

NBER WORKING PAPER SERIES

CAN ONLINE OFF-THE-SHELF LESSONS IMPROVE STUDENT OUTCOMES?
EVIDENCE FROM A FIELD EXPERIMENT

C. Kirabo Jackson
Alexey Makarin

Working Paper 22398
<http://www.nber.org/papers/w22398>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2016

This paper was previously circulated under the title “Simplifying Teaching: A Field Experiment with Online ‘Off-the-Shelf’ Lessons.” This paper was made possible by a grant from the Carnegie Corporation of New York through 100Kin10. We’re extremely grateful to Ginny Stuckey and Kate Novak at Mathalicious, and Sarah Emmons of the University of Chicago Education Lab, and Tracy Dell’Angela at the University of Chicago Urban Education Institute. We also thank math coordinators and the data management persons in Hanover, Henrico, and Chesterfield school districts. We thank Amy Wagner, Jenni Heissel, Hao Hu, and Mathew Steinberg for excellent research assistance. This paper benefited from comments from Jon Guryan, Eric Mbakop, Irma Perez-Johnson, Sergey V. Popov, Egor Starkov, and participants of APPAM 2015. The statements made and views expressed are solely the responsibility of the authors. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2016 by C. Kirabo Jackson and Alexey Makarin. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Can Online Off-The-Shelf Lessons Improve Student Outcomes? Evidence from A Field Experiment
C. Kirabo Jackson and Alexey Makarin
NBER Working Paper No. 22398
July 2016, Revised January 2017
JEL No. I20,J0,J48

ABSTRACT

There has been a proliferation of websites that warehouse instructional materials designed to be taught by teachers in a traditional classroom. While this new technology has revolutionized how most teachers plan their lessons, the potential benefits of this innovation are unknown. To present evidence on this, we analyze an experiment in which middle-school math teachers were randomly given access to high-quality “off-the-shelf” lessons. Only providing teachers with online access to the lessons increased students’ math achievement by 0.06 of a standard deviation, but providing teachers with online access to the lessons along with supports to promote their use increased students’ math achievement by 0.09 of a standard deviation. Benefits were much larger for weaker teachers, suggesting that weaker teachers compensated for skill deficiencies by substituting the lessons for their own efforts. Survey evidence suggests that these effects were mediated by both improvements in lesson quality and teachers having more time to engage in other tasks. We rationalize these results with a multitask model of teaching. The intervention is more scalable and cost effective than most policies aimed at improving teacher quality, suggesting a real benefit to making high-quality instructional materials available to teachers on the internet.

C. Kirabo Jackson
Northwestern University
School of Education and Social Policy
2040 Sheridan Road
Evanston, IL 60208
and NBER
kirabo-jackson@northwestern.edu

Alexey Makarin
Northwestern University
Department of Economics
302 Donald P. Jacobs Center
2001 Sheridan Road
Evanston, IL 60208
alexey.makarin@u.northwestern.edu

I Introduction

Teachers have been shown to have sizable effects on student test scores (Kane and Staiger 2008; Rivkin et al., 2005) and longer-run outcomes (Chetty et al., 2014a; Jackson, 2016b). However, relatively little is known about how to effectively improve teacher quality (Jackson et al., 2014). Teaching is a complex job that involves multiple tasks (Holmstrom and Milgrom, 1991) such as designing lessons, delivering them, managing the classroom, etc. However, most research on teacher effectiveness has been focused on how teachers deliver lessons (e.g. Pianta, 2011; Taylor and Tyler, 2012; Araujo et al., 2016) and stayed largely silent on the potentially important task of improving the lessons that teachers deliver. To help fill this space, in this paper we examine an intervention aimed at increasing the quality of the lessons used by teachers in the classroom. Specifically, we study the student achievement effects of providing teachers with free access to high-quality, off-the-shelf lessons on the Internet.

There has been a recent proliferation of lesson plans and instructional materials that can be accessed online. These lessons and materials are disseminated online but are usually designed to be taught by teachers in a traditional classroom. One of the early sites called **Teachers Pay Teachers** was launched in 2006 and allowed teachers to sell their lesson plans and instructional materials to other teachers. As of 2016, this site is estimated to have an active membership of approximately 4 million (this is more than all primary and secondary teachers in the United States which are estimated to be 3.5 million). Other major players in this product space provide mostly free and openly licensed instructional materials such as **LearnZillion**, **Pinterest**, and **Amazon Inspire** (Madda, 2016). There is considerable demand among teachers for these online resources. Opfer et al. (2016) found that over 90 percent of secondary teachers look to the internet for instructional materials when planning lessons. Even though this innovation may have little visible effect on how teachers deliver lessons, it has revolutionized how teachers plan and create the lesson content that they deliver.

Lesson sharing websites create a positive information externality such that all teachers, irrespective of geography or experience, may have access to high-quality lesson plans. These lesson plans may be designed by expert educators and may embody years of teaching knowledge and skills that most individual teachers do not possess themselves. In principle, through these websites, the creation of one high-quality lesson has the potential to improve the outcomes of millions of students. However, the extent to which providing teachers access to high-quality online instructional materials improves their student's performance is unknown. We present the first rigorous examination of this question. Specifically, we implemented a randomized field experiment in which middle-school

math teachers in three school districts were randomly provided access to high-quality off-the-shelf lessons, and we examine the effects on their students' subsequent academic achievement.

At the heart of our intervention are high-quality, off-the-shelf lessons. These lessons differ from those in most traditional math classrooms. In the typical US math class, teachers present definitions and show students procedures for solving specific problems. Students are then expected to memorize the definitions and practice the procedures (Stigler et al., 1999). In contrast, informed by education theory on inquiry-based instruction (Dostál, 2015), embedded learning (Lave and Wenger, 1991; Brown et al., 1989), classroom discussion (Bonwell and Eison, 1991), and scaffolding (Sawyer, 2005), the off-the-shelf lessons we used in this study were designed to promote deep understanding, improve student engagement, and promote retention of knowledge.¹ Under our experiment, teachers were randomly assigned to one of three treatment conditions. In the “license only” condition, teachers were given free access to these online lessons. Note that while these lessons were provided online, they are designed to be taught by teachers in a traditional classroom setting. To promote lesson adoption, some teachers were randomly assigned to the “full treatment” condition. In the full treatment, teachers were granted free access to the online lessons, received email reminders to use them, and were invited to an online social media group focused on lesson implementation. Finally, teachers randomly assigned to the control condition continued “business-as-usual.”

Because the treatments were assigned randomly, we identify causal effects using multiple regression. Students of teachers in the license only group and the full treatment group experienced a 0.06σ and 0.09σ test score increase relative to those in the control condition, respectively. The full treatment has a similarly sized effect as that of moving from an average teacher to one at the 80th percentile of quality, or reducing class size by 15 percent. Because the lessons and supports were all provided online, the marginal cost of this intervention is low. Moreover, the intervention can be deployed to teachers in remote areas where coaching and training personnel may be scarce, and there is no limit to how many teachers can benefit from it. Back-of-the-envelope calculations suggest a benefit-cost ratio above 900, and an internal rate of return greater than that of interventions such as the Perry Pre-School Program (Heckman and Masterov, 2007), Head Start (Deming, 2009), class size reduction (Chetty et al., 2014b) or increases in per-pupil school spending (Jackson et al., 2016).

A useful methodological innovation of this paper is that we demonstrate that, even with a single year of achievement data, one can test for heterogeneous treatment effects by teacher/classroom

¹While there is much observational evidence that teachers who engage in these best practices have better student outcomes (e.g., Pianta, 2011, Mihaly et al., 2013, Araujo et al., 2016), there is very little experimental evidence on how promoting best practices among existing teachers impacts achievement tests.

quality using conditional quantile regression models (Koenker and Bassett, 1978). We follow this approach. Even though information technology is complementary to worker skill in many settings (e.g. Katz and others, 1999, Akerman et al., 2015), our results reveal that the benefits of online lesson use are the largest for the least effective teachers, and decrease with effectiveness (as measured by teacher/classroom value added).

To highlight the economics at play, we conceptualize teaching as being a multitask job (Holmstrom and Milgrom, 1991) involving two complementary tasks: planning lessons and all other activities (lesson delivery, classroom management, etc.). We model student achievement as a function of both the quality of the lesson the teacher teaches and also how effective she is at other complementary activities (such as deliver the lesson, manage classroom behaviors, etc.). The best teachers are those that can perform both tasks. Off-the-shelf lessons are a technology that guarantees a minimum lesson quality for a fixed cost. By allowing teachers who use this technology to focus less of their time on lesson planning and more of their effort on lesson delivery (i.e. specialize), the technology simplifies the job of teaching. In the model, student outcomes may improve due to improvement in the quality of lessons, but also through teachers having more time to spend on other tasks. The larger effects for weaker teachers are consistent with our model in which (a) weaker teachers experience the largest direct improvements in lesson quality, (b) the amount of planning time saved by using the lessons is higher for less able teachers, and (c) the marginal effect of any time savings associated using off-the-shelf lesson is larger for weaker teachers (due to diminishing returns).

Looking to mechanisms, in our preferred specification, teachers who were only granted free access to the lessons looked at 1.59 more lessons, and report teaching 0.65 more lessons than control teachers, while, on average, fully-treated teachers (access plus supports) looked at 4.4 more lessons, and taught 1.9 more lessons than control teachers. This is consistent with the positive test score effects being driven largely by lesson use. The level of lesson use in the full treatment relates to about one-third of a years' worth of material. This pattern suggests that the effects are likely driven by increased lesson use. We also analyze effects on student surveys. Consistent with improved lesson quality and the aims of the intervention, treated students are more likely to say that teachers emphasize deep learning and more likely to feel that math has real life applications. Consistent with teachers spending more time on tasks complementary to lesson planning (as theorized in the model), treated teachers give students more individual attention. The facts that there were meaningful test score gains in the license only condition suggests that the improved outcomes in the full treatment condition are not driven by the additional supports to promote lesson use but by increased lesson use itself. To provide further evidence of this, we show that (a) the treatment arms with the largest increases in lesson use also had the largest test score improvements, (b) on average,

the test score effects increase with lesson use, and (c) conditional on lesson use, receiving the extra supports was unrelated to test scores.

Given the large documented benefits to lesson use, we explore why take-up was not more robust. Since lesson use was voluntary, the regular reminders and additional supports to use the lessons may have been important in the full treatment condition. Though suggestive, we uncover patterns in the data that are consistent with the relatively low levels of lesson adoption being due to teachers' behavioral biases. In our context, such biases may lead teachers to procrastinate and postpone exerting the effort to implement the lessons until it is too late (O'Donoghue and Rabin, 1999). While speculative, the notion that such biases are at play is supported by survey evidence, and the fact that lesson use dropped off most suddenly when the email reminders ceased.

The approach to improving instructional quality we study is a form of division of labor; classroom teachers focus on some tasks, while creating instructional content is (partially) performed by experts with particular skills in that domain. As such, this paper adds to a nascent literature exploring the potential productivity benefits of teacher specialization in schools (e.g. Fryer, 2016; Jacob and Rockoff, 2011). Moreover, our findings contribute to the education policy literature because the light touch approach we employ stands in contrast to more involved policy approaches that seek to improve the skills of the existing stock of teachers through training, selection, or changes in incentives (e.g. Taylor and Tyler, 2012; Muralidharan and Sundararaman, 2013; Rothstein and others, 2015). Also, while there is evidence certain kinds of instructional materials are *associated* with better student outcomes (Chingos and Whitehurst, 2012), we provide an experimental demonstration that an intervention that exogenously introduces high-quality instructional materials into existing classrooms has a sizable causal effect on student outcomes. The findings also contribute to the growing literature on the effective use of technology in education. Most existing studies of technology in education have focused on the effects of computer use among students (e.g. Beuermann et al., 2015; for a recent survey, see Bulman and Fairlie, 2016) or on the effects of specific educational software packages (e.g. Angrist and Lavy, 2002, Rouse and Krueger, 2004, Banerjee et al., 2007, Barrow et al., 2009, Taylor, 2015). In contrast, this paper examines whether technology can help teachers enhance their traditional teaching practices through the dissemination of teaching knowledge in a scalable and cost-effective way.² Finally, this study relates to the personnel economics and management literatures by presenting a context in which one can improve worker productivity by simplifying the jobs workers perform (Bloom et al., 2012; Jackson and Schneider, 2015; Anderson et al., 2001; Pierce et al., 2009).

²In related work, Comi et al. (2016) find that effectiveness of technology at school depends on teachers' ability to incorporate it into their teaching practices.

There are now thousands of lessons of varying quality available for download online. From a policy perspective, websites that make online lessons and instructional materials available to anyone with an internet connection may facilitate a large positive externality and have revolutionized how teachers source and create instructional content. Indeed, this change in how teachers create instructional materials has led to recent popular press headlines such as “*How the Internet is complicating the art of teaching*” and “*How did we teach before the Internet?*”. Our findings suggest that if districts can identify high-quality lessons, make them available to their teachers, and promote their use, the benefits could be as large, if not larger, than the positive effects we document here.

The remainder of the paper is as follows. Section II describes the off-the-shelf lessons used in the intervention and outlines the experiment. Section III describes the data. Section IV presents the empirical strategy and Section V describes the main results we obtained. Section VI provides a stylized model which is used to derive testable predictions, Section VII explores the mechanisms, and Section VIII concludes.

II The Intervention

II.1 The Off-the-Shelf Lessons

The job simplifying technology at the heart of the intervention is off-the-shelf lessons. These lessons are from the Mathalicious curriculum.³ Unlike a typical math lesson that would involve rote memorization of definitions provided by the teacher along with practicing of problem-solving procedures (Stigler et al., 1999), Mathalicious is an inquiry-based math curriculum for grades 6 through 12 grounded in real-world topics. All learning in these lessons is contextualized in real-world situations because students engage in activities that encourage them to explore and think critically about the way the world works.⁴ For example, in one of the more simple lessons titled “*New-Tritional Info*” (see Appendix K), students investigate how long LeBron James (a well-known National Basketball Association athlete) would have to exercise to burn off the calories in different McDonald’s menu items. This more simple lesson would likely be taught over the course of one or

³<http://www.mathalicious.com/about>

⁴Mathalicious lessons are designed for teaching applications of math. The Common Core defines rigorous mathematics instruction as having an equal emphasis on procedures, concepts, and applications. Teaching procedures involve showing students how to perform certain mathematical procedures, such as how to do long division. Teaching concepts would involve simple word problems that make the mathematical concept clear. Teaching applications are where students use math to explore multiple facets of some real-world question. In teaching applications, students would develop their own models, test and refine their thinking, and talk about it with each other. Model-eliciting activities (Lesh and Doerr, 2003) would fall into this category.

two class periods. Because most secondary school children are familiar with McDonalds and LeBron James, this lesson is interesting and relevant to their lives, and the math concepts presented are embedded in their everyday experiences. Also, because the lesson teaches students about rates through problem solving, students may gain an intuitive understanding of rates through experience rather than through rote memorization.

The lesson titled “*Xbox Xponential*” (see [Appendix L](#)) is a more complex, and representative, lesson that illustrates how students learn math through exploration of the real world. This lesson would be taught over three or four class periods. In the first part of the lesson, students watch a short video documenting the evolution of football video games over time. Students are asked to “*sketch a rough graph of how football games have changed over time*” and then asked to describe what they are measuring (realism, speed, complexity, etc). They are then guided by the teacher to realize that “*while a subjective element like ‘realism’ is difficult to quantify, it is possible to measure speed (in MHz) of a console’s processor.*” In the second part of the lesson, students are introduced to Moore’s 1965 prediction that computer processor speeds would double every two years. They are then provided with data on the processor speeds of game consoles over time (starting with the Atari 2600 in 1977 through to the XBOX 360 in 2005). Students are instructed to explain Moore’s law in real world terms and to use this law to predict the console speeds during different years. In the third part of the lesson, students are asked to sketch graphs of how game consoles speeds have actually evolved over time, come up with mathematical representations of the patterns in the data, and compare the predictions from Moore’s Law to the actual evolution of processor speeds over time. During this lesson, students gain an intuitive understanding of measurement, exponential functions, extrapolation, and regression through a topic that is very familiar to them - video games.

