	[0.380]	[0.327]	[0.378]	[0.350]		
Computer Lab	0.741	-0.000	-0.017	0.050	-0.083	0.859	85
	[0.441]	[0.148]	[0.161]	[0.153]	[0.153]		
Internet in School	0.729	0.150	0.101	-0.165	0.035	0.270	85
	[0.447]	[0.136]	[0.148]	[0.177]	[0.153]		
7th Grade Repetition	0.087	-0.018	-0.008	-0.013	-0.012	0.984	85
	[0.062]	[0.020]	[0.025]	[0.019]	[0.019]		
Not Urban	0.447	-0.050	0.110	-0.157	-0.157	0.235	85
	[0.500]	[0.148]	[0.137]	[0.157]	[0.163]		

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]-[5] shows the regression coefficients and the standard errors in square brackets corresponding to equation (1), a regression model that only include controls for strata. Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. Column [6] shows the p-value of a test of all coefficients jointly equal to zero. Column [7] shows the sample size.

Table 6: Gaming

	Average and S.D.	Difference w.r.t. Control (coeff and s.e.)				p-value	Sample
	All	Curriculum	Interactive Whiteboard	Computer Lab	One-to-One		Size
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Learned teaching assignment before lottery	0.837	-0.106	-0.118	0.035	-0.055	0.285	190
	[0.370]	[0.079]	[0.084]	[0.082]	[0.079]		
Class learned geometry 1st Term	0.016	-0.020	-0.045	-0.037	-0.041	0.794	190
	[0.125]	[0.050]	[0.049]	[0.041]	[0.044]		
Class learned 4 geo units in 1st Term	0.126	0.066	-0.034	-0.092	-0.117	0.113	190
	[0.333]	[0.122]	[0.109]	[0.098]	[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]-[5] shows the regression coefficients and the standard errors in square brackets corresponding to equation (1), a regression model that only include controls for strata. Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. Column [6] shows the p-value of a test of all coefficients jointly equal to zero. Column [7] shows the sample size.

Table 7: Treatment Take-up

	Difference w.r.t. Control (coeff and s.e.)				
	Curriculum	Interactive Whiteboard	Computer Lab	One-to-One	N
	[1]	[2]	[3]	[4]	[5]
Access/ reported use:					
Class materials	0.751	0.942	0.830	0.714	190
	[0.072]***	[0.059]***	[0.084]***	[0.077]***	
Interactive whiteboards	-0.007	0.983	-0.017	-0.040	190
	[0.034]	[0.022]***	[0.022]	[0.026]	
Students' laptops	-0.045	-0.041	0.941	0.923	190
	[0.044]	[0.047]	[0.049]***	[0.057]***	
Some technology in class	-0.052	0.943	0.924	0.883	190
	[0.057]	[0.055]***	[0.054]***	[0.063]***	
Observed use:					
Class uses student's workbook	0.840	1.056	0.965	1.036	153
	[0.064]***	[0.046]***	[0.079]***	[0.050]***	
Class uses teacher's manual	0.869	1.023	0.926	0.995	153
	[0.066]***	[0.060]***	[0.107]***	[0.061]***	
Class uses Geogebra software	-0.032	0.768	0.523	0.808	153
	[0.069]	[0.089]***	[0.103]***	[0.097]***	
Class uses internet	0.013	0.006	0.037	0.077	153
	[0.020]	[0.022]	[0.030]	[0.048]	
Class uses regular blackboard	-0.225	-0.223	-0.326	-0.460	135
	[0.133]*	[0.137]	[0.156]**	[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. Columns [1]-[4] shows 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 [5] shows the sample size.

Table 8: Class Dynamics

	Difference w.r.t. Control (coeff and s.e.)				
	Curriculum	Interactive Whiteboard	Computer Lab	One-to-One	N
	[1]	[2]	[3]	[4]	[5]
Active learning	0.028	0.093	0.049	0.084	4052
	[0.047]	[0.044]**	[0.053]	[0.037]**	
Classroom activity	0.121	0.223	0.117	0.141	4157
	[0.044]***	[0.049]***	[0.051]**	[0.053]***	
Exploration	0.328	0.422	0.568	0.457	153
	[0.086]***	[0.078]***	[0.089]***	[0.079]***	
Formalization	-0.087	-0.030	-0.093	-0.047	153
	[0.041]**	[0.048]	[0.048]*	[0.053]	
Practice	-0.241	-0.393	-0.475	-0.410	153
	[0.103]**	[0.098]***	[0.095]***	[0.098]***	
Class lecture	-0.046	-0.091	-0.056	0.020	153
	[0.038]	[0.037]**	[0.048]	[0.039]	
Class discussion	0.149	0.326	0.141	0.092	153
	[0.068]**	[0.066]***	[0.078]*	[0.062]	
Work in groups	-0.017	-0.070	0.010	-0.111	153
	[0.051]	[0.044]	[0.039]	[0.038]***	
Work in pairs	0.004	0.011	0.034	-0.034	153
	[0.040]	[0.034]	[0.043]	[0.038]	
Work individually	-0.091	-0.176	-0.129	0.033	153
	[0.073]	[0.074]**	[0.088]	[0.069]	
Math prescribed learning practices (Student)	0.308	0.649	0.566	0.757	153
	[0.218]	[0.216]***	[0.272]**	[0.224]***	
Math prescribed teaching practices (Teacher)	0.383	0.384	0.551	0.502	153
	[0.236]	[0.259]	[0.321]*	[0.247]**	

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]-[4] shows 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 [5] shows the sample size.

Table 9: Geometry Test Results

		Difference w.r.t. Control (coeff and s.e.)				
		Curriculum	Interactive Whiteboard	Computer Lab	One-to-One	N
		[1]	[2]	[3]	[4]	[5]
Geometry score		-0.171	-0.155	-0.210	-0.355	4157
		[0.080]**	[0.093]*	[0.118]*	[0.091]***	
Geometry score (Basic skills)		-0.142	-0.090	-0.175	-0.340	4157
		[0.079]*	[0.088]	[0.108]	[0.088]***	
Geometry score (Higher-order skills)		-0.126	-0.138	-0.273	-0.225	4157
		[0.054]**	[0.072]*	[0.086]***	[0.066]***	
Dependent Variable: Geometry score						
Student:	(A) Low ability	-0.041	-0.068	-0.173	-0.193	1658
		[0.080]	[0.078]	[0.090]*	[0.069]***	
	(B) High ability	-0.248	-0.176	-0.152	-0.411	1620
		[0.122]**	[0.136]	[0.164]	[0.134]***	
	(C) Males	-0.155	-0.175	-0.191	-0.332	2032
		[0.090]*	[0.094]*	[0.125]	[0.099]***	
	(D) Females	-0.188	-0.142	-0.216	-0.378	2125
		[0.084]**	[0.100]	[0.128]*	[0.100]***	
Teacher:	(E) Low experience	-0.317	-0.099	-0.353	-0.443	2182
		[0.136]**	[0.151]	[0.142]**	[0.136]***	
	(F) High experience	0.001	-0.235	0.044	-0.180	1862
		[0.110]	[0.067]***	[0.150]	[0.085]**	
	(G) Low quality	-0.147	-0.190	-0.197	-0.262	1929
		[0.066]**	[0.065]***	[0.080]**	[0.080]***	
	(H) High quality	-0.139	-0.127	-0.213	-0.408	1867
		[0.123]	[0.093]	[0.101]**	[0.082]***	

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]-[4] shows 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 [5] shows the sample size.