Teachers during these lessons do not serve as instructors to present facts (as is typical in most classroom settings), but serve as facilitators who guide students to explore and discover facts about the world on their own. The idea that math should be learned in real world contexts (situated learning) through exploration (inquiry-based learning) has been emphasized by education theorists for years ([Lave and Wenger, 1991](#); [Brown et al., 1989](#); [Dostál, 2015](#)). However, because the existing empirical studies on this topic are observational, this paper presents some of the first experimental evidence of a causal link between inquiry-based situated math instruction and student achievement outcomes.

Because the Mathalicious lessons are memorable and develop mathematical intuition through experience, they serve as “anchor lessons” that teachers can build upon during the year when introducing formal math ideas. For example, after teaching New-Tritional Info, teachers who are introducing the idea of rates formally would say, “*Remember how we figured out how long it takes*

for LeBron to burn off a Big Mac? This was a rate!” and students would use the intuition built up during the anchor lesson to help them understand the more formal lesson about rates (which may occur days or weeks later). Each of these “anchor lessons” touches on several topics and may serve as an anchor for as much as two months of math classes. This is particularly true for the more complex lessons such as “Xbox Xponential” that provides an intuitive introduction to several math concepts. When the Mathalicious curriculum is purchased by a school district, each Mathalicious lesson lists the grade and specific topics covered in that lesson, and proposed dates when each lesson might be taught. Full fidelity with the curriculum entailed teaching 5 to 7 lessons each year.

In addition to the lessons, one treatment arm of the intervention involved an additional component to facilitate lesson use, called Project Groundswell. Project Groundswell allowed teachers to interact with other teachers using Mathalicious lessons online through Edmodo (a social networking platform designed to facilitate collaboration among teachers, parents, and students).⁵ Through Edmodo, Project Groundswell provided a private online space to have asynchronous discussions with both Mathalicious developers and also other Mathalicious teachers concerning lesson implementation. Project Groundswell also included webinars (about 7 per year) created by Mathalicious developers. During these webinars, Mathalicious personnel would walk teachers through the narrative flow of a lesson, highlight key understandings that should result from each portion of the lesson, anticipate student responses and misconceptions, and model helpful language to discuss the math concepts at the heart of the lesson. In sum, Project Groundswell entailed online supports to facilitate Mathalicious lesson use.

II.2 The Experiment

Three Virginia school districts participated in this study: Chesterfield, Henrico, and Hanover. Across all grade levels, 59,186 students were enrolled in 62 Chesterfield public schools; In total, 50,569 students were enrolled in 82 Henrico public schools; and 18,264 students were enrolled in 26 Hanover public schools in the 2013-2014 school year (NCES). All grades 6 through 9 math teachers in these districts were part of the study. Teachers were placed into one of the three conditions described below:

Treatment Condition 1: Full Treatment (Mathalicious subscription and Project Groundswell)

Full treatment teachers were granted access to both the Mathalicious lessons and also Project Groundswell. They were invited to an in-person kickoff event where Mathalicious personnel re-

⁵<http://www.edmodo.com/>

viewed the on-line materials, introduced Project Groundswell, provided a schedule of events for the year, and assisted teachers through the login processes. During the first few months, full treatment teachers received email reminders to attend webinars in real time or watch recordings. Under Project Groundswell, teachers were enrolled in one of four grade-level Edmodo groups (grade 6, 7, 8, and 9). Teachers were encouraged to log in on a regular basis, watch the webinars, use their peers as a resource in implementing the lessons, and to reflect on their practice with Mathalicious developers and each other.⁶ Importantly, participation in all components of the treatment was entirely voluntary.

Treatment Condition 2: License Only Treatment (Mathalicious subscription only)

Teachers who were assigned to the license only treatment were only provided with a subscription to the Mathalicious curriculum. These teachers received the same basic technical supports available to all Mathalicious subscribers. However, they were not invited to participate in Project Groundswell (i.e. they were not invited to join an Edmodo group and did not receive email reminders). In sum, at the start of the school year, these teachers were provided access to the lessons, given their login information, and left to their own devices.

Treatment Condition 3: Control Condition (business-as-usual)

Teachers who were randomly assigned to the control condition continued “business-as-usual.” That is, control teachers continued to use the non-Mathalicious curriculum of their choice. They were not offered the Mathalicious lessons, nor were they invited to participate in Project Groundswell. Because these school districts had not been offered Mathalicious lessons before the intervention, control teachers would not have been familiar with the curriculum and would not have been using it. Insofar as any spillovers did occur (through treatment teachers sharing materials with colleagues in the control group), our estimated effects would be attenuated toward zero. In any case, our survey evidence on Mathalicious lesson use suggests that any spillover effects, if they exist, are negligible.

Assignment of Teachers to Treatment Conditions

Prior to conducting the study, for each district, the research team and Mathalicious decided on a predetermined number of licenses that could be allocated to teachers in each district. In summer 2013 (the summer before the intervention), the research team received a list of all math teachers eligible for this study from each district, with teacher qualifications and demographics. To facilitate district participation in the study, two of the districts were allowed to pre-select certain teachers that

⁶The Project Groundswell model is based on the notion that effective teacher professional development is sustained over time, embedded in everyday teacher practice (Pianta, 2011) and enables teachers to reflect on their practice with colleagues (Darling-Hammond et al., 2009).

they wished to receive access to the Mathalicious licenses (i.e., receive either Treatment Condition 1 or Treatment Condition 2). We refer to these teachers as “requested” teachers. All requested teachers were identified and removed from the control condition. All of the remaining unrequested licenses in each district were allocated randomly to the remaining teachers.⁷ As such, among those that were not requested teachers, whether a teacher received a license was random. In a second stage, among all teachers who had licenses (i.e. both those who were pre-selected and those who received the license by random chance) we randomly assigned half to receive the full treatment (i.e. Treatment Condition 2). Among non-requested teachers, treatment status is random conditional on district, and among requested teachers, assignment to the full treatment is random conditional on district. As such, treatment assignment was random conditional on *both* “requested” status and district, and the interaction between the two. Accordingly, all models condition on district and “requested” status and their interaction.⁸ Moreover, our main results are robust to excluding the requested teachers.⁹ The use of randomization ensured that conditional on requested status and district, teachers (and their students) had no control over their treatment condition and therefore reduced the plausibility of alternative explanations for any observed *ex post* differences in outcomes across treatment groups (Rosenbaum, 2002).

Table 1 shows the average baseline characteristics for teachers and students in each treatment condition. Baseline characteristics are similar across treatment conditions. To test for balance, we test for equality of the means for each characteristic across all three treatment conditions within each district conditional on requested status. We present the *p*-value for the hypothesis that the groups’ means are the same. Across the 17 characteristics, only one of the models yields a *p*-value below 0.1. This is consistent with sampling variability and indicates that the randomization was successful.

III Data

The data used in this study come from a variety of sources. The universe is all middle school teachers in the three school districts and their students (363 teachers and 27,613 students). Our first data sources are the administrative records for these teachers and their students in the 2013-4 academic year (the year of the intervention). The teacher records included total years of teaching

⁷Because the number of unrequested licenses varied across districts, the probability of being randomly assigned to the license condition varied by district. Note that all the empirical models include district fixed effects to account for such differences.

⁸Table A1 of Appendix A summarizes teacher participation by district, requested status, and treatment condition.

⁹See Appendix B for our main test score results that exclude requested teachers from the sample.

experience, gender, race, highest degree received, age, and years of teaching experience in the district. The administrative student records included grade level, gender, and race. Students were linked to their classroom teachers. These pre-treatment student and teacher attributes are shown in [Table 1](#).

The key outcome for this study is student math achievement (as measured by test scores). We obtained student results on the math portion of the Virginia Standards of Learning (SoL) assessment for each district for the academic years 2012-13 and 2013-14. These tests comprise the math content that Virginia students were expected to learn and were required for students in grades 3-8, Algebra I, Geometry, and Algebra II. These test scores were standardized to be mean-zero unit-variance in each grade and year.¹⁰ Reassuringly, like for all other incoming characteristics, [Table 1](#) shows that incoming test scores are balanced across the three treatment conditions. Note that test scores in 2013 are similar between students in the control and full treatment groups (a difference of 0.04σ) but in 2014 are 0.163σ higher in the full treatment condition relative to the control condition.¹¹ The relative improvement in math scores over time is $0.163-0.04=0.123\sigma$ between the full treatment and the control group. By comparison, the relative improvement in English scores over time (where there should be no effect) between the full treatment and the control group is 0.003σ . These simple comparisons telegraph the more precise multiple regression estimates we present in [Section V](#).

We supplement administrative records with data from other sources to measure lesson use and to uncover underlying mechanisms. Using teacher survey data, we observe the self-reported lessons they taught and read. Because these data are from surveys, using them will automatically have zeros for those individuals who do not complete the surveys - leading to an underestimate of the effect of the treatments on lesson use. We describe how we address this problem in [Section V](#). As such, we supplement these data with the more objective measure of lessons downloaded. Based on both these data sources, our two measures of Mathalicious lesson use are (a) the number of lessons looked at, and (b) the number of lessons taught. They are constructed as follows. For each lesson, we record whether it was downloaded for each teacher's account using tracker data from the Mathalicious website. For each lesson, we code up a lesson as having been looked at if either the tracker indicated that it was downloaded or if the teacher reported reading or teaching that lesson. The lessons taught measure comes exclusively from survey reports.

¹⁰In Hanover district, the exam codes were not provided so that the test scores are standardized by grade and year only. In our preferred specification, we control for the interaction between incoming test scores and district indicators.

¹¹So that we can include all students with math scores in 2014 in regression models, students with missing 2013 math scores are given an imputed standardized score of zero. To account for this in regression models we also include an indicator denoting these individuals in all specifications.

To explore causal mechanisms, surveys were given to students.¹² Survey questions were designed by the research team in conjunction with Mathalicious to measure changes in factors hypothesized to be affected by the intervention (see [Appendix C](#) for survey items). The student surveys were administered in the middle and at the end of the intervention year in two of the districts. We will focus on the end of year surveys. The student surveys were designed to measure student attitudes toward mathematics and academic engagement. While both teacher and student survey items are linked to individual teachers, the student surveys were anonymous. The survey items are discussed in greater detail in [Section VII](#).

IV Empirical Strategy

We aim to identify the effect of treatment status on various teacher and student outcomes. Owing to the random assignment of teachers, one can obtain consistent estimates of the treatment effect by comparing mean outcomes across the treatment conditions. As pointed out in [Bloom et al. \(2005\)](#), the statistical precision of estimated randomized treatments in education settings is dramatically improved by adjusting the outcomes for differences in pre-treatment covariates.¹³ Accordingly, to improve statistical precision, we compare outcomes across treatment categories in a multiple regression framework while controlling for a variety of student and teacher characteristics.

Because randomization took place at the teacher level, for the teacher-level outcomes, we estimate the following regression equation using ordinary least squares:

$$Y_{dt} = \alpha_d + \beta_1 License_{dt} + \beta_2 Full_{dt} + X_{dt} \delta_d + \pi_d Req_{dt} + \epsilon_{dt} \quad (1)$$

Y_{dt} is the outcome measure of interest for teacher t in district d , $License_{dt}$ is an indicator variable equal to 1 if teacher t was randomly assigned to the license only condition, and $Full_{dt}$ is an indicator variable equal to 1 if teacher t was randomly assigned to the full treatment condition (license plus supports). Accordingly, β_1 and β_2 represent the differences in outcomes between the control and the license only groups, and between the control and the full treatment groups, respectively. The

¹²We also administered teacher surveys for this study. However, due to high differential attrition rates the results are inconclusive and we do not discuss effects on these data in the main text. Teacher surveys were administered in the middle and at the end of the intervention year in all three school districts. They were designed to measure teacher job satisfaction and classroom practices. Results on the teacher surveys are presented in [Appendix I](#).

¹³Intuitively, even though groups may have similar characteristics on average, the precision of the estimates is improved because covariates provide more information about the potential outcomes of each individual participant. The increased precision can be particularly large when covariates are strong predictors of the outcomes (e.g. lagged test scores are very strong predictors of current test scores).

treatment assignment was random within districts and after accounting for whether the teacher was requested for a Mathalicious license. Consequently, all models include a separate dummy variable for each district to absorb the district effects, α_d , and we include an indicator variable Req_{dt} denoting whether teacher t requested a license in district d . To improve precision, we also include X_{dt} , a vector of teacher covariates (these include teacher experience, gender, ethnicity, and grade level taught) and student covariates averaged at the teacher level (average incoming student math and English test scores, and the proportion of males, and the proportion of black, white, Hispanic, and Asian students).

Our main outcome of interest is student test scores in mathematics. For this outcome, we estimate models at the individual student level and employ a standard value added model (Todd and Wolpin, 2003) that includes individual lagged test scores as a covariate (in addition to other individual student-level demographic controls and also classroom averages of all the student-level characteristics). Specifically, where students are denoted with the subscript i , in our test score models, we estimate the following regression equation using OLS:

$$Y_{idt} = \rho Y_{idt-1} + \alpha_d + \beta_1 License_{dt} + \beta_2 Full_{dt} + X_{idt} \delta_d + \pi_d Req_{dt} + \varepsilon_{idt} \quad (2)$$

In (2), X_{idt} includes student race, student gender, teacher level averages of the student-level covariates (including lagged math and English test scores), as well as all of the teacher-level covariates from (1). Because treatment status is randomly assigned at the teacher level, the use of student-level covariates primarily serves to improve statistical precision. However, including student-level covariates could also help to account for any potential imbalances across treatment groups. Standard errors are adjusted for clustering at the teacher level in all student-level models.

V Main Results

V.1 Effects on Student Achievement in Mathematics

To measure the effect of the intervention on student achievement in mathematics (our main outcome of interest), we use two forms of math test scores – raw and standardized scores. Raw test scores are measured on a 0-600 scale, while the standardized test scores refer to the raw scores standardized by exam. Test scores are analyzed at the individual student level, and standard errors are adjusted for clustering at the teacher level.

Results for math test scores are summarized in [Table 2](#). The results reveal positive effects on math test scores from simply providing licenses, and even larger positive and statistically significant effects for the full treatment. The first model (columns 1 and 3) includes the key conditioning variables (district fixed effects interacted with requested status) and the average lagged math scores in the classroom interacted with the district. Looking at raw test scores, this parsimonious model (Column 1) shows that scores were 2.65 points higher ($p\text{-value}>0.1$) among teachers in the license only condition, and 7.899 points higher among teachers in the full treatment condition than the control condition ($p\text{-value}<0.01$). Because standardized scores are easier to interpret, and they adjust for differences across grades and exam types, we focus our discussion on these standardized scores. In this model (column 3), teachers who only had access to the lessons had test scores that were 5% of a standard deviation higher than those in the control condition ($p\text{-value}>0.1$), and teachers with access to both Mathalicious lessons and extra supports increased their students' test scores by 10.5% of a standard deviation relative to those in the control condition ($p\text{-value}<0.05$). One cannot reject that the full treatment teachers have outcomes different from those in the license only group, but one can reject that they have the same outcomes as teachers in the control group.

Columns 2 and 4 present models that also include all teacher and classroom level controls. While the point estimates are similar, the standard errors are about 15 percent smaller. In the preferred student-level model in Column 5 (all student-level, teacher level, and classroom level controls), teachers who only had access to the lessons had test scores that were 6% of a standard deviation higher than those in the control condition ($p\text{-value}<0.1$). This modest positive effect indicates that merely providing access to high-quality lessons can improve outcomes. Looking at the full treatment condition, teachers with access to both Mathalicious lessons and extra supports increased their students' test scores by 8.6% of a standard deviation relative to those in the control condition ($p\text{-value}<0.05$). To ensure that the student and teacher level models tell the same story, we estimate the teacher level model where average test scores are the dependent variable (column 6). Because randomization took place at the teacher level, this is an appropriate model to run. In such models (with all teacher and classroom level controls), teachers in the license only condition increased their students' test scores by 5.5% of a standard deviation relative to those in the control condition ($p\text{-value}<0.1$), and full treatment condition increased their students' test scores by 9.3% of a standard deviation relative to those in the control condition ($p\text{-value}<0.01$). In sum, across all the models, there is a robust positive effect of both the license only treatment and the full treatment (relative to the control condition) on student test scores of roughly 6 and 9 percent of a standard deviation, respectively. Also, across all models, the full treatment is associated with larger and more precisely estimated math test score gains than the license only treatment.