Samples of low/high ability students, low/high ability teachers, low/high quality teachers are described in Section 6.4.

Table 10: Student Mediation

	Difference w.r.t. Control (coeff and s.e.)				
	Curriculum	Interactive Whiteboard	Computer Lab	One-to-One	N
	[1]	[2]	[3]	[4]	[5]
(A) Bad behavior	0.089	0.096	0.002	0.090	4030
	[0.056]	[0.056]*	[0.063]	[0.066]	
(B) Avoid novelty	0.072	0.050	0.092	0.115	3943
	[0.053]	[0.053]	[0.065]	[0.061]*	
(C) Academic engagement	-0.040	0.047	-0.003	-0.003	3973
	[0.075]	[0.078]	[0.085]	[0.085]	
(D) Academic press	-0.011	-0.048	-0.014	-0.030	3917
	[0.048]	[0.046]	[0.046]	[0.047]	
(E) Preference for math	-0.140	-0.025	-0.085	-0.066	3970
	[0.077]*	[0.068]	[0.076]	[0.076]	
Student Combined Scale (-A-B+C+D+F)	-0.070	-0.034	-0.039	-0.061	3970
	[0.041]*	[0.041]	[0.045]	[0.053]	
Dependent Variable: Student Combined Scale					
Low Ability	-0.034	-0.003	-0.020	0.006	1978
	[0.045]	[0.053]	[0.049]	[0.061]	
High Ability	-0.105	-0.077	-0.052	-0.138	1992
	[0.053]**	[0.044]*	[0.052]	[0.050]***	

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]-[4] shows 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 [5] shows the sample size.

Samples of low/high ability students are described in Section 6.4. Construction of the combined scale is described in Section 6.5.

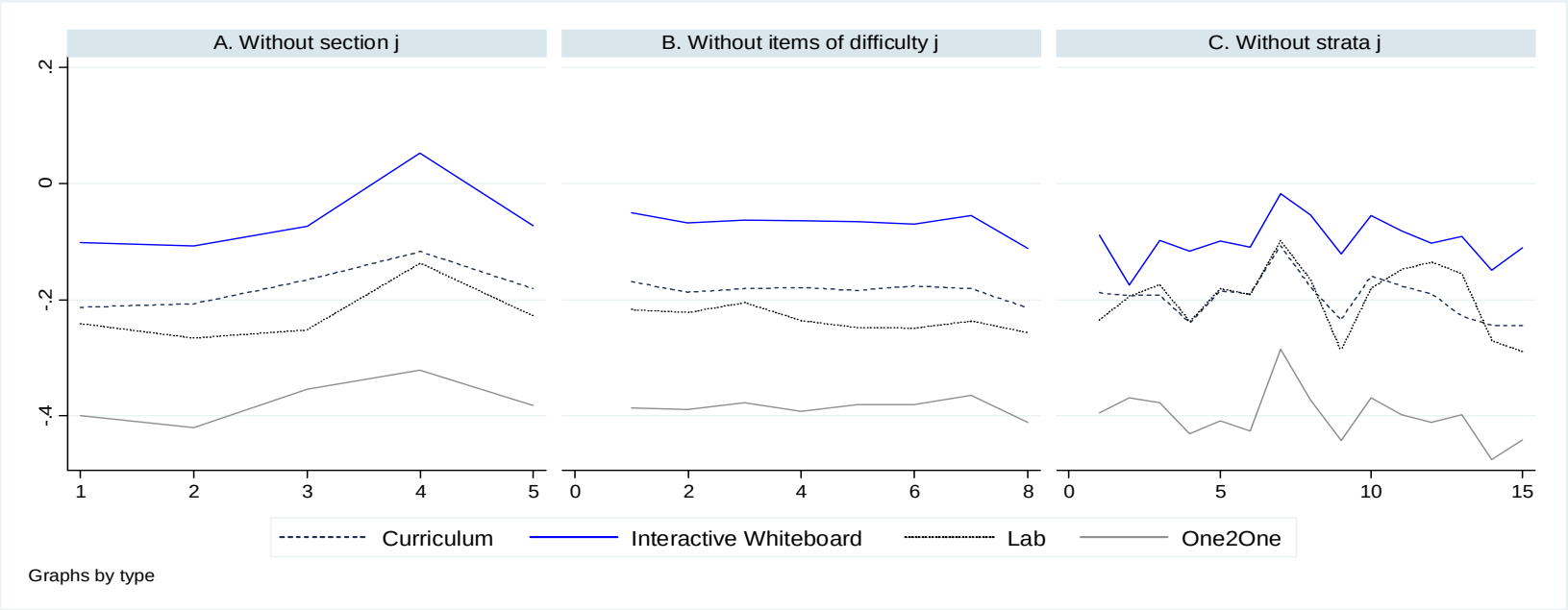
Table 11: Teachers Mediation

	Difference w.r.t. Control (coeff and s.e.)				
	Curriculum	Interactive Whiteboard	Computer Lab	One-to-One	N
	[1]	[2]	[3]	[4]	[5]
(A) Access to new ideas	0.204	0.246	0.536	0.382	184
	[0.268]	[0.279]	[0.303]*	[0.240]	
(B) Innovation	0.377	0.082	0.027	0.253	184
	[0.236]	[0.283]	[0.261]	[0.235]	
(C) Reflective dialogue	0.326	0.420	0.469	0.471	185
	[0.240]	[0.267]	[0.286]	[0.239]**	
(D) Teaching mediation	-0.667	-0.311	-1.195	-0.320	153
	[0.371]*	[0.313]	[0.486]**	[0.269]	
(E) Teaching efficacy	-0.108	0.007	-0.144	-0.214	187
	[0.201]	[0.234]	[0.244]	[0.195]	
Teacher Innovation Scale (A+B+C)	0.241	0.227	0.309	0.330	184
	[0.165]	[0.172]	[0.185]*	[0.157]**	
Teacher Mediation Scale (D+F)	-0.513	-0.343	-0.715	-0.347	153
	[0.206]**	[0.190]*	[0.246]***	[0.159]**	
Dependent variable: Innovation Scale					
Low Quality	0.356	0.564	0.272	0.498	86
	[0.282]	[0.314]*	[0.300]	[0.241]**	
High Quality	0.119	0.027	0.366	0.120	98
	[0.179]	[0.216]	[0.208]*	[0.217]	
Dependent variable: Mediation Scale					
Low Quality	-0.521	-0.023	-0.879	-0.419	74
	[0.346]	[0.358]	[0.357]**	[0.308]	
High Quality	-0.384	-0.443	-0.375	-0.179	79
	[0.284]	[0.283]	[0.407]	[0.227]	

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]-[4] shows 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 [5] shows the sample size.