To assuage concerns that the estimated effects are spurious, we report a falsification exercise

with English test scores as the main outcome in Columns 7 and 8. Even though assignment to treatment was random, one may worry that treated students, *by chance*, received a positive shock for reasons unrelated to the treatment. Alternatively, one may worry that there was something else that could drive the positive math test score effects that is correlated with the random treatment assignment. To test for these possibilities, we use data on English test scores at the end of the experiment. Because the Mathalicious website provided lessons only for math curriculum, test scores for English are a good candidate for a falsification test – if it were the lessons that drove our findings in Columns 1-6, not some unobserved characteristic that differed across experimental groups, then we would observe a positive effect for math scores and no effect for English scores. This is precisely what one observes. There are no statistically or economically significant differences in English test scores across treatment groups. This reinforces the notion that the improved math scores are due to increased lesson use and are not driven by student selection, Hawthorne effects, or John Henry effects.

V.2 Effect Heterogeneity by Teacher Quality

In principle, because effective teaching requires both good lesson plans and good lesson delivery, these off-the-shelf lessons may be complementary to teacher effectiveness. Conversely, weaker teachers who are relatively ineffective at improving student performance may benefit greatly from the provision of off-the-shelf lessons. To test which scenario holds empirically, we see if the marginal effect of the treatment is larger or smaller for teachers lower down in the quality distribution. Following the teacher quality literature, we conceptualize teacher quality as the ability to raise average test scores. Because we only have a single year of data, we cannot distinguish between classroom quality and teacher quality *per se*; however, we know from prior research that the two are closely related. As such, following [Chetty et al. \(2014a\)](#), we proxy for teacher quality with classroom quality. As is typical in the value-added literature, we define a high-quality classroom as one that has a large positive residual (i.e. a classroom that does better than would be expected based on observed characteristics) and we define a low-quality classroom as one that has a large average negative residual.

To test for effects by teacher effectiveness, ideally one would estimate teacher effectiveness using some pre-experimental data, and then interact the randomized treatment with the teacher's pre-treatment effectiveness. Unfortunately, we only have access to a single year of achievement data so that we take a different, but closely related, approach. To test for different effects for classrooms at various points in the distribution of classroom quality, we employ conditional quantile

regression. Conditional quantile regression models provide marginal effect estimates at particular quantiles of the residual distribution (Koenker and Bassett, 1978). As we formally show in Appendix D, when average test scores at the teacher level are the dependent variable, the teacher-level residual from (1) is precisely the standard value-added measure of classroom quality. That is, Appendix D shows that the marginal effect of the treatment at the p -th percentile from the conditional quantile regression of equation (1) is the marginal effect of the treatment for teachers at the p -th percentile of effectiveness. The interpretation of the point estimates from conditional quantile regression models applied to the teacher-level test score regressions are intuitive and fall naturally out of the empirical setup.¹⁴ To estimate the marginal effect of the full treatment for different percentiles of the classroom quality distribution, we aggregate test scores to the teacher level and estimate conditional quantile regressions for the 10th through 90th percentiles in intervals of 5 percentile points. We then plot the marginal effects of the full treatment against the corresponding quantiles along with the 90 percent confidence interval for each regression estimate. This plot is presented for math test scores in Figure 2.

Even though the relationship is non-linear, Figure 2 exhibits a clear declining pattern indicating larger benefits for low-quality classrooms than for high-quality classrooms. The estimated slope through the data points is -0.00073 (p -value <0.01) which implies that as one goes from a classroom/teacher at the 75th percentile to one at the 25th percentile, the marginal effect of the full treatment increases by $50 \times 0.00073 = 0.0365\sigma$. To more accurately model this non-linear relationship, we fit a piece-wise linear function with a structural break at the 60th percentile. At and below the 60th quantile the slope is 0.0003 and not statistically significant (p -value $=0.243$), while above the 60th quantile the slope is -0.00314 (p -value $=0.001$). The data indicate that for the bottom 60 percent of teachers, the marginal effect of the full treatment is 0.11σ , and that the full treatment is only ineffective for the most able teachers in the top ten percent of the effectiveness distribution. This is consistent with a model where off-the-shelf lessons and teacher quality are substitutes in the production of student outcomes such that they may be very helpful for the least effective teachers.

Given this decline, one may worry that the intervention might reduce effectiveness for high-quality classrooms. Indeed such patterns were observed for computer-aided instruction in Taylor (2015). However, even at the 95th and 99th percentile of classroom quality, the semi-parametric point estimates are positive (albeit not statistically different from zero). To ensure that these patterns are real, as a falsification exercise, we estimate the same quantile regression model for English test scores (see Appendix E). As one would expect, there is no systematic relationship for English

¹⁴To assuage concerns that the teacher-level model yields different results from the student-level model, Appendix F shows that the OLS test score regressions aggregated to the teacher level yield nearly identical results to those at the student level across all specifications and falsification tests.

scores, and the estimated point estimates for English are never statistically significantly different from zero at the ten percent level. This provides further evidence that the estimated effects on math scores are causal, and that the pattern of larger treatment effect for the less able teachers is real.

VI Model

The fact that providing online off-the-shelf lessons improved the outcomes of teachers on average, with very large benefits at the bottom of the teacher effectiveness distribution and small but positive effects at the top suggests that certain economic forces and mechanisms may be at play. To facilitate interpretation of our results, we lay out a stylized model of teacher behavior that yields these same predictions regarding the treatment effects on student outcomes. The model also yields some additional empirical predictions regarding the underlying mechanisms that we will test empirically. For ease of exposition, we provide an intuitive graphical presentation of the model below, and provide formal mathematical arguments in [Appendix G](#).

We model teachers as being akin to firms (consumers) choosing the ‘profit’ (utility) maximizing mix of inputs (goods) given a fixed total cost (budget) and fixed input prices (prices of goods). Teachers produce student test scores (y_i), where i is a student in the teacher’s class. Student i ’s test scores depend on the teacher’s allocation of time (T) between creating lessons (d) and all other tasks (n). Variable n captures all tasks complementary to creating lessons which include lesson delivery, classroom management, and others. For simplicity, student i ’s test scores take a form of a Cobb-Douglas function in d and n with homogeneous output elasticities and heterogeneous individual shocks. Teacher abilities to create lessons and perform all other tasks are modeled as input “prices” (p_d) and (p_n). These prices denote the amount of time needed to create one unit of lesson quality and produce one unit of other tasks, respectively. All teachers have the same time allocation (T), but higher ability teachers have lower p ’s such that they can produce more overall learning per unit of time.

Teachers maximize their students’ weighted average test scores by choosing how much time to spend on other tasks ($n \geq 0$) and how much time to spend on lesson creation ($d \geq 0$), subject to the time constraint (T) and the prices they face (p_n and p_d). To illustrate the model visually, the optimal allocation is depicted in Panel (a) of [Figure 1](#). The isocost curve is depicted by the straight line segment with slope $-p_n/p_d$, and different levels of average test scores are represented by different indifference curves. Higher average test scores are on higher indifference curves (up and to the right). At the optimal mix of inputs d^* and n^* , the teacher’s indifference curve (i.e.

average test scores) is tangent to the isocost curve such that test scores are maximized with this mix of inputs given the time constraints faced by the teacher. As we show in [Appendix G](#), at the output maximizing allocation, average test scores are decreasing in both p_n and p_d . That is, average test scores increase with teacher ability. We define d^* and n^* as the optimal allocation with no lessons.

We model off-the-shelf lessons as a technology that guarantees a minimum level of lesson quality (\underline{d}) at some fixed time cost F . Panel B of [Figure 1](#) depicts the scenario in which the fixed cost is equal to zero (i.e $F = 0$). Because teachers can always spend more time on lesson quality than the minimum, the new technology shifts the isocost curve up by \underline{d} . However, the maximum time allocation for other tasks is unchanged, so that the isocost curve is vertical at (T/p_n) . The teacher adopts the technology if the average test scores she can attain under this technology are greater than that without. We define \tilde{d} and \tilde{n} as the optimal allocation after adopting the lessons.

With no cost of lesson adoption, for teachers with equilibrium lesson quality above the minimum guaranteed by the lessons (i.e. $d^* > \underline{d}$), the lessons simply increase the total amount of time that can be spent on either more lesson planning or other tasks, leading to a positive ‘income’ effect. The new indifference curve IC_2 shows that a higher level of standard scores is achieved by using off-the-shelf lessons, with both \tilde{d} and \tilde{n} being higher than without lessons. For ease of exposition, we make the simplifying assumption that no teacher would find it optimal to locate at the kink with lesson use.¹⁵ That is, we make the realistic assumption that it will always be optimal to spend some of one’s own time planning lessons, even if the online lessons are high quality.

Now consider the case where the fixed cost of lesson adoption is non-zero. With some fixed time cost ($F > 0$) to adoption, the isoquant no longer shifts up by \underline{d} , but shifts up only by $\underline{d} - F/p_d < \underline{d}$. Specifically, the isoquant shifts out by the guaranteed lesson quality minus the loss in lesson quality associated with the time (F) spent adopting the lessons. It is straightforward to see graphically that the outward shift in the isoquant is smaller for high ability teachers (with low p_d and p_n) and larger for lower ability teachers (with higher p_d and p_n). Because the upward shift in the isoquant is larger for the low ability teachers, the potential benefits to lesson use are larger for the low ability teachers. Intuitively, the high ability teachers have a higher opportunity cost of time so that, for the same fixed time cost F , the high types lose more in potential student achievement than the low types. This differential shift and the corresponding larger benefit to lesson use for less able teachers are depicted in panels C (higher ability teacher) and D (lower ability teacher). Note that if the production function exhibits diminishing returns to scale, this would further increase the difference in benefits between high and low ability teachers.

¹⁵For teachers with equilibrium lesson quality $d^* < \underline{d}$, some will locate at the kink such that both d and n are higher with lesson use. A minimal assumption to prevent this behavior is listed in [Appendix G](#). However, for simplicity, one could also make a sufficient assumption that at the optimum without lessons all teachers are such that $d^* \geq \underline{d}$.

The simple framework depicted in [Figure 1](#) produces some useful predictions that can be tested empirically. The first, most obvious, prediction is that, among teachers who use the lessons, the optimal lesson quality (d) should increase. This occurs either because the lessons used are of higher quality than what the teacher would have created on her own (a direct lesson use effect), or because the time savings afforded by the lessons allows teachers to improve their lesson quality (a time savings effect). While it is obvious that lesson quality will increase if the online lessons are sufficiently high quality, the model highlights the fact that even lessons of modest quality can benefit even the most able teachers because of the time savings mechanism. The second, less obvious prediction from the model is that because one of the benefits of the lessons is time savings, teachers who adopt the lessons will spend more time doing other complementary tasks (n). The model also makes some predictions regarding the overall effect of lesson adoption and also regarding which teachers are most likely adopt lessons. In sum, the model yields the following four predictions:

Prediction 1: If teachers know their ability, among those who chose to adopt lessons voluntarily, the gains in average test scores from using the off-the-shelf lessons are non-negative.

Prediction 2: Among those who chose to adopt lessons, teacher time spent on n (that is, all tasks complementary to lesson planning) should increase.

Prediction 3: Among teachers who chose to adopt lessons, the effect on lesson quality d is positive.

Prediction 4: The gains to using off-the-shelf lessons are decreasing in teacher effectiveness (as measured by ability to raise average test scores).¹⁶

VII Mechanisms

Section [V](#) established that student achievement in mathematics improved in the license only condition relative to the control condition and with even larger improvements in the full treatment condition. In this section, we shed light on the underlying causal mechanisms, and empirically test the remaining predictions from the model.

¹⁶If one is willing to make the additional assumption that teachers have full information regarding their ability, another prediction from our model is that absent external nudges or reminders (i.e. the license only condition), less effective teachers (as measured by the ability to raise average test scores) will be more likely to use the lessons. We do not include this as a formal prediction because, due to data limitations, we are unable to convincingly test this prediction empirically. We discuss the data limitation in more detail in [Footnote 17](#).

VII.1 Effects on Mathalicious lesson use

The first mechanism we explore is the extent to which the test score effects are driven by increased Mathalicious lessons use. We have two sources of data to measure Mathalicious use, both of which are imperfect. First, we rely on self-reported measures of which Mathalicious lessons were taught or read. This information was reported by teachers during the mid-of-year and end-of-year surveys. As such, these data may suffer from bias due to survey non-response. Second, we use the data received from Mathalicious site logs on whether a teacher downloaded a certain lesson or not (based on login email). While the lessons downloaded measure is not subject to bias due to survey non-response, the download tracker may understate lessons downloaded for two reasons. First, the download tracker was not available for the first month of the experiment so that overall downloads are under-recorded. Second, the tracker was only able to track downloads for teachers that used their official public school email address. While teachers were urged to use their school email accounts, there was nothing preventing teachers from using their personal email accounts. With these imperfect sources of information on lesson use, we construct three measures: the number of Mathalicious lessons taught (as reported by the teacher across both surveys), the number of Mathalicious lessons the teacher looked at (either reported as taught, reported as read, or tracked as downloaded), and the the number of Mathalicious lessons tracked as downloaded. To gain a sense of whether teachers made use of the extra supports to facilitate Mathalicious lesson use, we also employ data on webinars attended in real-time. While the webinars were designed to facilitate real-time interaction among teachers and Mathalicious facilitators, they were recorded and made available for asynchronous viewing. As such, this measure may not capture the extent to which teacher *viewed* webinars, and may understate the extent to which teacher used these additional supports.

We analyze the effect of the treatment on these measures of use in [Table 3](#). Because our measures of lessons taught and viewed are (partially) obtained from survey data, we only have complete lesson use for teachers who completed the surveys during both waves. Because lesson use is zero in the control condition, imputing zero lesson use for those who did not fill in both the mid-year and the end-of-year surveys will mechanically lead to a downward bias for those in the partial or full treatment conditions. As such, in Panel A, we report estimated effects only on those teachers for whom we have complete survey data (i.e. data for both the mid-year and end-of-year surveys). Only 20 percent of all teachers have completed survey data in both waves so that the estimated effects are rather imprecise. However, the patterns are instructive and are robust across all measures of lesson use for which we have more data. All models include the full set of controls mentioned in [Section IV](#).

The point estimates in the top panel reveal that among teachers with complete survey data, those in the license only condition looked at 1.4 more lessons, taught 0.092 more lessons, downloaded 1.969 more lessons, and attended 0.168 more webinars than control teachers. None of these differences is significant at the ten percent level, but the magnitudes are instructive. Teachers in the full treatment condition looked at 5.1 more lessons, taught 2.28 more lessons, downloaded 3.699 more lessons, and attended 0.499 more webinars than teachers in the control condition. Only the effect on webinars attended is statistically significantly different from zero at the 5 percent level. While using teachers with complete data deals with downward bias due to survey-non-response, it may also introduce upward bias if those teachers who complete surveys tend to have higher levels of use than those who do not. We test for this formally in [Appendix H](#), where we show that conditional on treatment status, survey participation is unrelated to lessons downloaded. This would suggest that the results that use data only for teachers with complete surveys, while imprecise, should yield an accurate estimate of the effect of the treatment on lesson use.

To improve precision and use data obtained from a larger number of teachers, we also compute lesson use based on teachers that completed either the mid-year survey or the end of year survey. Because teachers who do not complete one of the surveys are automatically assigned zero use for *that survey wave*, these results are biased toward zero. As such, these estimates are likely to be lower in magnitude than the real effects. Panel B presents the estimated effects among the 60 percent of teachers with at least partially complete survey data (i.e. survey data in at least one of the two waves). Among teachers with partially complete survey data, teachers in the license only condition looked at 1.396 more lessons, taught 0.466 more lessons, downloaded 1.034 more lessons and attended no more webinars than teachers in the control condition. The effects on lessons looked at and lessons downloaded are significant at the 5 percent level. Consistent with the larger effect on test scores, the effects on use are larger in the full treatment condition. Teachers in the full treatment condition looked at 2.618 more lessons, taught 0.983 more lessons, downloaded 2.134 more lessons and attended 0.097 more webinars than teachers in the control condition. While the point estimates are smaller than results using the 20 percent of teachers with full data (as expected), all the marginal effects are meaningful and significant at the 5 percent level for the full treatment condition.