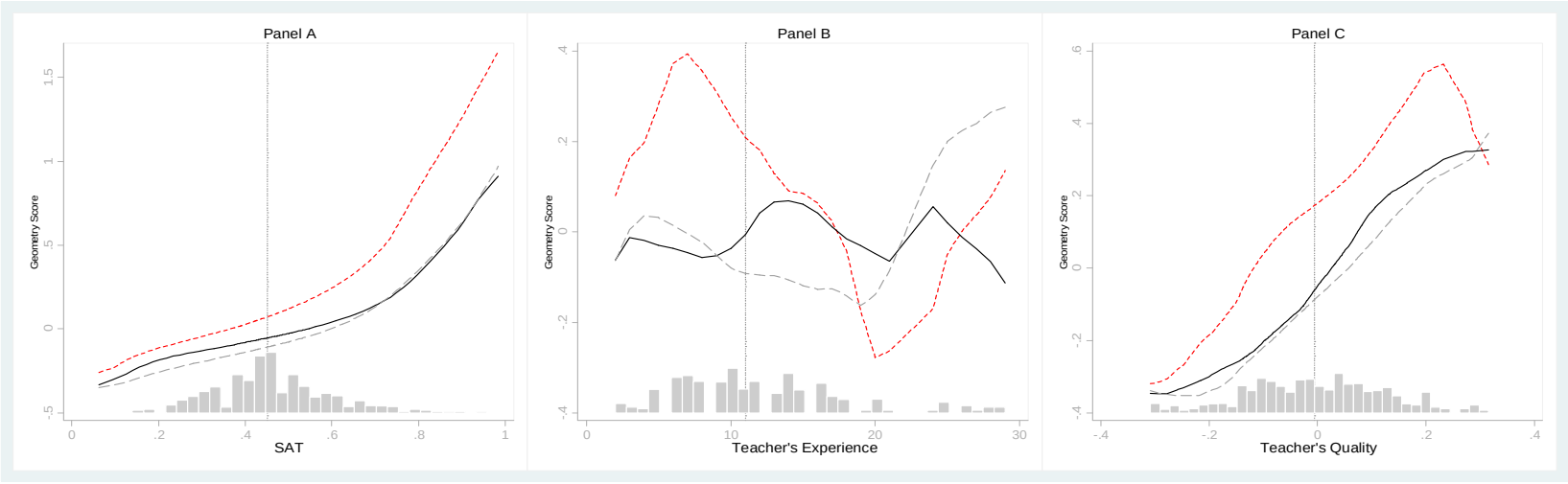
Samples of low/high ability teachers, low/high quality teachers are described in Section 6.4. Construction of the combined scales is described in Section 6.5.

Figure 1: Robustness of results on scores



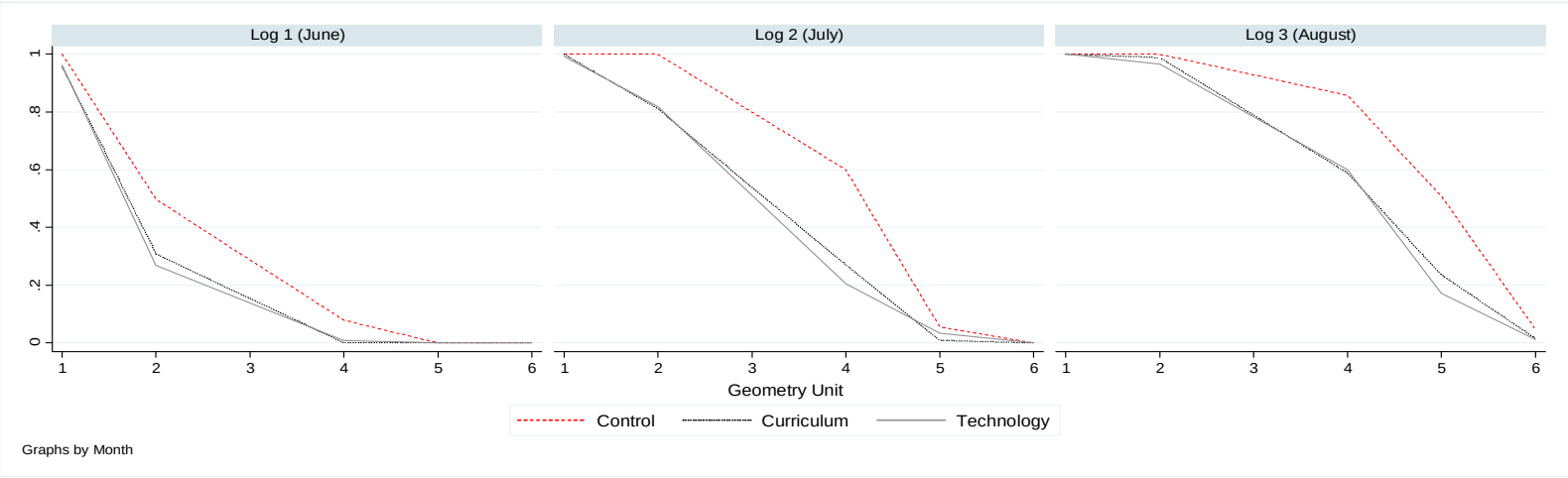
Note: The y-axis shows the treatment effect of a standardized geometry test score on treatment dummies estimated following equation (2). Panel A shows estimates obtained by removing items that belong to one (syllabus) section at a time. Panel B shows estimates obtained by removing items of one difficulty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

Figure 2: Treatment Effect Heterogeneity (Geometry Score)



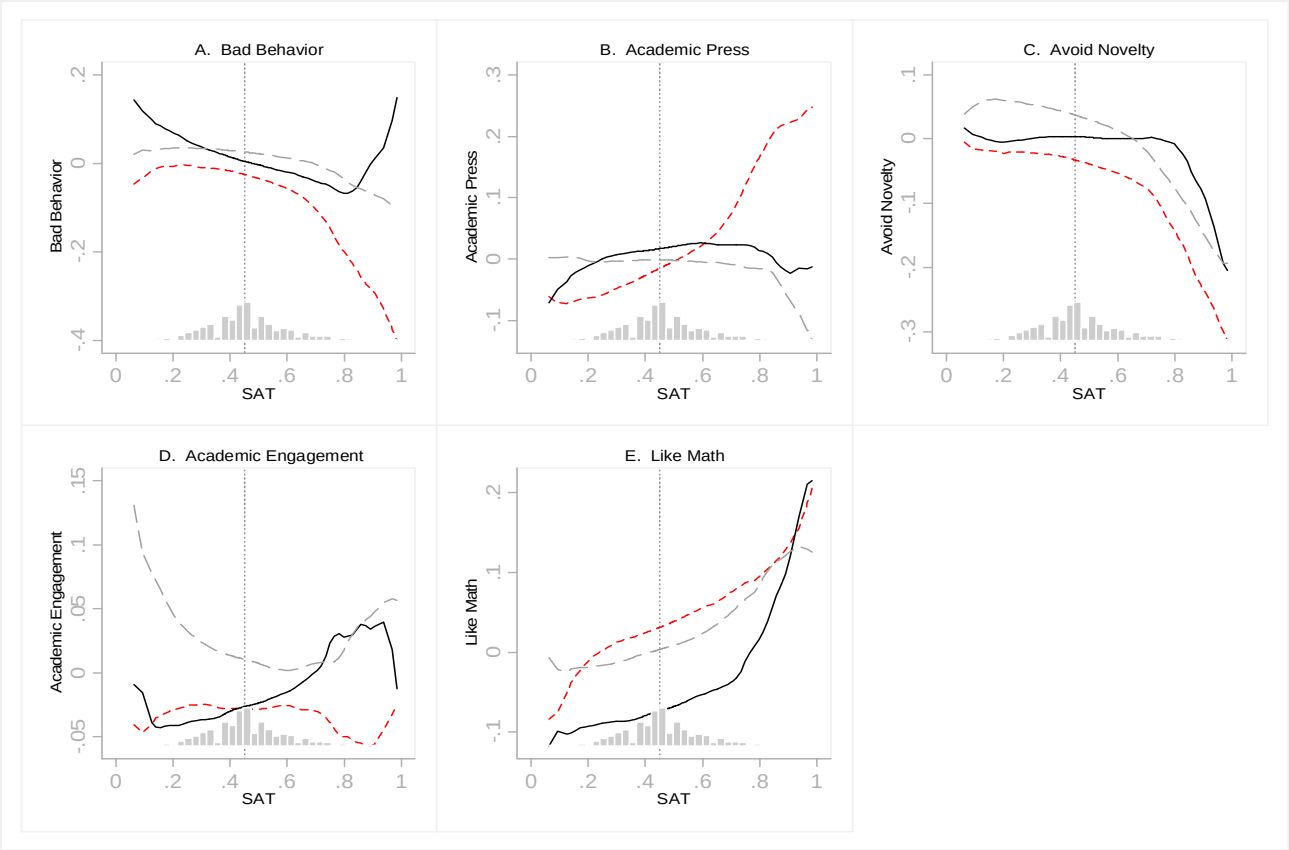
Note: Each line presents a local polynomial regression of the geometry test-scores (y-axis) --controlling for strata fixed effects-- on a mediating variable (x-axis): student pre-treatment SAT (panel A), teacher experience (panel B) and teacher quality (panel C). The red dashed line is for the control group, the black solid line is for those students in the curriculum 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. 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 3: 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 4: 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 curriculum 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. The local polynomial regressions were estimated using an Epanechnikov with a bandwidth of 0.15.

Table A1: Description of Test Items

Basic skills	Higher-order skills
Recognize geometric figures, given measurements, angles (e.g., recognize whether three lengths make up a triangle).	Identify parts of a geometric figure
Calculate an unknown given numbers or explicit information (big topic)	Classify geometric figures given properties in (angles, lengths)
Draw geometric figures given certain properties	Visualize (geometrically) theorems, problems, solutions
Construct figures to exemplify geometric situations	Justify theorems, using geometric figures, definitions or theorems
Recognize formal demonstrations of theorems.	Deduce patterns / formulate conjectures given examples
Write definitions	Read and use conventions and formal vocabulary

Table A2: Student Scales

Scale name and reliability measures		Scale survey question	Factor Loadings
[1]		[2]	[3]
Active learning (SRI)	1	In the math class I write math problems for other students to solve	0.2712
	2	I write a few lines to explain how I solved a math problem	0.5376
Eigenvalue: 1.437	3	I use math in situations outside the classroom	0.4883
Cronbach's Alpha: 0.640	4	I solve math problems that involve many steps and take more than 20 minutes	0.3722
	5	I explain to the class how to solve a math problem	0.6083
	6	I discuss possible solutions to math problems with other students	0.5723
Preference for math (SRI)	7	How much do you like mathematics?	0.7040
Eigenvalue: 1.925	8	Think about the most recent unit in your math class. Think about the activities and the math you learned. How much did you enjoy your math class during this unit?	0.8922
Cronbach's Alpha: 0.827	9	Think about the most recent unit in your math class. If math classes were always like this, would you be excited to take math classes in the future?	0.7961
Academic engagement (Chicago)	10	I often count down the minutes until class is over.	-0.4517
	11	What I am learning in class is so interesting, I don't want class to end.	0.6685
Eigenvalue: 1.72	12	I usually look forward to this class.	0.7009
Cronbach's Alpha: 0.678	13	I usually get bored with what we are learning in class.	-0.4733
	14	The topics we are studying are interesting and challenging.	0.5231
	15	I work hard to do my best in this class.	0.2916
Academic press (Chicago)	16	Nobody wastes time in class.	0.0870
	17	Usually this is a difficult class.	0.1833
Eigenvalue: 1.531	18	Usually the teacher asks difficult questions in class.	0.2144
Cronbach's Alpha: 0.638	19	Usually the teacher asks difficult questions on tests.	0.2241
	20	Usually this class challenges me.	0.4187
	21	This class really makes me think.	0.4070
	22	Generally this class requires me to work hard to do well.	0.2865
	23	The teacher expects everyone do their best all the time.	0.7205
	24	The teacher expects everyone to work hard.	0.6723
Avoid novelty (PALS-UM)	25	During class I prefer to work on tasks that are familiar to me rather than to learn how to do new ones.	0.2469
	26	I don't like to learn a lot of concepts during class.	0.3265
Eigenvalue: 0.896	27	I prefer to do my work as usual rather than to try something new.	0.4208
Cronbach's Alpha: 0.542	28	I like academic concepts that are familiar to me rather than ones I have never heard before.	0.4747
	29	I would rather chose to work on something I already know how to do rather than something I have never done before.	0.5711
Bad behavior (PALS-UM)	30	Sometimes I bother my teacher during class.	0.6775
	31	Soemtimes I get in trouble with my teacher during class.	0.6385
Eigenvalue: 2.124	32	Soemtimes I behave in a way that upsets my teacher during class.	0.6154
Cronbach's Alpha: 0.800	33	Sometimes I do not follow my teacher's instructions during class.	0.7239
	34	Sometimes I cause disorder during class.	0.5953
Classroom activity		Last week, how often during math class did the teacher conduct the following activities?	
	35	Lectures / demonstrations.	0.5150
Eigenvalue: 1.152	36	Class discussions.	0.5799
Cronbach's Alpha: 0.638	37	Work individually or in pairs.	0.6379
	38	Work in groups.	0.3791