Because using teachers with partially complete survey data mechanically leads to a downward bias in our estimated effects on lesson taught (because missing data in non-completed surveys is assumed to be zero), we address the missing data problem more rigorously using multiple imputation ([Rubin, 2004](#); [Schafer, 1997](#)) to impute lesson use for those individuals who did not complete the surveys. Within each multiple imputation sample, we impute the missing numbers of lessons looked at and lessons taught using predicted values for other teachers in the same treatment condi-

tion from a Poisson regression (note that these are count data). Recall that [Appendix H](#) indicates that the lesson download behavior of teachers in the same treatment condition is unrelated to having complete survey data so that this imputation method is likely valid. For the lessons looked at, we conduct multiple imputation for the survey responses before combining it with the tracker data. The regression results based on imputed use (for missing data) are presented in Panel C of [Table 3](#). The results are very similar to those in Panel A that uses teachers with complete survey data. However, now all the differences between the treatment groups and the control group are statistically significant at the one percent level. These are our preferred estimates because they use data for the full sample of teachers. Note that standard errors are corrected for multiple imputation using the method in [Rubin \(2004\)](#). The point estimates indicate that teachers in the license only condition looked at 1.586 more lessons and taught 0.657 more lessons than teachers in the control condition, while teachers in the full treatment condition looked at 4.4 more lessons and taught 1.925 more lessons than teachers in the control condition.¹⁷

Across all models, teachers who received only the Mathalicious licenses looked at and taught more lessons than control teachers, while teachers who received the full treatment looked at and taught more lessons than either the control teachers or those who received licenses only. To avoid overstating the effects of lesson use on outcomes, it is important that we do not understate the increases in lessons use. Accordingly, we take the conservative approach and focus on the larger and more credible estimates in Panels A and C that rely on teachers with complete data on lesson use. To put these estimates into perspective, each Mathalicious lesson provides intuition for topics that span between 3 and 6 weeks. As such, teachers in the full treatment looked at Mathalicious lessons that could impact about one-half of the school year and report teaching lessons that could impact about two-thirds of the school year. Accordingly, while the full treatment group never reached full fidelity with the Mathalicious model (which is between 5 and 7 lessons per year), the increased lesson use likely translated into changes in instruction for a sizable proportion of the school year. In [Section VII.4](#), we present evidence on why the usage may not have been as widespread.¹⁸ Another noteworthy result is that the attendance at webinars was very low in the full

¹⁷Finally, as a robustness check of the reliability of the survey data, we also analyze results only using the online tracker data on the full sample. Any missing data is imputed to be zero, so that these are lower bound estimates. These results are presented in Panel D. Teachers in the license only condition looked at at least 1.115 more lessons and downloaded at least 0.916 more lessons than those in the control condition. Both effects are significant at the 5 percent level. Teachers in the full treatment condition looked at at least 2.236 more lessons and downloaded at least 1.900 more lessons than those in the control condition. Importantly, both of these differences is statistically significant at the 1 percent level.

¹⁸ One prediction from our model is that absent external incentives, *if teachers are aware of their own ability*, lesson adoption should be highest among the least effective teachers. Unfortunately, we cannot test this convincingly due to data limitations. As discussed previously, in order to test for treatment effect heterogeneity by teacher effectiveness, one would ideally employ historical data on teachers (and their students) and then interact the treatment with measures of teacher effectiveness obtained out of sample. Because we only have one year of achievement data linked to lagged

treatment condition even though lesson use was higher. This suggests that the increased use in the full treatment condition was not driven by the additional supports *per se*, but may have been driven by the regular reminders to use the lessons. We present suggestive evidence of this in Section VII.4.

VII.2 Effects on Student Perceptions and Attitudes

The specific aims of the Mathalicious lessons were to promote deeper student understanding, instill a sense that math has real world applications, and develop greater student interest and engagement in the subject. As such, by changing the lessons teachers deliver, the intervention lessons could alter student attitudes toward mathematics. To test this, we analyze effects on student responses on an anonymous survey given at the end of the Fall semester (December) and also at the end of the experiment (May). These survey responses cannot be linked to individual students, but are linked to the math teacher. Due to permission restrictions, these survey data were collected for Chesterfield and Hanover only. On the surveys, we asked several questions on a Likert scale and used factor analysis to extract common variation from similar items. After grouping similar questions, we ended up with 6 distinct factors.¹⁹ Each factor is standardized to be mean zero, unit variance.

Teachers are only partially treated at the time of the mid-year survey, while responses at the end of the year reflect exposure to the intervention for the full duration. To account for this, among those in the license only treatment, we code the variable $License_{dt}$ to be 1 during the end-of-year survey and 1/3 in the mid-year survey. Similarly, among those in the full treatment, we code the variable $Full_{dt}$ to be 1 during the end-of-year survey and 1/3 in the mid-year survey.²⁰ Using data from both surveys simultaneously, we estimate the effect on student responses to the survey items using the following equation, where all variables are defined as in (1) and $Post_{idt}$ is an indicator that is equal to 1 for the end-of-year survey and zero otherwise.

$$Y_{idt} = \alpha_d + \beta_1 License_{dt} + \beta_2 Full_{dt} + X_{dt} \delta_d + \pi_d Req_{dt} + \gamma Post_{idt} + \varepsilon_{idt} \quad (3)$$

As with test scores, we analyze the student surveys at the student level. Table 4 presents results from models that include no controls (Panel A) and models that include the full set of controls

outcomes, this is not feasible. We are able to get around this data limitation by exploiting the specific interpretation of conditional quantile regression models when testing for achievement effects (as shown in Appendix D). However, the conditional quantile models applied to lesson use do not have the same interpretation. As such, we are unable to provide any credible empirical evidence on this prediction.

¹⁹To avoid any contamination associated with the treatments, we only used data for the control group in forming the factors. When grouping questions measuring the same construct, each group is explained by only one underlying factor. Factor loadings for each individual question are presented in Appendix C.

²⁰Note that our results are robust to using fractions of similar magnitude, e.g., 1/2 or 1/4.

(Panel B).

In order for the estimation to be credible, it requires that the survey response rates are similar across all treatment arms. As such, the first column is a model where the dependent variable is the survey response rate computed at the teacher level.²¹ The analytic sample in this model is all students in the testing file (irrespective of whether they completed a survey) in the two participating districts. Overall, the survey response rate was 66 percent. Importantly, there are no statistically significant differences in survey response rates across the three treatment arms. In fact, the model with no controls, the survey response rate is slightly *lower* in the treatment arms than in the control group, while in the model with full controls the survey response rate is slightly *higher* in the treatment arms than in the control group. Moreover, the estimated treatment effects on the survey questions are similar in models with and without controls (for which the direction of the response rates are opposite in sign), so that any differences in response to questions are not likely driven by differential non-response.

Because the estimated effects on the factors are so similar in models with and without controls, we focus our discussion on the models with all controls (Panel B). The first factor measures whether students believe that math has real life applications. The results in Column 9 of [Table 4](#) show that, while there is no effect for the license only condition, students in the full treatment condition are more likely to report that math has real life application than students in the control group. Specifically, students of the full treatment teachers agree that math has real world applications 0.162σ more than those of control teachers ($p\text{-value} < 0.05$). This is consistent with the substance and stated aims of the Mathalicious lessons and confirms our priors that their content was more heavily grounded in relevant real-world examples than what teachers would have been teaching otherwise.

The next three factors measure student interest in math class, effort in math class, and motivation to study in general, respectively. Even though none of these are directly targeted by the intervention, the lessons may increase interest in math, and such benefits could spill over into broad increases in academic engagement. There is weak evidence of this. Students with full treatment teachers report meaningfully higher levels of interest in math (0.087σ). However, this effect is not statistically significant at traditional levels. The estimated coefficient on effort in math class is 0.045σ for the license only condition and a zero for the full treatment condition. In the full treatment, there is a small positive effect on the general motivation to study and a small negative effect on motivation to study in the license only condition. None of the effects on these three factors are

²¹For each teacher we use the test score data to determine how many students could have completed a survey. We then compute the percentage of students with completed surveys for each teacher and weight the regressions by the total number of students with the teacher.

statistically significant, but the magnitudes and direction of the estimates are suggestive.

The next two factors relate to student perceptions of their math teacher and allow us to test two of the predictions from the model. The fifth factor measures whether students believe their math teacher emphasizes deep understanding of concepts. This relates directly to the specific aims of the Mathalicious lessons. The model predicts that the optimal lesson quality should increase under the treatment so that we should see increases in agreement with statements regarding the teacher promoting deeper understanding. The sixth factor measures whether students feel that their math teacher gives them individual attention. Our model predicts that off-the-shelf lessons may free up teacher time toward other tasks that are complementary to lesson planning. There are many such tasks, but we hypothesize that providing one-on-one time is one. As such, one would expect that the additional time afforded by the lessons may allow teachers to provide students with more one-on-one instruction.²² The results support the premise of our model that teacher who used the Mathalicious lessons improved lesson quality. Students from the full treatment group are 0.175σ ($p\text{-value}<0.05$) more likely to agree that their math teacher promotes deep understanding. Also, consistent with off-the-shelf lessons freeing up teacher time to exert more effort in the classroom toward other complementary tasks, student agreement with statements indicating that their math teacher spends more one-on-one time with them is 0.033σ higher in the license only treatment condition ($p\text{-value}>0.1$) and 0.144σ higher in the full treatment condition than in the control condition ($p\text{-value}<0.05$).

In sum, the survey evidence shows that, among students whose teachers used the Mathalicious lessons most robustly (i.e. full treatment teachers), student perceptions regarding math and their math teachers changed in the expected directions. However, we do not find strong evidence of effects on these outcomes among students in the license only condition. This may either reflect no movement on these survey measures in the license only condition or that effects of the license only condition that are too small to detect. In any case, students of teachers in the full treatment say that there are more real life applications of math, and report somewhat higher levels of interest in math class. Moreover, they report that their teachers promote deep understanding and spend more one-on-one time with students. These patterns are consistent with the aims of the intervention, are consistent with some of the key predictions of the model, and are consistent with the pattern of positive test score effects.²³

²²Jackson (2016a) also uses more one-on-one time as a measure of teacher time. He finds that in more homogeneous classrooms, teachers spend more one-on-one time with students likely due to time savings.

²³We also analyze teachers' survey responses to assess whether the intervention had any effect on teachers' attitudes toward teaching, or led to any changes in their classroom practices. Although the response rate on the teacher survey was similar to that of the student surveys (61.43 percent), the response rates were substantially higher among teachers in the full treatment condition. As such, the results on the teacher surveys are inconclusive. Moreover, we do not find any systematic effects on any of the factors based on the teacher survey items. We present a detailed discussion of the

VII.3 Are the Effects Driven By Lesson Use *Per Se*?

The full treatment, which involved both lesson access and additional supports, led to the largest improvement in test scores. The extra supports were not general training, but were oriented toward implementing specific Mathalicious lessons. As such, it is unlikely that the gains were driven by the extra supports and not the lessons themselves. The fact that we find meaningful positive effects in the license only condition confirms that this is the case. Also, the fact that webinar attendance was so low overall suggests that many teachers in the full treatment were not using the additional online supports. The evidence presented thus far suggests that the improvements are due to lesson use rather than the extra supports, but we present more formal tests of this possibility in this section.

If the benefits of the intervention were driven by Mathalicious lesson use, then those treatments that generated the largest increases in lesson use should also have generated the largest test score increases. To test for this, using our preferred student level models, we estimate the effects of each treatment arm (license only or full) in each of the three districts (i.e. six separate treatments) relative to the control group in each district.²⁴ Figure 3 presents the estimated effects on lessons taught against the estimated effects on math test scores for each of the six treatments. Each data point is labeled with the district and the treatment arm (1 denotes the license only treatment and 2 denotes the full treatment). It is clear that the treatments that generated the largest increases in lesson use were also those that generated the largest test score gains. There is a very robust positive linear relationship. To test more formally whether the extra supports provided in the full treatment explain the pattern of treatment effects, we estimate a regression line through these 6 data points predicting the estimated test score effect using the estimated effect on lessons taught and an indicator for whether the treatment arm was the full treatment. In this model, conditional on the treatment type, the estimated slope for lessons taught on test scores is 0.047 (p-value<0.01). To use this variation more formally, we estimate instrumental variables models predicting student math test scores and using the individual treatment arms as instruments for lessons taught (detailed in Appendix J). The preferred instrumental variables regression model yields a coefficient on lessons taught of 0.033, suggesting that for every additional lesson taught test scores increase by 0.033σ . Importantly, in this model, one cannot reject the null hypothesis that the marginal effect of the full treatment is zero conditional on the lessons taught effect. These patterns indicate that (a) those treatments with larger effects on lesson use had larger test score gains and (b) the reason the full treatments had a larger effect on test scores is that they had a larger effect on lesson use.

teacher survey results in Appendix I.

²⁴When estimating effects on lessons taught, we use multiple imputation as outlined in Section V.

VII.4 Patterns of Lesson Use Over Time

Given the sizable benefits to using the off-the-shelf lessons, one may wonder why lesson use was not even more widespread. To gain a sense of this, we present some graphical evidence of lesson use over time. [Figure 4](#) shows the number of lessons downloaded by license only and full treatment groups in different months. As expected, lesson use was much larger in the full treatment condition than that in the license only condition. However, [Figure 4](#) reveals a few other interesting patterns. There was a steady decline in the number of lessons downloaded over time within groups. While there were 97 downloads in the full treatment in November 2014, there were only 8 downloads in May 2015. Similarly, in the license only group, while there were 59 downloads in the November 2014, there were only 4 downloads in May 2015. To determine whether this decline is driven by the same number of teachers using Mathalicious less over time, or a decline in the number of teachers using Mathalicious over time, we also plot the number of teachers downloading lessons by treatment group over time. There is also a steady decline in the number of teachers downloading lessons so that the reduced use is driven by both reductions in downloads among teachers, and a reduction in the number of teachers downloading lessons over time.

Even though we have no dispositive evidence on why lesson use was not higher, or why lesson use dropped off over time, we speculate that it may have to do with behavioral biases and time management. The patterns of attrition from lesson downloads over time are remarkably similar to the patterns of attrition at online courses ([Koutropoulos et al., 2012](#)), gym attendance ([DellaVigna and Malmendier, 2006](#)), and fitness tracker use ([Ledger and McCaffrey, 2014](#)). Economists hypothesize that such behaviors may be due to individuals underestimating the odds that they will be impatient in the future and then procrastinate ([O'Donoghue and Rabin, 1999](#); [Duflo et al., 2011](#)). Similar patterns in [Figure 4](#) provide a reason to suspect that similar behaviors may be at play. In our context, these patterns may reflect teachers being optimistic about their willpower to use the lessons such that they started out strong, but when the time came, they procrastinated and did not make the time to implement them later on. However, it is also possible that as teachers use the lessons, they perceive that they are not helpful and decide to discontinue their use after downloading the first few lessons. Most of the empirical patterns support the former explanation. First, the rate of decay of lesson use is more rapid in the license only treatment than in the full treatment group. Specifically, without the additional supports to implement the lessons, the drop-off in lesson use was more rapid. In the full treatment group, downloads fell by about 45 percent between Nov/Dec and January/Feb, while it fell by over 80 percent during that same time period in the license only group. If the reason for the drop-off was low lesson quality, drop-off should have been similarly rapid for both groups. The second piece of evidence is that there is a sizable reduction in lessons downloaded in the

full treatment condition after February when Mathalicious ceased sending out email reminders to teachers, while lesson use was stable in the license only condition. The third piece of evidence comes from surveys. We employed data from the end of year survey that asked treated teachers why they did not use off-the-shelf lessons more. Looking specifically at the question of whether the lessons were low quality, only 2 percent of teachers mentioned this was a major factor and almost 89% stated that it was not a factor at all. In sum, poor lesson quality does not explain the drop-off in lesson use, being reminded mattered, and the patterns of drop-off are very similar to other contexts in which behavioral biases played a key role – suggesting that procrastination is a plausible explanation.