Note: Each panel presents information on a given scale. Column [1] shows the scale name, its source between parenthesis, and reliability measures: Cronbach's alpha and the estimated eigenvalue obtained in a maximum likelihood principal components estimator with only one retained latent factor. Column [2] shows the survey question used to build each scale. Column [3] shows the loading factor that each survey question received in the maximum likelihood principal components estimator.

Table A3: Teachers Scales

Scale name and reliability measures		Scale survey/log/class observarion question	Factor Loadings
[1]		[2]	[3]
Innovation (Chicago)		The teachers in this school…	
	1	Are really trying to improve their teaching	0.7910
Eigenvalue: 1.986	2	Are willing to take risks to make the school better	0.4675
Cronbach's Alpha: 0.823	3	Are eager to try new ideas	0.5954
	4	Have a positive "I can do" attitude	0.5695
	5	Are continually learning and seeking new ideas	0.4747
	6	Are encouraged to "grow" professionally	0.4877
Reflective dialogue (Chicago)		In this school year, have you had conversations with your colleagues more than twice about…	
	7	What helps students learn the best	0.7587
Eigenvalue: 4.204	8	The mathematics curriculum	0.7635
Cronbach's Alpha: 0.847	9	The goals of this school	0.6615
	10	Managing classroom behavior	0.7542
	11	Teaching styles and learning	0.6942
	12	Teachers in this school discuss instruction in the teachers' lounge, faculty meetings, etc	0.7616
	13	Teachers in this school share and discuss student work with other teachers	0.7200
	14	Experienced teachers invite new teachers to observe their class, provide feedback, etc	0.4446
	15	The teacher body at this school makes new teachers feel welcomed	0.5108
Access to New Ideas (Chicago)		Usually…	
	16	I have discussed curriculum/instruction matters with an outside group	0.2069
Eigenvalue: 2.196	17	I have attended professional development activities organized by my school	0.3366
Cronbach's Alpha: 0.756	18	I have taken college/university courses relative to improving my school	0.3143
	19	I have participated in a network with teachers outside my school	0.4769
	20	I have worked with other teachers to develop materials or activities for specific classes	0.8238
	21	I have observed another teacher's class to obtain ideas about how to teach my class	0.2865
	22	I have reviewed my students' evaluations with other teachers to make decisions about teaching	0.4554
	23	I have observed another teacher's class to provide them with feedback	0.5530
	24	I have worked on teaching strategies with other teachers	0.6633
Teaching efficacy (Chicago)	25	With enough effort I can even make students with the most difficulty understand the subject	0.5199
	26	Events I can not control have a greater influence on the performance of my students than I do	-0.0341
Eigenvalue: 1.786	27	I am good at helping my students achieve significant improvements	0.8185
Cronbach's Alpha: 0.563	28	Some students will not make much progress this year, regardless of what I do	0.1693
	29	I am sure I can make a difference in the lives of my students	0.6724
	30	There is little I can do to ensure that all my students achieve significant progress this year	-0.0957
	31	I perform well under any teaching challenge	0.5959
Math prescribed learning practices -Student		Mark if the following prescribed activities are observed in class. Students:	
(Class observations)	32	Make mathematical conjectures	0.1924
Eigenvalue: 1.605	33	Explain the validity of a conjecture	0.6859
Cronbach's Alpha: 0.651	34	Describe or explain their reasoning	0.0898
	35	Work on proposed activities	0.0289
	36	Explain relations between different concepts	0.0703
	37	Maniputlate propositions and expressions using mathematical language	0.0416
	38	Discover a math rule or proposition from a pattern	0.0658
Math prescribed learning practices -Teachers		Mark if the following prescribed activities are observed in class. The teacher:	
(Class observations)	39	Foster students to derive math conjectures	0.7295
Eigenvalue: 1.561	40	Ask students to explain their reasoning	0.4586
Cronbach's Alpha: 0.682	41	Answer students questions with other questions that promote knowledge buidling	0.5064
	42	Builds mathematical language	0.3965
	43	Ask students to verify their conjectures	0.6363
Teacher mediation		Mark if you observe or don't the following teacher-students interactions:	
(Class observations)	44	Maintain class order/discipline	0.6031
Eigenvalue: 1.007	45	Offers students clear instructions	0.5533
Cronbach's Alpha: 0.454	46	Answer students questions	-0.0871
	47	Students follow instructions without difficulty	0.5696
	48	Students ask questions when they need to	0.0712