The last piece of evidence to support the procrastination hypothesis also comes from the survey evidence shown in [Figure 5](#). The main reason cited for not using more lessons was a lack of time. Taken at face value, one might argue that the pressures on teacher time increased over the course of the year such that lesson use declined over time. However, this cannot explain the large differences in the trajectory of lesson use over time across the treatment arms. The explanation that best fits the observed patterns and the survey evidence is that, without the reminders and extra supports (i.e. Edmodo groups), teachers were unable to hold themselves to make the time to implement the lessons. The patterns also suggest that providing ways to reduce procrastination during the school year (such as sending constant reminders or providing some commitment mechanism) may be fruitful ways to increase lesson use. Other simple approaches may reduce the incentive to procrastinate at the moment by providing designated lesson planning time, or granting lesson access the summer before the school year when the demands on teachers' time may be lower.

VIII Discussion and Conclusions

Teaching is a complex job that requires that teachers perform several complementary tasks. One important task is planning lessons. In the past few years, the availability of lesson plans and instructional material for use in the traditional classroom that can be downloaded from the internet has increased rapidly. Today over 90 percent of secondary teachers look to the Internet for instructional materials when planning lessons [Opfer et al. \(2016\)](#) and lesson warehouse sites such as [Teachers Pay Teachers](#) have more active user accounts than teachers in the United States. Teacher use of these online lessons is a high-tech form of division of labor; classroom teachers focus on some tasks while creating instructional content is (partially) performed by others. If this technological change now provides all teachers access to high-quality lessons, the social benefits could be very large. However, the extent to which providing teachers access to high-quality online instructional

materials improves their student's performance had not previously been explored. To speak to this question, we implemented a randomized field experiment in which middle-school math teachers in three school districts were randomly provided access to high-quality, off-the-shelf lessons, and we examine the effects on their students' subsequent academic achievement.

The online "off-the-shelf" lessons provided in our intervention were not typical of ordinary mathematics lesson plans. The off-the-shelf lessons were experiential in nature, made use of real-world examples, promoted inquiry-based learning, and were specifically designed to promote students' deep understanding of math concepts. Though education theorists hypothesize that such lessons improve student achievement, this is among the first studies to test this idea experimentally.

Offering the lessons for free had modest effects on lesson use and modest (but economically meaningful) effects on test scores (0.06σ). However, fully-treated teachers (who also received online supports to promote lesson use) used the lessons more and improved their students' test scores by about 0.09σ relative to teachers in the control condition. These positive effects appear to have been mediated by students feeling that math had more real life applications, and having deeper levels of understanding. There is also evidence that as teachers substituted the lessons for their own lesson planning efforts, they were able to spend more time on other tasks such as providing one-on-one time with students. Consistent with our multitask model of teaching, the positive test score effects are largest for the weaker teachers. We hypothesize that the relatively low levels of lesson use may be due to behavioral biases among teachers such that they put off taking the time to implement the lessons until it is too late (i.e. they may procrastinate). Our findings imply that regular reminders and supports were helpful to keep teachers engaged enough to implement the lesson through the school year.

Because the lessons and supports were all provided online, the per teacher costs of the intervention are low. An upper bound estimate of the cost of the program is \$431 per teacher.²⁵ Chetty et al. (2014a) estimate that a teacher who raises test scores by 0.14σ generates marginal gains of about \$7,000 per student in present value future earnings. Using this estimate, the test score effect of about 0.09σ would generate roughly \$4,500 in present value of future earnings per student. While this may seem like a modest benefit, consider that each teacher has about 90 students in a given year so that each teacher would generate \$405,000 in present value of students' future earnings. This implies a benefit-cost ratio of 939. Because of the low marginal cost of the intervention, it is extraordinarily cost effective. Furthermore, because the lessons and supports are provided on

²⁵The price of an annual Mathalicious subscription is \$320. The cost of providing the additional supports (e.g., extra time for Mathalicious staff time to run Project Groundswell) was \$25,000. With 225 treated teachers, this implies an average per teacher cost of \$431. Because the subscription partly recovers fixed costs, the marginal cost is lower than this. One can treat this as an upper bound of the marginal cost.

the Internet, the intervention is highly scalable and can be implemented in remote locations where other policy approaches would be infeasible.

Our findings show that by providing teachers with access to high-quality, off-the-shelf lessons on the Internet is a viable and cost-effective alternative to the typical policies that seek to improve the skills of the existing stock of teachers through training, selection, or changes in incentives (e.g. [Taylor and Tyler, 2012](#); [Muralidharan and Sundararaman, 2013](#); [Rothstein and others, 2015](#)). Our findings also suggest that policies aiming to modify the production technology of teaching (such as changes in curriculum design, innovative instructional materials, and others) may be fruitful avenues for policymakers to consider.

References

- A. Akerman, I. Gaarder, and M. Mogstad. The Skill Complementarity of Broadband Internet. *The Quarterly Journal of Economics*, 130(4):1781–1824, 2015.
- N. Anderson, D. S. Ones, H. K. Sinangil, and C. Viswesvaran. *Handbook of Industrial, Work & Organizational Psychology: Volume 1: Personnel Psychology*. Sage, 2001.
- J. Angrist and V. Lavy. New evidence on classroom computers and pupil learning. *The Economic Journal*, 112(482):735–765, 2002.
- M. C. Araujo, P. Carneiro, Y. Cruz-Aguayo, and N. Schady. Teacher Quality and Learning Outcomes in Kindergarten. *The Quarterly Journal of Economics*, 131(3):1415–1453, 2016.
- A. V. Banerjee, S. Cole, E. Duflo, and L. Linden. Remedying Education: Evidence from Two Randomized Experiments in India. *The Quarterly Journal of Economics*, 122(3):1235–1264, 2007.
- L. Barrow, L. Markman, and C. E. Rouse. Technology’s Edge: The Educational Benefits of Computer-Aided Instruction. *American Economic Journal: Economic Policy*, pages 52–74, 2009.
- D. W. Beuermann, J. Cristia, S. Cueto, O. Malamud, and Y. Cruz-Aguayo. One Laptop per Child at Home: Short-Term Impacts from a Randomized Experiment in Peru. *American Economic Journal: Applied Economics*, 7(2):53–80, 2015.
- H. S. Bloom, L. Richburg-Hayes, and A. R. Black. Using Covariates to Improve Precision: Empirical Guidance for Studies That Randomize Schools to Measure the Impacts of Educational Interventions. *MDRC Working Papers on Research Methodology*, 2005.
- N. Bloom, C. Genakos, R. Sadun, and J. Van Reenen. Management Practices Across Firms and Countries. *The Academy of Management Perspectives*, 26(1):12–33, 2012.

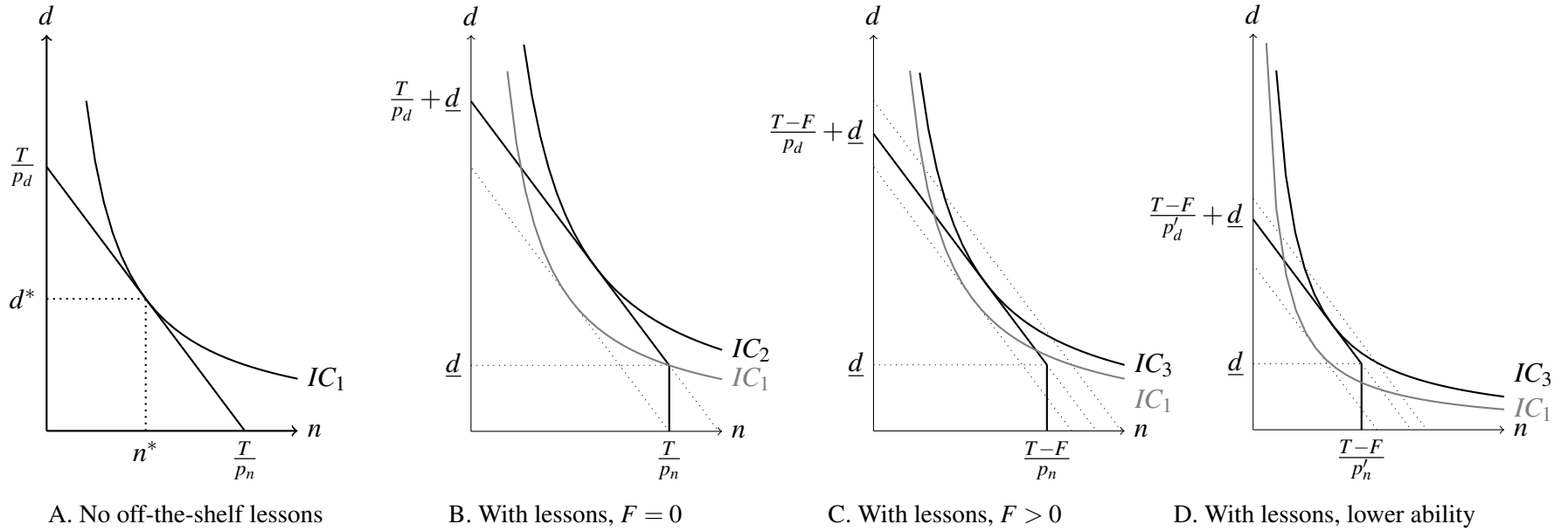
- C. C. Bonwell and J. A. Eison. *Active Learning: Creating Excitement in the Classroom*. 1991 ASHE-ERIC Higher Education Reports. ERIC, 1991.
- J. S. Brown, A. Collins, and P. Duguid. Situated cognition and the culture of learning. *Educational Researcher*, 18(1):32–42, 1989.
- M. Buchinsky. Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of Human Resources*, pages 88–126, 1998.
- G. Bulman and R. W. Fairlie. Technology and Education: Computers, Software, and the Internet. *NBER Working Paper w22237*, May 2016.
- R. Chetty, J. N. Friedman, and J. E. Rockoff. Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*, 104(9):2593–2632, 2014a.
- R. Chetty, J. N. Friedman, and J. E. Rockoff. Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, 104(9):2633–2679, 2014b.
- M. M. Chingos and G. J. Whitehurst. Choosing Blindly: Instructional Materials, Teacher Effectiveness, and the Common Core. *Brookings Institution*, 2012.
- S. Comi, M. Gui, F. Origo, L. Pagani, and G. Argentin. Is it the way they use it? Teachers, ICT and student achievement. *DEMS Working Paper No. 341*, June 2016.
- L. Darling-Hammond, R. C. Wei, A. Andree, N. Richardson, and S. Orphanos. Professional learning in the learning profession. 2009.
- S. DellaVigna and U. Malmendier. Paying not to go to the gym. *The American Economic Review*, pages 694–719, 2006.
- D. Deming. Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics*, pages 111–134, 2009.
- J. Dostál. *Inquiry-based instruction : Concept, essence, importance and contribution*. Palacky University Olomouc, 2015.
- E. Duflo, M. Kremer, and J. Robinson. Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya. *The American Economic Review*, pages 2350–2390, 2011.
- R. G. Fryer. The 'Pupil' Factory: Specialization and the Production of Human Capital in Schools. *NBER Working Paper w22205*, 2016.
- J. J. Heckman and D. V. Masterov. The productivity argument for investing in young children. *Applied Economic Perspectives and Policy*, 29(3):446–493, 2007.
- B. Holmstrom and P. Milgrom. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, pages 24–52, 1991.

- C. K. Jackson. The Effect of Single-Sex Education on Academic Outcomes and Crime: Fresh Evidence from Low-Performing Schools in Trinidad and Tobago. *NBER Working Paper w22222*, May 2016a.
- C. K. Jackson. What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes. *NBER Working Paper w22226*, May 2016b.
- C. K. Jackson and H. S. Schneider. Checklists and Worker Behavior: A Field Experiment. *American Economic Journal: Applied Economics*, 7(4):136–168, 2015.
- C. K. Jackson, J. E. Rockoff, and D. O. Staiger. Teacher Effects and Teacher-Related Policies. *Annual Review of Economics*, 6(1):801–825, 2014.
- C. K. Jackson, R. C. Johnson, and C. Persico. The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms. *The Quarterly Journal of Economics*, 131(1):157–218, 2016.
- B. A. Jacob and J. E. Rockoff. *Organizing schools to improve student achievement: Start times, grade configurations, and teacher assignments*. Brookings Institution, Hamilton Project, 2011.
- T. J. Kane and D. O. Staiger. Estimating teacher impacts on student achievement: An experimental evaluation. *NBER Working Paper w14607*, Dec. 2008.
- L. F. Katz and others. Changes in the wage structure and earnings inequality. *Handbook of Labor Economics*, 3:1463–1555, 1999.
- R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, pages 33–50, 1978.
- A. Koutropoulos, M. S. Gallagher, S. C. Abajian, I. de Waard, R. J. Hogue, N. O. Keskin, and C. O. Rodriguez. Emotive vocabulary in MOOCs: Context & participant retention. *European Journal of Open, Distance and E-Learning*, 15(1), 2012.
- J. Lave and E. Wenger. *Situated learning: Legitimate peripheral participation*. Cambridge University Press, 1991.
- D. Ledger and D. McCaffrey. Inside wearables: How the science of human behavior change offers the secret to long-term engagement. *Endeavour Partners*, 2014.
- R. Lesh and H. Doerr. Foundations of a model and modeling perspective on mathematics teaching, learning, and problem solving. 2003.
- M. Madda. Amazon Launches 'Inspire,' a Free Education Resource Search Platform for Educators. *EdSurge*, June 2016.
- K. Mihaly, D. F. McCaffrey, D. O. Staiger, and J. Lockwood. A composite estimator of effective teaching. *Seattle, WA: Bill & Melinda Gates Foundation*, 2013.
- K. Muralidharan and V. Sundararaman. Contract teachers: Experimental evidence from India. *NBER Working Paper w19440*, Sept. 2013.

- T. O'Donoghue and M. Rabin. Doing it now or later. *American Economic Review*, pages 103–124, 1999.
- V. D. Opfer, J. H. Kaufman, and L. E. Thompson. Implementation of K–12 State Standards for Mathematics and English Language Arts and Literacy. Technical report, RAND Corporation, 2016.
- R. C. Pianta. Teaching Children Well: New Evidence-Based Approaches to Teacher Professional Development and Training. *Center for American Progress*, 2011.
- J. L. Pierce, I. Jussila, and A. Cummings. Psychological ownership within the job design context: revision of the job characteristics model. *Journal of Organizational Behavior*, 30(4):477–496, 2009.
- S. G. Rivkin, E. A. Hanushek, and J. F. Kain. Teachers, schools, and academic achievement. *Econometrica*, pages 417–458, 2005.
- P. R. Rosenbaum. *Observational studies*. Springer, 2002.
- J. Rothstein and others. Teacher Quality Policy When Supply Matters. *American Economic Review*, 105(1):100–130, 2015.
- C. E. Rouse and A. B. Krueger. Putting computerized instruction to the test: a randomized evaluation of a “scientifically based” reading program. *Economics of Education Review*, 23(4): 323–338, 2004.
- D. B. Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.
- R. K. Sawyer. *The Cambridge handbook of the learning sciences*. Cambridge University Press, 2005.
- J. L. Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997.
- J. W. Stigler, P. Gonzales, T. Kwanaka, S. Knoll, and A. Serrano. The TIMSS Videotape Classroom Study: Methods and Findings from an Exploratory Research Project on Eighth-Grade Mathematics Instruction in Germany, Japan, and the United States. A Research and Development Report. 1999.
- E. S. Taylor. New Technology and Teacher Productivity. 2015.
- E. S. Taylor and J. H. Tyler. The effect of evaluation on teacher performance. *The American Economic Review*, 102(7):3628–3651, 2012.
- P. E. Todd and K. I. Wolpin. On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113(485):3–33, 2003.