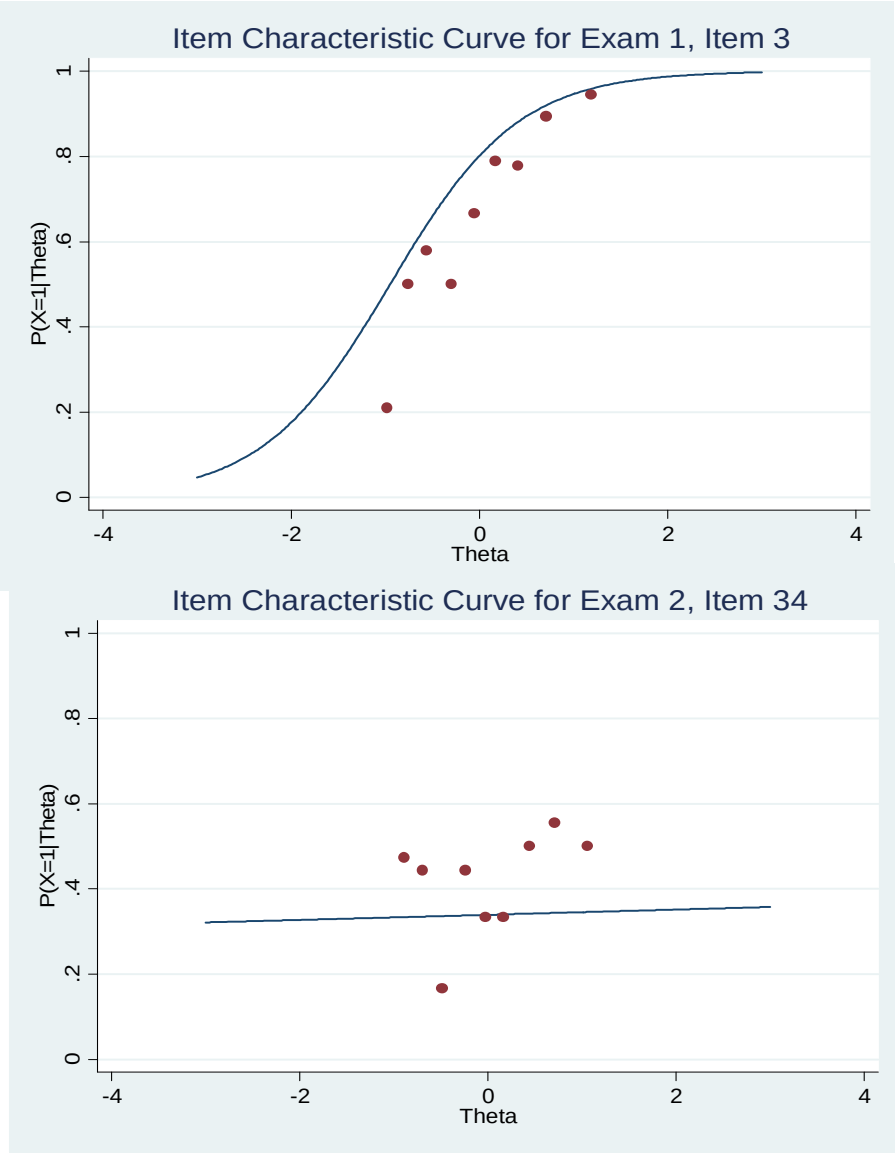
Note: Each panel presents information on a given scale. Column [1] shows the scale name, its source between parenthesis, and reliability measures: Cronbach's alpha and the estimated eigenvalue obtained in a maximum likelihood principal components estimator with only one retained latent factor. Column [2] shows the survey question used to build each scale. Column [3] shows the loading factor that each survey question received in the maximum likelihood principal components estimator.

Table A4: One Side Tests

Dependent Variable:	Without Controls	With Controls
p-values of one side test H1:	[1]	[2]
One2One <= Lab	0.055	0.093
One2One <= Curriculum	0.008	0.008
One2One <= Interactive whiteboard	0.002	0.012
Lab <= Interactive board	0.225	0.320
Lab <= Curriculum	0.500	0.366
Curriculum <= Interactive whiteboard	0.119	0.411

Note: standard errors in brackets are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. Individual controls include gender, age, mom educ, books, SAT. Teacher controls include gender, age, experience. School controls include # students in 7th grade, # classrooms in 7th grade, Lab in school, region dummies. Dependent variables: score is the % correct score and standardized IRT-score (was produced using the IRT parameteres in the control sample.) Both scores are standardizes using meand and s.d. of the control

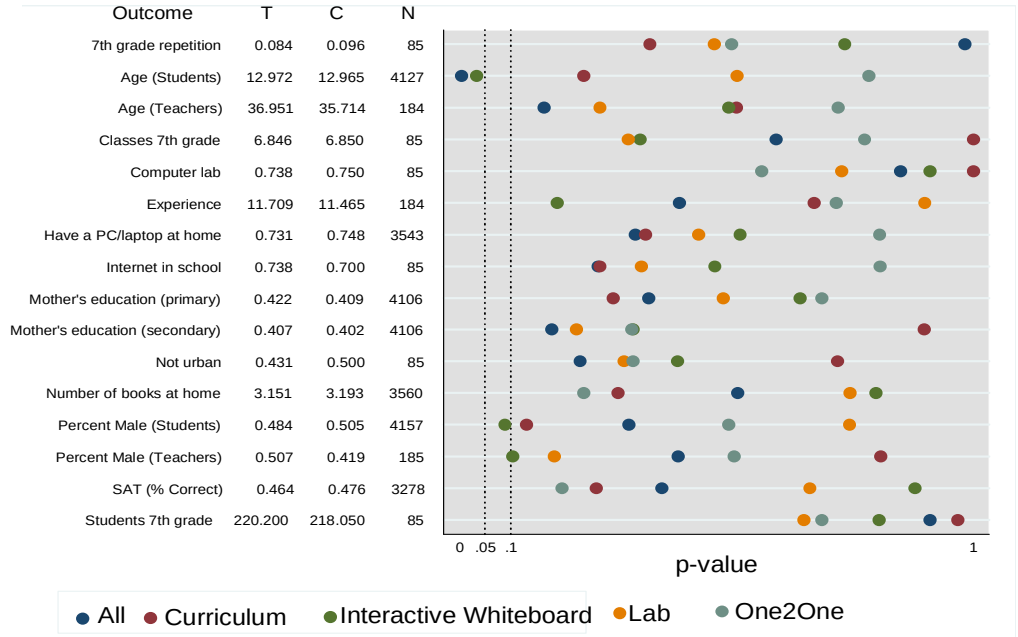
Appendix Figure 1



Note: Each figure shows the item characteristic curve for a test item. Student scores are split into deciles representing the 10th, 20th, etc. percentiles of ability scores, θ (x-axis). The proportion of students answering the item correctly in each decile is plotted with a red dot (y-axis). The curve is a two-parameter logistic fit to the data via maximum likelihood.