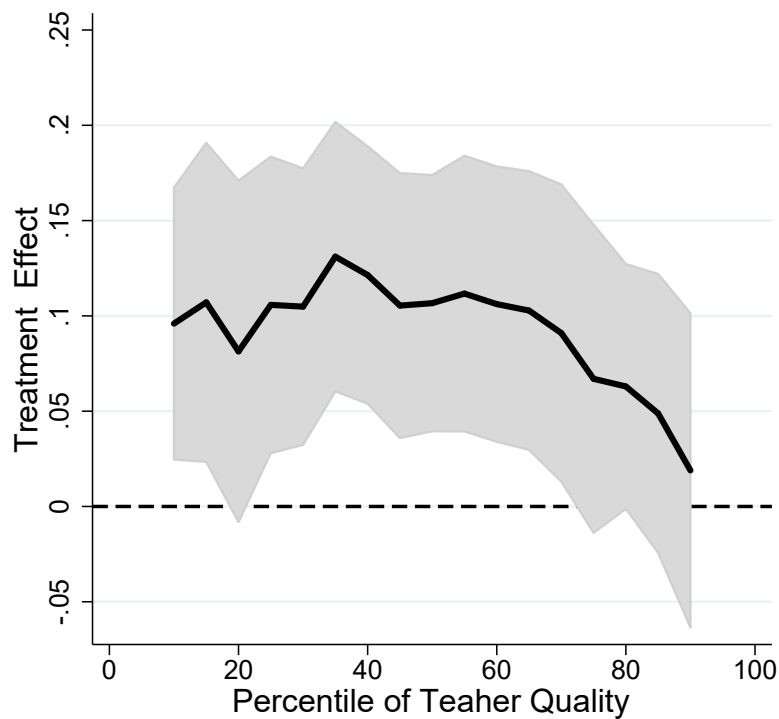
Tables and Figures

Figure 1: Illustration of the Model.



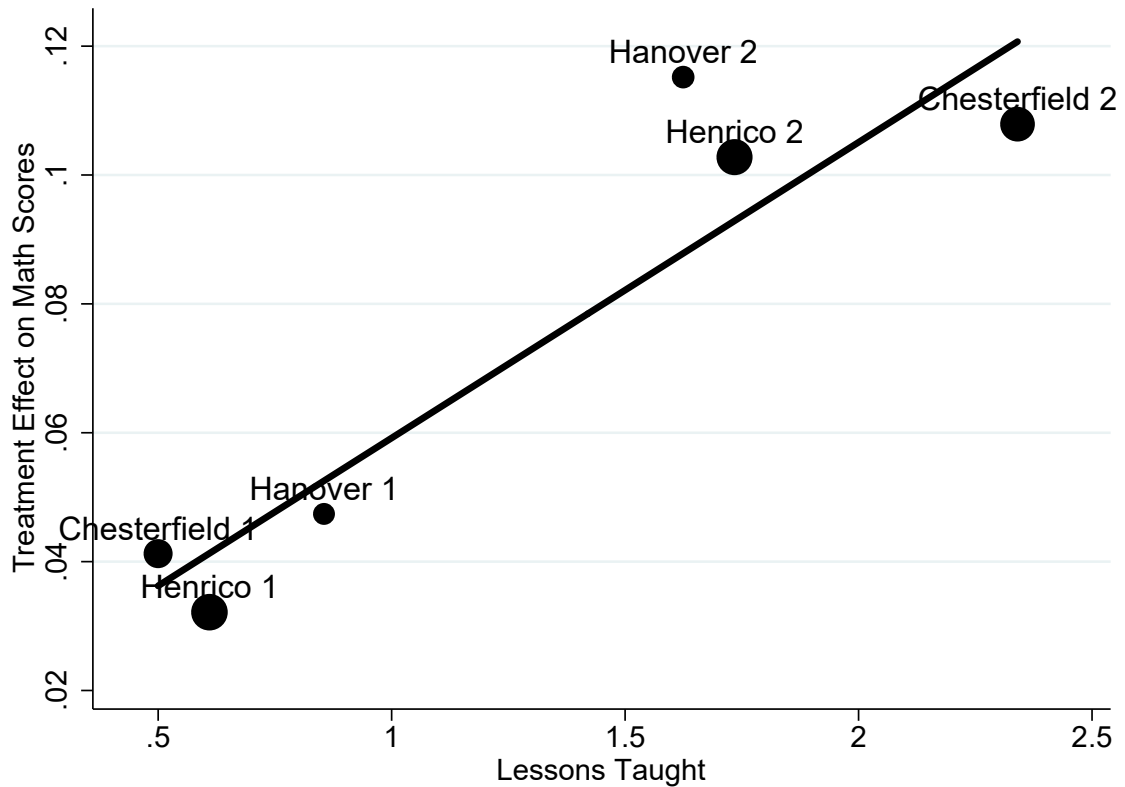
Notes: This is an Illustration of the stylized model presented in Section G2.

Figure 2. Marginal Effect of the Full Treatment by Teacher Quality.
Mathematics Test Scores.



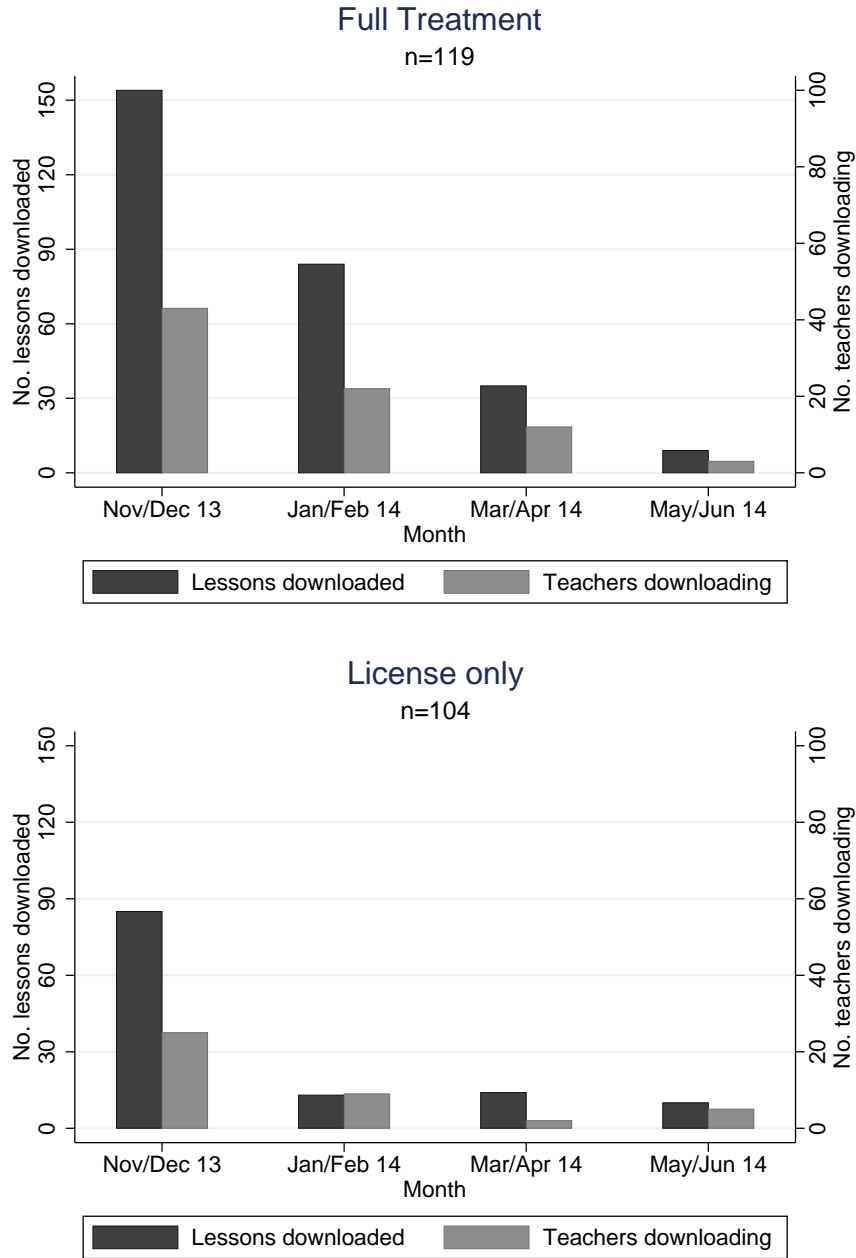
Notes: The solid black line represents the treatment effect estimates from estimating equation (1) using conditional quantile regression. The dependent variable is the teacher-level average standardized 2014 math test scores. The shaded area depicts the 90% confidence interval for each conditional quantile regression estimate. For a formal discussion of the method, see [Appendix D](#). The specification includes controls for the requested indicator, average teacher-level 2013 math and reading test scores, and teacher-level shares of students with missing 2013 math and reading test scores - all interacted with district fixed effects. Other controls include teachers' education level, years of experience, sex, race, grade fixed effects, as well as the percentage of male, black, white, Asian, and Hispanic students in their class.

Figure 3. Estimated Effect on Math Test Scores by Estimated Effect on Lessons Taught



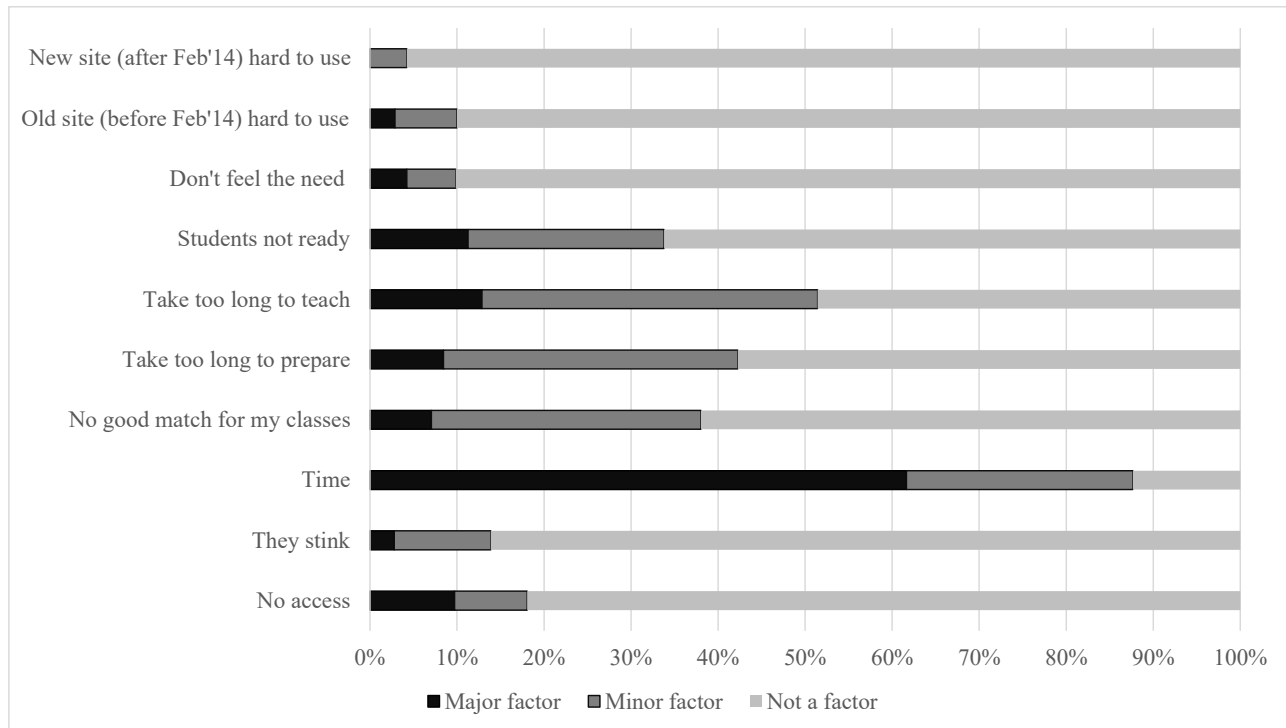
Notes: This figure plots average treatment effects on lesson use and standardized math scores, separately by district and by treatment. Chesterfield, Hanover, and Henrico are the school districts in Virginia where the intervention took place. The ‘License only’ treatment is denoted by the number 1, and the ‘Full Treatment’ is denoted by the number 2. The Y-axis displays coefficients for specifications identical to those estimated in Columns (5) of Table 2. The X-axis displays coefficients for specifications similar to those estimated in Panel C Column (10) of Table 3. However, all regressions are estimated based on a restricted sample within each district that compares each treatment group to the control group in the same district. For example, the ‘Chesterfield 1’ label means that the corresponding point displays the coefficients from the aforementioned regressions estimated within Chesterfield only and without the ‘Full Treatment’ teachers. The black line represents the best linear prediction based on six points displayed on each graph. The size of the dots corresponds to the relative size of the district-treatment groups in terms of the number of students.

Figure 4. Downloads of Mathalicious Lessons Over Time



Notes: Data on lesson downloads come from the teachers' individual accounts on the Mathalicious website. Mathalicious ceased to send out email reminders to teachers in the Full Treatment group after February 2014.

Figure 5. Reasons for Lack of Mathalicious Lesson Use.
License Only and Full Treatment Teachers Combined (n=71).



Notes: Data come from teacher responses to the following question on an end-of-year teacher survey: ‘Which of the following kept you from teaching a Mathalicious lesson this year?’. There were 10 reasons provided as non-mutually exclusive options. We report the percentage of completed responses that cite each of the 10 reasons. We combine the responses of both treatments in a single figure because the patterns are very similar in the license only and full treatment conditions.

Table 1. Summary Statistics.

Variable	N	Mean	SD	Mean (Control)	Mean (License Only)	Mean (Full Treatment)	P-value for balance hypothesis (w/district Fixed Effects and Requested)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
Teachers' characteristics	Has MA degree	363	0.424	0.495	0.386	0.433	0.462	0.767
	Has PhD degree	363	0.008	0.091	0.007	0.010	0.008	0.863
	Teacher is female	363	0.802	0.399	0.793	0.769	0.840	0.852
	Years teaching ^a	363	11.730	8.628	12.150	11.130	11.750	0.425
	Teacher is white	363	0.884	0.320	0.879	0.865	0.908	0.622
	Teacher is black	363	0.096	0.296	0.114	0.096	0.076	0.745
	Grade 6	363	0.311	0.464	0.300	0.240	0.387	0.503
	Grade 7	363	0.366	0.482	0.343	0.413	0.353	0.169
	Grade 8	363	0.342	0.475	0.321	0.356	0.353	0.746
	Participation across webinars	363	0.014	0.117	0	0	0.042	0.005***
	Total no. Mathalicious lessons the teacher taught	236 ^b	0.818	2.123	0.275	0.750	1.519	0.053*
	Total no. Mathalicious lessons the teacher taught or read	236 ^b	1.030	2.884	0.275	0.853	2.078	0.034**
	Total no. Mathalicious lessons the teacher downloaded	363	1.132	3.221	0.064	1.173	2.353	0.004***
Total no. Mathalicious lessons the teacher downloaded, read, or taught	256 ^c	2.184	4.458	0.337	2.107	4.157	0.001***	
Students' chars (student level)	Student is male	27613	0.516	0.074	0.515	0.519	0.513	0.798
	Student is black	27613	0.284	0.249	0.293	0.300	0.259	0.652
	Student is white	27613	0.541	0.261	0.534	0.535	0.553	0.588
	Student is Asian	27613	0.054	0.063	0.055	0.046	0.059	0.044**
	Student is Hispanic	27613	0.083	0.078	0.081	0.078	0.089	0.395
	Student is of other race	27613	0.036	0.025	0.034	0.036	0.037	0.209
	Math SOL scores, standardized by exam type, 2013	24112 ^d	0.0521	0.979	0.037	0.043	0.076	0.644
	Math SOL scores, standardized by exam type, 2014	27613	-0.002	1.001	-0.071	-0.021	0.092	0.887
	Reading SOL scores, standardized by grade, 2013	24878 ^d	0.015	0.997	-0.010	-0.025	0.077	0.690
	Reading SOL scores, standardized by grade, 2014	24409 ^e	0.008	0.997	-0.021	-0.027	0.068	0.969

Notes: *** - significance at less than 1%; ** - significance at 5%, * - significance at 10%. ^a Using years in district for Henrico. ^b The number of lessons taught and read were reported by teachers in the mid-year and end-of-year surveys. 127 teachers did not take part in either of the surveys, hence the missing values. ^c See (b) for an explanation of attrition. 20/127 teachers with missing values in (b) had non-zero values for the number of lessons downloaded. ^d A small share of students have no recorded 2013 test scores. This is likely due to transfers into the district. ^e 18 teachers did not have students with reading scores that year. Other comments: The test of equality of the group means is performed using a regression of each characteristic on treatment indicators and the district fixed effects interacted with the requested indicator. P-values for the joint significance of the treatment indicators are reported in Column (7). For student-level characteristics, standard errors are clustered at teacher level.

Table 2. Effects on Student Test Scores.

	Mathematics						Falsification: English	
	2014 Raw Score	2014 Raw Score	2014 Standardized Score	2014 Standardized Score	2014 Standardized Score	2014 Standardized Score	2014 Raw Score	2014 Standardized Score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
License Only	2.653	3.583*	0.050	0.061*	0.060*	0.055*	1.105	0.025
	[2.136]	[1.926]	[0.040]	[0.034]	[0.033]	[0.032]	[1.041]	[0.019]
Full Treatment	7.899***	7.057***	0.105**	0.094**	0.086**	0.093***	0.460	0.008
	[2.662]	[2.308]	[0.046]	[0.038]	[0.038]	[0.035]	[1.223]	[0.022]
District FE x Requested	Y	Y	Y	Y	Y	Y	Y	Y
District FE x Teacher-Level Lagged Test Scores	Y	Y	Y	Y	Y	Y	Y	Y
District FE x Individual Lagged Test Scores	N	N	N	N	Y	N	Y	Y
All controls	N	Y	N	Y	Y	Y	Y	Y
Observations	27,613	27,613	27,613	27,613	27,613	363	25,038	25,038
Unit of Observation	Student	Student	Student	Student	Student	Teacher	Student	Student

Notes: *** - significance at less than 1%; ** - significance at 5%; * - significance at 10%. Standard errors clustered at the teacher level are reported in square brackets. All specifications include controls for the requested indicator, average teacher-level 2013 math and reading test scores, and teacher-level shares of students with missing 2013 math and reading test scores - all interacted with district fixed effects. So that we can include all students with math scores in 2014 in regression models, students with missing 2013 standardized math and reading scores were given an imputed score of zero. To account for this in regression models, we also include indicators denoting these individuals in all specifications. Results are robust to restricting the sample to students with complete data. Columns (5), (7), and (8) control for individual-level 2013 math and reading test scores. Additional student-level controls include race, and gender. Additional teacher-level controls include teachers' educational attainment, years of experience, sex, race, grade fixed effects, as well as the percentage of male, black, white, Asian, and Hispanic students in the classroom. Standardized scores refer to the raw scores standardized by exam type. In the absence of exam type data for Hanover, test scores for that district were standardized by grade.

Table 3. Effects on Lesson Use.

Panel A: Subsample of Teachers Who Answered Both Mid-Year and End-of-Year Surveys (~20%).				
	Lessons Looked	Lessons Taught	Lessons Downloaded	Webinars Viewed
	(1)	(2)	(3)	(4)
License Only	1.404 [5.018]	0.092 [1.650]	1.969 [4.178]	0.168 [0.175]
Full Treatment	5.103 [5.021]	2.284 [1.912]	3.699 [4.225]	0.499** [0.231]
All controls	Y	Y	Y	Y
Observations	69	69	69	69
Panel B: Subsample of Teachers Who Answered either Mid-Year or End-of-Year Survey (~60%).				
	Lessons Looked	Lessons Taught	Lessons Downloaded	Webinars Viewed
	(5)	(6)	(7)	(8)
License Only	1.396** [0.700]	0.466 [0.407]	1.034** [0.490]	-0.027 [0.018]
Full Treatment	2.618*** [0.720]	0.983** [0.390]	2.134*** [0.588]	0.097** [0.041]
All controls	Y	Y	Y	Y
Observations	236	236	236	236
Panel C: Multiple Imputation Estimates. Missing Outcome Data From Panel A Imputed Using Multiple Imputation.				
	Lessons Looked	Lessons Taught	Lessons Downloaded	Webinars Viewed
	(9)	(10)	(11)	(12)
License Only	1.586*** [0.418]	0.657*** [0.191]		
Full Treatment	4.404*** [0.605]	1.925*** [0.282]	N/A	N/A
All controls	Y	Y		
Observations	363	363		
Panel D: Full Sample Estimates. Missing Data for Lessons Looked and Taught Replaced with Zero (Lower Bound).				
	Lessons Looked	Lessons Taught	Lessons Downloaded	Webinars Viewed
	(13)	(14)	(15)	(16)
License Only	1.115*** [0.422]	0.262 [0.221]	0.916*** [0.328]	-0.013 [0.009]
Full Treatment	2.236*** [0.506]	0.573** [0.238]	1.900*** [0.457]	0.048** [0.022]
All controls	Y	Y	Y	Y
Observations	363	363	363	363

Notes: *** - significance at less than 1%; ** - significance at 5%; * - significance at 10%. Robust standard errors are reported in square brackets. Standard errors in Panels C are corrected for multiple imputation according to Rubin (2004). All specifications include controls for the requested indicator, average teacher-level 2013 math and reading test scores, and teacher-level shares of students with missing 2013 math and reading test scores - all interacted with district fixed effects. Additional controls include teachers' education level, years of experience, sex, race, grade fixed effects, as well as the percentage of male, black, white, Asian, and Hispanic students in their class. The data on lessons downloaded and webinars watched are available for all 363 teachers. The number of lessons taught or read was missing for some teachers because of survey non-response: 69 teachers completed both mid-year and end-of-year surveys, 236 teachers completed either of the two. Panel A restricts the sample to 69 teachers who completed both surveys. Panel B restricts the sample to 236 teachers who completed either survey. Panel C uses data from 69 teachers to impute the missing values using multiple imputation (Rubin, 2004). Multiple imputation is performed using a Poisson regression (outcomes are count variables) and 20 imputations. Imputed values in each imputation sample is based on the predicted values from a Poisson regression of lesson use on treatment and requested status. Panel D studies all 363 teachers, replacing missing data for lessons looked and taught with zeros.

Table 4. Students' Post-Treatment Survey Analysis (Chesterfield and Hanover only).

	Share of Completed Surveys	Standardized Factors					
		Math has Real Life Application	Increased Interest in Math Class	Increased Effort in Math Class	Increased Motivation for Studying in General	Math Teacher Promotes Deeper Understanding	Math Teacher Gives Individual Attention
Panel A. No Controls.							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
License Only	-0.052 [0.083]	-0.017 [0.072]	-0.030 [0.075]	0.010 [0.046]	-0.021 [0.053]	0.052 [0.076]	0.085 [0.078]
Full Treatment	-0.036 [0.095]	0.158** [0.076]	0.058 [0.074]	0.030 [0.045]	0.036 [0.050]	0.204** [0.081]	0.187*** [0.072]
End-of-Year Indicator	Y	Y	Y	Y	Y	Y	Y
District FE x Requested	N	N	N	N	N	N	N
All controls	N	N	N	N	N	N	N
Observations	27,450	18,013	17,855	18,010	17,822	17,899	18,503
Panel B. With All Controls.							
	(8)	(9)	(10)	(11)	(12)	(13)	(14)
License Only	0.100 [0.082]	-0.012 [0.060]	-0.018 [0.062]	0.045 [0.035]	-0.021 [0.035]	0.001 [0.065]	0.033 [0.063]
Full Treatment	0.012 [0.099]	0.162** [0.063]	0.087 [0.074]	0.003 [0.044]	0.039 [0.035]	0.175** [0.070]	0.144** [0.069]
End-of-Year Indicator	Y	Y	Y	Y	Y	Y	Y
District FE x Requested	Y	Y	Y	Y	Y	Y	Y
All controls	Y	Y	Y	Y	Y	Y	Y
Observations	27,450	17,959	17,799	17,954	17,768	17,843	18,443

Notes: *** - significance at less than 1%; ** - significance at 5%, * - significance at 10%. Standard errors clustered at the teacher level are reported in square brackets. For details on the estimating strategy, see (3). Each outcome, except for the share of completed surveys, is a result of factor analysis and encompasses variation from several individual questions. For details on how the factors were formed, see Appendix C. The specifications in Panel A do not contain any covariates other than the treatment and end-of-year indicators. The specifications in Panel B add controls for district fixed effects, average teacher-level 2013 math and reading test scores, and teacher-level shares of students with missing 2013 math and reading test scores (all interacted with the requested indicator), as well as teachers' education level, years of experience, sex, race, grade fixed effects, and the percentage of male, black, white, Asian, and Hispanic students in their class. The fact that the survey was anonymous prevented us from including any student-level covariates. The regressions presented in Column (1) are estimated at the teacher level. The share of completed surveys for each teacher was calculated by comparing the number of completed student surveys with the number of students with complete data on math test scores.

For Online Publication.

Appendix A. Treatment Allocation.

Table A1. Total Number of Teachers Participating, by District and Treatment Condition.

	Treatment By District			Total	Requested
	Control	License Only	Full Treatment		
Hanover	19	18	19	56	0
Henrico	46	46	43	135	89
Chesterfield	75	40	57	172	33
Total	140	104	119	363	122

Appendix B. Main Result without Requested Teachers.

Table B1. Main Result without Requested Teachers

	Mathematics						Falsification: English	
	2014 Raw Score	2014 Raw Score	2014 Standardized Score	2014 Standardized Score	2014 Standardized Score	2014 Standardized Score	2014 Raw Score	2014 Standardized Score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
License Only	2.125	2.684	0.039	0.042	0.043	0.048	-0.688	-0.012
	[2.111]	[2.023]	[0.038]	[0.036]	[0.036]	[0.035]	[1.050]	[0.019]
Full Treatment	9.382***	8.714***	0.124***	0.108**	0.101**	0.117***	1.880	0.030
	[2.904]	[2.692]	[0.046]	[0.045]	[0.044]	[0.043]	[1.450]	[0.026]
District FE x Requested	Y	Y	Y	Y	Y	Y	Y	Y
District FE x Teacher-Level Lagged Test Scores	Y	Y	Y	Y	Y	Y	Y	Y
District FE x Individual Lagged Test Scores	N	N	N	N	Y	N	Y	Y
All controls	N	Y	N	Y	Y	Y	Y	Y
Observations	16,883	16,883	16,883	16,883	16,883	241	14,427	14,427
Unit of Observation	Student	Student	Student	Student	Student	Teacher	Student	Student

Notes: *** - significance at less than 1%; ** - significance at 5%; * - significance at 10%. Standard errors clustered at the teacher level are reported in square brackets. Only teachers who did not request a license are included in the analysis. All specifications include controls for the requested indicator, average teacher-level 2013 math and reading test scores, and teacher-level shares of students with missing 2013 math and reading test scores - all interacted with district fixed effects. So that we can include all students with math scores in 2014 in regression models, students with missing 2013 standardized math and reading scores were given an imputed score of zero. To account for this in regression models, we also include indicators denoting these individuals in all specifications. Results are robust to restricting the sample to students with complete data. Columns (5), (7), and (8) control for individual-level 2013 math and reading test scores. Additional student-level controls include race, and gender. Additional teacher-level controls include teachers' educational attainment, years of experience, sex, race, grade fixed effects, as well as the percentage of male, black, white, Asian, and Hispanic students in the classroom. Standardized scores refer to the raw scores standardized by exam type. In the absence of exam type data for Hanover, test scores for that district were standardized by grade.

Appendix C. Construction of Factors for The Student Survey.

Factor 1: Math has Real Life Application	Factor 2: Increased Interest in Math Class	Factor 3: Increased Effort in Math Class	Factor 4: Increased Motivation for Studying in General	Factor 5: Math Teacher Promotes Deeper Understanding	Factor 6: Math Teacher Gives Individual Attention
My math teacher often connects what I am learning to life outside the classroom (0.570)	I usually look forward to this class (0.644)	I work hard to do my best in this class (0.212)	I set aside time to do my homework and study (0.320)	My math teacher encourages students to share their ideas about things we study in class (0.621)	My math teacher is willing to give extra help on schoolwork if I need it (0.605)
In math how often do you apply math situations in life outside of school (0.584)	Sometimes I get so interested in my work I don't want to stop (0.610)	Lower bound hours per week studying/working on math outside class (0.212)	I try to do well on my schoolwork even when it isn't interesting to me (0.373)	My math teacher encourages us to consider different solutions or points of view (0.652)	My math teacher notices if I have trouble learning something (0.605)
In math how often do your assignments seem connected to the real world (0.628)	The topics are interesting/challenging (0.562)		I finish whatever I begin. Like you? (0.617)	My math teacher wants us to become better thinkers, not just memorize things (0.574)	
Do you think math can help you understand questions or problems that pop up in your life? (0.507)	Times per week you talk with your parents or friends about what you learn in math class (0.373)		I am a hard worker. Like you? (0.691)	In math how often do you talk about different solutions or points of view (0.501)	
	Number of students in math class who feel it is important to pay attention in class (0.305)		I don't give up easily. Like you? (0.623)	My math teacher explains things in a different way if I don't understand something in class (0.595)	

Notes: Each factor is represented in a different column. The individual questions used to create each factor are presented. The rotated factor loadings are presented in parentheses under each question.

Appendix D. Effect Heterogeneity by Teacher Quality.

As a start, we use the teacher value-added model as presented in [Jackson et al. \(2014\)](#).²⁶ We show that marginal effects in this standard value-added model, when aggregated up to the teacher level, yield a very intuitive interpretation in conditional quantile regression models. Specifically, we will show that when average student test scores (at the teacher level) are used as an outcome, the estimated coefficient of a randomized treatment using conditional quantile regression at quantile τ , is the estimated effect of that treatment on teachers at the τ th percentile of the teacher quality distribution.

The standard teacher effects model states that student test scores are determined as below:

$$Y_{it} = X_{it}\beta + \mu_t + \theta_c + \varepsilon_{it}$$

Here Y_{it} is student i 's test score, where student i is being taught by teacher t . X_{it} are observable student covariates, ε_{it} is the idiosyncratic student-level effect, θ_c is the classroom fixed effect, and, finally, μ_t is the teacher t 's value added. That is, a teacher's value added is the average increase (relative to baseline) in student test scores caused by the teacher. Let us aggregate this model to the teacher level by taking averages.

$$\bar{Y}_t = \frac{1}{S} \sum_{i=1}^S Y_{it} = \bar{X}_t\beta + \mu_t + \theta_c + \bar{\varepsilon}_t$$

Our hypothesis is that teacher effects are impacted by the treatment. That is, we posit that:

$$\mu_t = \beta_T T_t + v_t$$

, where β_T is the influence of Mathalicious lessons on the teacher's value added, while v_t is the teacher fixed effect before introducing the treatment. The full model is now:

$$\bar{Y}_t = \beta_T T_t + \bar{X}_t\beta_{-T} + v_t + \theta_c + \bar{\varepsilon}_t \quad (4)$$

Now note that treatment was randomized across teachers. In terms of our model this means that T_t is independent of all other random variables in the model, i.e. $T_t \perp \{\bar{X}_t, v_t, \theta_c, \bar{\varepsilon}_t\}$. Now, assuming that β_T and β_{-T} may vary with the quantile τ , let us apply the quantile function to the equation above:

$$Q_\tau(\bar{y}_t|T, \bar{X}) = \beta_T(\tau)T_j + \bar{X}_t\beta_{-T}(\tau) + Q_\tau(v_t(\tau) + \theta_c(\tau) + \bar{\varepsilon}_t(\tau)|\bar{X})$$

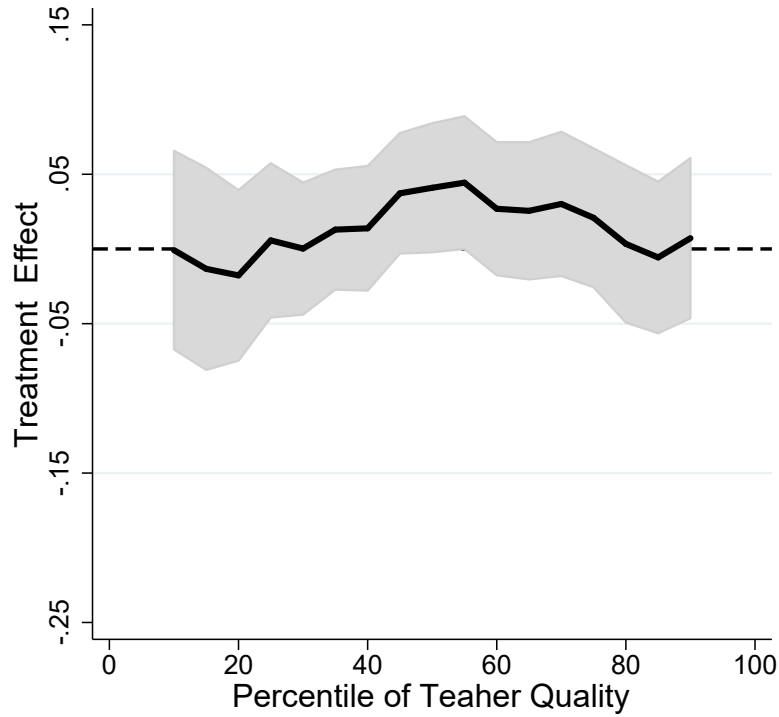
Now, assuming $Q_\tau(v_t(\tau) + \theta_c(\tau) + \bar{\varepsilon}_t(\tau)|T, \bar{X}) = 0$ for each quantile τ ,²⁷ the quantile regression coefficient $\hat{\beta}_T(\tau)$ is a consistent estimate of $\beta_T(\tau)$ in the model 4. Moreover, it is asymptotically normal. This can be proven by putting the moments into a GMM framework, e.g. see [Buchinsky \(1998\)](#). To conclude, conditional quantile regression model provides marginal effect estimates at particular quantiles of the distribution of the residual, which in our case can be interpreted as teacher value-added.