Appendix Figure 2: Balance Graph

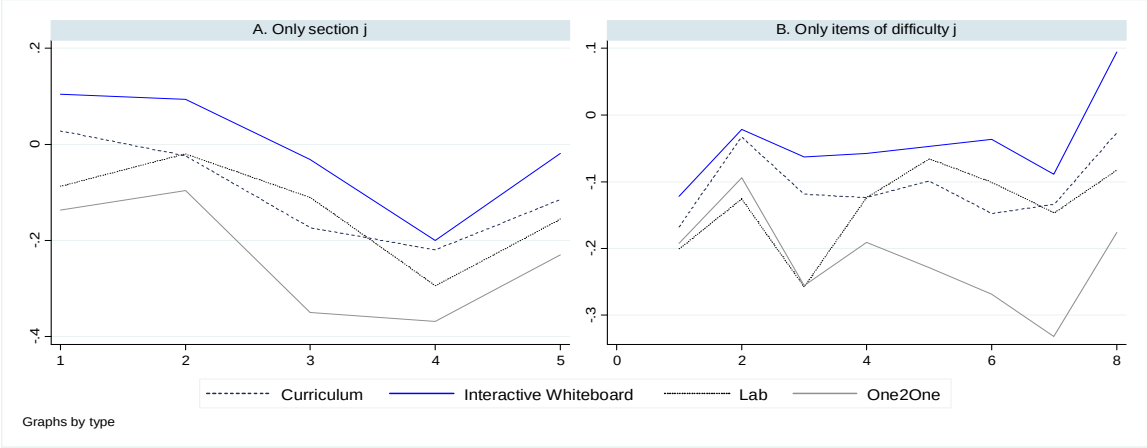
p-values of the test of equality of means in five treatment arms (joint and pair-wise tests)



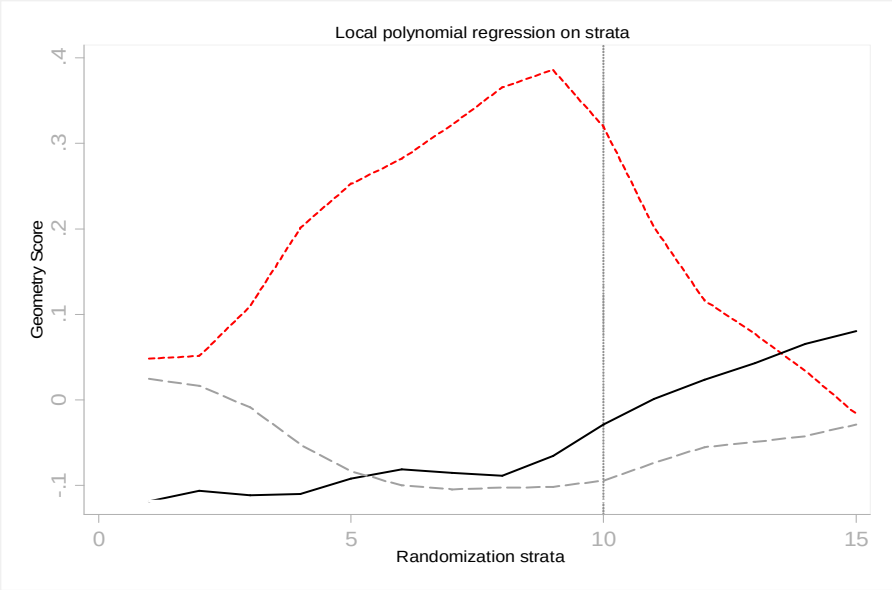
Note: Column "Outcome" shows the covariate, column "T" the mean among the treated (curriculum, interactive whiteboard, lab and one-to-one schools), column "C" the mean among the controls, column "N" the sample size. Each dot is the p-value of the t-test of the null hypothesis that the regression coefficient in equation (1) is equal to zero. The dots labeled "All" show the p-value of the null that all four point estimates are jointly equal to zero.

Appendix Figure 3: Treatment Effect Heterogeneity

Panel A and B: Test



Panel C: Sample



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 only on items of sections 1, 2…, 5. Panel B shows estimates on each subset of items of a given difficulty. Panel C shows a local polynomial regression of the geometry test-scores (y-axis) on strata with Epanechnikov kernel and a bandwidth of 2. The red dashed line is for the control group, the black solid line is for those students in the curriculum condition, and the grey long-dashed line is for those students in the three technology groups.

Table B7: Treatment Take-up (without controls)

	Difference w.r.t. Control (coeff and s.e.)				
	Curriculum	Interactive Whiteboard	Computer Lab	One-to-One	N
	[1]	[2]	[3]	[4]	[5]
Access					
Class materials	0.784	0.898	0.873	0.666	190
	[0.069]***	[0.052]***	[0.087]***	[0.062]***	
Interactive whiteboards	0.002	0.950	-0.045	-0.049	190
	[0.038]	[0.029]***	[0.027]	[0.029]*	
Students' laptops	-0.051	-0.048	0.922	0.917	190
	[0.050]	[0.049]	[0.060]***	[0.056]***	
Some technology in class	-0.050	0.903	0.877	0.869	190
	[0.061]	[0.056]***	[0.065]***	[0.061]***	
Use					
Class uses student's workbook	0.797	1.008	0.930	0.977	153
	[0.060]***	[0.029]***	[0.068]***	[0.032]***	
Class uses teacher's manual	0.830	0.978	0.889	0.949	153
	[0.057]***	[0.048]***	[0.089]***	[0.045]***	
Class uses regular blackboard	-0.305	-0.308	-0.428	-0.516	135
	[0.108]***	[0.138]**	[0.126]***	[0.104]***	
Class uses Geogebra software	-0.034	0.780	0.545	0.840	153
	[0.042]	[0.074]***	[0.108]***	[0.068]***	
Class uses internet	-0.000	0.004	0.001	0.061	153
	[0.012]	[0.012]	[0.014]	[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]-[4] shows the regression coefficients and the standard errors in square brackets corresponding to equation (1), a regression model which only includes strata controls. Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. Column [5] shows the sample size.

Table B8: Class Dynamics (without controls)

	Difference w.r.t. Control (coeff and s.e.)				
	Curriculum	Interactive Whiteboard	Computer Lab	One-to-One	N
	[1]	[2]	[3]	[4]	[5]
Active learning	0.035	0.112	0.064	0.054	4052
	[0.054]	[0.055]**	[0.056]	[0.048]	
Classroom activity	0.106	0.225	0.119	0.126	4157
	[0.049]**	[0.055]***	[0.059]**	[0.056]**	
Exploration	0.294	0.378	0.516	0.427	153
	[0.082]***	[0.068]***	[0.081]***	[0.068]***	
Formalization	-0.118	-0.054	-0.080	-0.074	153
	[0.044]***	[0.048]	[0.049]	[0.054]	
Practice	-0.176	-0.324	-0.435	-0.353	153
	[0.099]*	[0.084]***	[0.091]***	[0.074]***	
Class lecture	-0.087	-0.104	-0.065	-0.047	153
	[0.045]*	[0.038]***	[0.052]	[0.042]	
Class discussion	0.127	0.333	0.137	0.076	153
	[0.056]**	[0.065]***	[0.068]**	[0.062]	
Work in groups	0.017	-0.057	0.011	-0.078	153
	[0.044]	[0.041]	[0.039]	[0.033]**	
Work in pairs	0.014	0.020	0.063	-0.027	153
	[0.032]	[0.032]	[0.035]*	[0.026]	
Work individually	-0.071	-0.191	-0.146	0.076	153
	[0.060]	[0.068]***	[0.083]*	[0.063]	
Math prescribed learning practices (Student)	0.253	0.551	0.362	0.661	153
	[0.243]	[0.212]***	[0.269]	[0.226]***	
Math prescribed teaching practices (Teacher)	0.403	0.451	0.588	0.567	153
	[0.224]*	[0.209]**	[0.271]**	[0.216]***	

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]-[4] shows the regression coefficients and the standard errors in square brackets corresponding to equation (1), a regression model which only includes strata controls. Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. Column [5] shows the sample size.

Table B9: Geometry Test Results (without controls)

		Difference w.r.t. Control (coeff and s.e.)				
		Curriculum	Interactive Whiteboard	Computer Lab	One-to-One	N
		[1]	[2]	[3]	[4]	[5]
Geometry score		-0.202	-0.100	-0.202	-0.412	4157
		[0.120]*	[0.129]	[0.154]	[0.128]***	
Geometry score (Basic skills)		-0.178	-0.049	-0.174	-0.398	4157
		[0.120]	[0.128]	[0.143]	[0.123]***	
Geometry score (Higher-order skills)		-0.145	-0.081	-0.268	-0.267	4157
		[0.083]*	[0.096]	[0.103]***	[0.092]***	
Dependent Variable: Geometry score						
Student:	(A) Male	-0.184	-0.110	-0.195	-0.370	2032
		[0.118]	[0.126]	[0.157]	[0.126]***	
	(B) Female	-0.219	-0.093	-0.213	-0.456	2125
		[0.134]	[0.144]	[0.166]	[0.145]***	
	(C) Low ability	-0.016	-0.006	-0.115	-0.173	1658
		[0.083]	[0.088]	[0.102]	[0.079]**	
	(D) High ability	-0.371	-0.211	-0.275	-0.646	1620
		[0.198]*	[0.199]	[0.214]	[0.188]***	
Teacher:	(E) Low experience	-0.344	-0.051	-0.321	-0.531	2182
		[0.185]*	[0.239]	[0.198]	[0.208]**	
	(F) High experience	0.057	-0.109	-0.053	-0.165	1862
		[0.114]	[0.133]	[0.198]	[0.115]	
	(G) Low quality	-0.116	-0.095	-0.215	-0.256	1929
		[0.090]	[0.083]	[0.102]**	[0.108]**	
	(H) High quality	-0.292	-0.217	-0.214	-0.566	1867
		[0.219]	[0.220]	[0.203]	[0.174]***	

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]-[4] shows the regression coefficients and the standard errors in square brackets corresponding to equation (1), a regression model which only includes strata controls. Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. Column [5] shows the sample size.

Samples of low/high ability students, low/high ability teachers, low/high quality teachers are described in Section 6.4.

Table B10: Student Mediation (without controls)

	Difference w.r.t. Control (coeff and s.e.)				
	Curriculum	Interactive Whiteboard	Computer Lab	One-to-One	N
	[1]	[2]	[3]	[4]	[5]
Student Engagement Scale (-A-B+C+D+F)	-0.041	-0.001	-0.022	-0.054	3970
	[0.046]	[0.045]	[0.045]	[0.067]	
Student Engagement Scale Component:					
(A) Bad behavior	0.068	0.081	0.037	0.093	4030
	[0.070]	[0.068]	[0.072]	[0.078]	
(B) Avoid novelty	0.051	0.037	0.083	0.112	3943
	[0.053]	[0.047]	[0.070]	[0.068]*	
(C) Academic engagement	0.010	0.112	0.038	0.008	3973
	[0.083]	[0.085]	[0.078]	[0.114]	
(D) Academic press	0.006	-0.031	0.018	-0.010	3917
	[0.049]	[0.050]	[0.049]	[0.048]	
(E) Preference for math	-0.101	0.034	-0.044	-0.060	3970
	[0.079]	[0.075]	[0.070]	[0.101]	
Dependent Variable: Student Engagement Scale					
Low Ability	-0.028	0.016	-0.006	0.004	1978
	[0.046]	[0.056]	[0.047]	[0.079]	
High Ability	-0.057	-0.039	-0.060	-0.124	1992
	[0.064]	[0.053]	[0.059]	[0.066]*	

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]-[4] shows the regression coefficients and the standard errors in square brackets corresponding to equation (1), a regression model which only includes strata controls. Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. Column [5] shows the sample size.

Samples of low/high ability students are described in Section 6.4. Construction of the combined scale is described in Section 6.5.

Table B11: Teachers Mediation (without controls)

	Difference w.r.t. Control (coeff and s.e.)				
	Curriculum	Interactive Whiteboard	Computer Lab	One-to-One	N
	[1]	[2]	[3]	[4]	[5]
Teacher Innovation Scale (A+B+C)	0.264	0.191	0.210	0.307	184
	[0.170]	[0.176]	[0.172]	[0.166]*	
Teacher Mediation Scale (D+F)	-0.541	-0.355	-0.846	-0.375	153
	[0.239]**	[0.188]*	[0.237]***	[0.173]**	
Teacher Scale Component:					
(A) Access to new ideas	0.228	0.289	0.432	0.337	184
	[0.261]	[0.257]	[0.254]*	[0.246]	
(B) Innovation	0.235	-0.057	-0.170	0.180	184
	[0.202]	[0.250]	[0.200]	[0.208]	
(C) Reflective dialogue	0.329	0.341	0.369	0.403	185
	[0.220]	[0.215]	[0.219]*	[0.212]*	
(D) Teaching mediation	-0.800	-0.481	-1.385	-0.457	153
	[0.427]*	[0.317]	[0.473]***	[0.285]	
(E) Teaching efficacy	-0.282	-0.229	-0.307	-0.294	187
	[0.177]	[0.160]	[0.198]	[0.203]	
Dependent variable: Innovation Scale					
Low Quality	0.489	0.412	0.244	0.541	86
	[0.321]	[0.284]	[0.331]	[0.282]*	
High Quality	0.121	0.018	0.174	0.089	98
	[0.201]	[0.215]	[0.161]	[0.231]	
Dependent variable: Mediation Scale					
Low Quality	-0.545	0.067	-1.048	-0.392	74
	[0.337]	[0.279]	[0.372]***	[0.314]	
High Quality	-0.424	-0.534	-0.476	-0.236	79
	[0.268]	[0.270]**	[0.402]	[0.235]	

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]-[4] shows the regression coefficients and the standard errors in square brackets corresponding to equation (1), a regression model which only includes strata controls. Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1. Column [5] shows the sample size.

Samples of low/high ability teachers, low/high quality teachers are described in Section 6.4. Construction of the combined scales is described in Section 6.5.