²⁶We suppress the time subscript, as there is no time dimension in our application.

²⁷This is a standard assumption in the quantile regression literature. For a reference, see e.g. [Buchinsky \(1998\)](#)

Appendix E. Quantile Regression: English Test Scores.

Figure E1. Marginal Effect of the Full Treatment by Classroom Quality.
Falsification Test: English Test Scores.



Notes: The solid black line represents treatment effect estimates that result from model (1) being evaluated at different quantiles of teacher quality using conditional quantile regression. Teacher-level average standardized 2014 English test scores serve as the main outcome. The shaded area depicts the 90% confidence interval for each regression estimate. For a formal discussion of the method, see [Appendix D](#).

Appendix F. Test Score Regressions - Teacher Level.

Table F1. Effect on Student Math Scores, Aggregated to the Teacher Level.

	Mathematics				Falsification: English	
	2014 Raw Score	2014 Raw Score	2014 Standardized Score	2014 Standardized Score	2014 Raw Score	2014 Standardized Score
	(1)	(2)	(3)	(4)	(5)	(6)
License Only	1.669 [2.087]	4.291** [2.072]	0.017 [0.034]	0.055* [0.032]	2.096 [5.874]	0.015 [0.022]
Full Treatment	8.401*** [2.431]	7.905*** [2.234]	0.093** [0.039]	0.093*** [0.035]	1.637 [3.826]	0.003 [0.024]
District FE x Requested	Y	Y	Y	Y	Y	Y
District FE x Lagged Test Scores	Y	Y	Y	Y	Y	Y
All controls	N	Y	N	Y	Y	Y
Observations	363	363	363	363	363	363
Unit of Observation	Teacher	Teacher	Teacher	Teacher	Teacher	Teacher

Notes: *** - significance at less than 1%; ** - significance at 5%; * - significance at 10%. Robust standard errors are reported in square brackets. All specifications include controls for the requested indicator, average teacher-level 2013 math and reading test scores, and teacher-level shares of students with missing 2013 math and reading test scores - all interacted with district fixed effects. Other controls include teachers' education level, years of experience, sex, race, grade fixed effects, as well as the percentage of male, black, white, Asian, and Hispanic students in their class. Standardized scores refer to the raw scores standardized by exam type. In the absence of exam type data for Hanover, test scores for that district were standardized by grade.

Appendix G. Stylized Model of Teacher Multitasking.

G1. General set-up

Let us consider the general optimization problem for a teacher and then lay out the parametric assumptions we impose. In our model, a teacher cares about her students' test scores (y_i , where i is a student from a class of size s). In turn, student i 's test score depends on the teacher's allocation of time (T) between planning lessons (d) and other tasks (n). Teacher's (in)ability to plan lessons is modeled as a 'price' p_d that amplifies the time needed to achieve d units of lesson quality. Similarly, 'price' p_n denotes the teacher's ability to achieve n units of other teaching tasks. Note that the higher teacher's abilities are, the lower are her corresponding p 's. Formally, we write:

$$\begin{aligned} U \left(\left\{ y_i(n, d) \right\}_{i=1}^s \right) &\rightarrow \max_{\{n, d\}} & (5) \\ \text{s.t. } p_n n + p_d d &\leq T \\ n &\geq 0 ; d \geq 0 \end{aligned}$$

We model off-the-shelf lessons as a technology that guarantees a minimum quality of lesson planning \underline{d} at a fixed time cost F . Teachers can either stick to their own efforts or delegate part of lesson planning to off-the-shelf lessons. If a teacher chooses to pay a fixed cost F and adopt off-the-shelf lessons, she is now able to spend the time saved from adopting lessons ($p_d \underline{d}$) on improving the lessons further or on other tasks. Thus, the optimization problem of a teacher with off-the-shelf lessons could be formally written as follows:

$$\begin{aligned} U \left(\left\{ y_i(n, d) \right\}_{i=1}^s \right) &\rightarrow \max_{\{n, d\}} & (6) \\ \text{s.t. } p_n n + p_d d &\leq T + p_d \underline{d} - F \\ n &\geq 0 ; d \geq \underline{d} \end{aligned}$$

G2. Special Case with Functional Form Assumptions

For the ease of exposition, we will consider a special case of the model with two functional form assumptions. Let U be a weighted average of students' test scores:

$$U \left(\left\{ y_i(n, d) \right\}_{i=1}^s \right) = \frac{1}{s} \sum_{i=1}^s w_i y_i(n, d)$$

Furthermore, let y_i be a Cobb-Douglas-type function with a common elasticity $\alpha \in [0, 1]$, but with a student-level heterogeneity parameter A_i :²⁸

$$y_i(n, d) = A_i n^\alpha d^{1-\alpha}$$

Finally, we assume that the stock of time is large enough, so that:²⁹

$$T > \frac{\alpha}{1-\alpha} p_d \underline{d} + F$$

This assumption allows us to rule out the possibility that some teachers will choose to locate at the kink of the budget line with off-the-shelf lessons. Instead, we can focus on teachers with a budget line that continues as a straight line after reaching the kink (drawn as a dotted line on Panels C and D in Figure 1).³⁰

G2.1. Solving the model without off-the-shelf lessons

After solving the problem in (5) using standard arguments for a Cobb-Douglas utility functions, we get the following optimal allocation:

$$n^* = \frac{\alpha T}{p_n} ; d^* = \frac{(1-\alpha)T}{p_d}$$

The optimal level of test scores is:

$$U(n^*, d^*) = \sum_{i=1}^s \frac{w_i A_i T}{s} \left[\frac{\alpha^\alpha (1-\alpha)^{1-\alpha}}{p_n^\alpha p_d^{1-\alpha}} \right] \quad (7)$$

²⁸ $A_i > 0$ can be interpreted as student i 's ability or an individual shock parameter. Technically, A_i is indistinguishable from the weight w_i with which a teacher values student i 's test score.

²⁹Under $\alpha = 1/2$ this assumption is equivalent to a statement that the stock of time each teacher possesses is large enough to cover both the fixed cost of lesson adoption (F) and the time needed to achieve \underline{d} on her own ($p_d \underline{d}$). Given that a teacher has to spend much time on tasks unrelated to designing lessons (n), this assumption is reasonably weak. It also captures an intuition that the fixed cost of adopting the lessons are not likely to be prohibitively large. Finally, this assumption leads to a realistic outcome that it will always be optimal to spend some of one's own time planning lessons, even if the online lessons are high quality.

³⁰Many of these assumptions are made solely for the illustrative purposes and could be relaxed. For instance, we could obtain similar predictions under the general utility and test score functions with reasonable assumptions on derivatives, as the main driving forces behind our results would remain unchanged. One could also weaken the assumption that the cost of adopting a lessons F does not depend on teacher ability - to the extent that the wedge between the cost differential is not as sharp as a difference between $p_d d^*$ and $p'_d d^*$ where $p'_d > p_d$, our predictions will still go through. (This latter conjecture is intuitive because low ability teachers will likely be much closer to high ability teachers in adopting lessons than in creating lessons of similar quality from scratch.) Similarly, one could weaken the assumption that teachers do not locate at the kink of the budget line, as this assumption does not change the fact that (i) the direct increase in lesson quality will be more important for the low ability teachers, (ii) the time savings from off-the-shelf lessons will be higher for the low ability teachers, and (iii) the law of diminishing returns will reinforce the differences in benefits across teacher skill.

Note that average test scores decrease both in p_n and p_d , i.e. increase in teacher ability. This is in line with our intuition that, *all else equal*, higher ability teachers would have students with higher test scores on average.

G2.2. Solving the model with off-the-shelf lessons

Applying similar calculations in Section G2.1 to the problem in (6), and ignoring the constraint of $d \geq \underline{d}$, we obtain the following optimal allocation:³¹

$$\tilde{n} = \frac{\alpha(T + p_d \underline{d} - F)}{p_n} ; \tilde{d} = \frac{(1 - \alpha)(T + p_d \underline{d} - F)}{p_d}$$

The optimal level of test scores is:

$$U(\tilde{n}, \tilde{d}) = \sum_{i=1}^s \frac{w_i A_i (T + p_d \underline{d} - F)}{s} \left[\frac{\alpha^\alpha (1 - \alpha)^{1-\alpha}}{p_n^\alpha p_d^{1-\alpha}} \right] \quad (8)$$

G2.3. Adoption of off-the-shelf lessons

A teacher adopts off-the-shelf lessons whenever student test scores under such technology are greater or equal to the test scores without it. Since teachers are heterogeneous in parameters p_n and p_d , let us find a set of threshold values $\hat{p} = \{\hat{p}_n, \hat{p}_d\}$ such that teachers with \hat{p} are indifferent between using off-the-shelf lessons and sticking to their own effort. Threshold values \hat{p} are defined by the following equation:

$$\sum_{i=1}^s \frac{w_i A_i T}{s} \left[\frac{\alpha^\alpha (1 - \alpha)^{1-\alpha}}{\hat{p}_n^\alpha \hat{p}_d^{1-\alpha}} \right] = \sum_{i=1}^s \frac{w_i A_i (T + \hat{p}_d \underline{d} - F)}{s} \left[\frac{\alpha^\alpha (1 - \alpha)^{1-\alpha}}{\hat{p}_n^\alpha \hat{p}_d^{1-\alpha}} \right] \quad (9)$$

, with the optimal utility level without off-the-shelf lessons (7) on the left-hand side and the optimal utility with off-the-shelf lessons (8) on the right-hand side. After canceling repeating parameters, we get a threshold value of $\hat{p}_d = F/\underline{d}$. Specifically, teachers *choose to adopt* off-the-shelf lessons whenever $p_d > F/\underline{d}$ and *choose not to* if $p_d \leq F/\underline{d}$. These calculations are intuitive as teachers choose to adopt the lessons whenever time savings from off-the-shelf lessons ($p_d \underline{d}$) are greater or equal to the associated time costs (F). Importantly, under our assumptions, this adoption rule mechanically leads to $\tilde{d} \geq \underline{d}$. To conclude, in our model, the adoption decision fully depends on the teacher (in)ability to develop lesson plans.

³¹As will be shown in Section G2.3., ignoring the constraint of $d \geq \underline{d}$ is valid in this model. The reason is that the teachers who choose to adopt the lessons are those for whom, under our assumptions, the optimal level of lesson planning \tilde{d} without this constraint is larger than \underline{d} .

G2.4. Evaluating the predictions

Prediction 1: *If teachers know their ability, among those who chose to adopt lessons voluntarily, the gains in average test scores from using the off-the-shelf lessons are non-negative.*

This prediction holds by construction, see Section G2.3.

Prediction 2: *Among those who chose to adopt lessons, teacher time spent on n (that is, all tasks complementary to lesson planning) should increase.*

Indeed, for those who chose to adopt off-the-shelf lessons, the teacher time spent on other teaching tasks strictly increases. This follows directly from the adoption rule:

$$T - F + p_d \underline{d} > T \implies \tilde{n} = \frac{\alpha(T - F + p_d \underline{d})}{p_n} > \frac{\alpha T}{p_n} = n^*$$

Prediction 3: *Among teachers who chose to adopt lessons, the effect on lesson quality d is positive.*

Using the same proof procedure as for Prediction 2, one can show that $\tilde{d} > d^*$ for the teachers who chose to adopt off-the-shelf lessons.

Prediction 4: *The gains to using off-the-shelf lessons are decreasing in teacher effectiveness (as measured by ability to raise average test scores).*

First, one can show that the gains from adopting the lessons $U(\tilde{n}, \tilde{d}) - U(n^*, d^*)$ are strictly increasing in p_d :

$$\frac{\partial [U(\tilde{n}, \tilde{d}) - U(n^*, d^*)]}{\partial p_d} = \sum_{i=1}^s \frac{w_i A_i}{s} \left[\frac{\alpha^\alpha (1 - \alpha)^{1-\alpha}}{\hat{p}_n^\alpha \hat{p}_d^{1-\alpha}} \right] \left[\alpha \underline{d} + \frac{(1 - \alpha)F}{p_d} \right] > 0$$

Moreover, one can prove that, when both p_d and p_n are increased *simultaneously by the same percentage*, the difference $U(\tilde{n}, \tilde{d}) - U(n^*, d^*)$ strictly increases. Specifically, after taking the exact differential of $U(\tilde{n}, \tilde{d}) - U(n^*, d^*)$, we show that simultaneous increases of p_d and p_n by the same percentage (i.e. such that $dp_d/p_d = dp_n/p_n = \varepsilon$) lead to an increase of the total difference:

$$\begin{aligned} d[\tilde{U} - U^*] &= \sum_{i=1}^s \frac{w_i A_i}{s} \left[\frac{\alpha^\alpha (1 - \alpha)^{1-\alpha}}{\hat{p}_n^\alpha \hat{p}_d^{1-\alpha}} \right] \left[[\alpha p_d \underline{d} + (1 - \alpha)F] \frac{dp_d}{p_d} - [\alpha(p_d \underline{d} - F)] \frac{dp_n}{p_n} \right] = \\ &= \sum_{i=1}^s \frac{w_i A_i}{s} \left[\frac{\alpha^\alpha (1 - \alpha)^{1-\alpha}}{\hat{p}_n^\alpha \hat{p}_d^{1-\alpha}} \right] F \varepsilon > 0 \end{aligned}$$

To conclude, our model predicts bigger gains from off-the-shelf lessons for less effective teachers.

Appendix H. Survey Response and Lesson Downloads.

Table H1. Survey Response and Lessons Downloads.

	1 = Participated in Both Surveys	1 = Participated in Both Surveys	1 = Participated in Either Survey	1 = Participated in Either Survey
	(1)	(2)	(3)	(4)
Lessons Downloaded	0.008 [0.009]	0.011 [0.008]	0.003 [0.009]	0.004 [0.009]
Treatment Status	Y	Y	Y	Y
District FE x Requested	Y	Y	Y	Y
All controls	N	Y	N	Y
Observations	363	363	363	363

Notes: *** - significance at less than 1%; ** - significance at 5%, * - significance at 10%. Robust standard errors are reported in square brackets. The outcomes are indicators for participation in both (either) mid-year and (or) end-of-year teacher surveys. All specifications include controls for the treatment indicators and the requested indicator interacted with district fixed effects. Other controls include average teacher-level 2013 math and reading test scores interacted with district fixed effects, teacher-level shares of students with missing 2013 math and reading test scores interacted with district fixed effects, teachers' education level, years of experience, sex, race, grade fixed effects, as well as the percentage of male, black, white, Asian, and Hispanic students in their class.

Appendix I. Teacher Survey.

This appendix explores the effects of providing teachers with licenses for off-the-shelf lessons, with or without complementary supports, on teacher behavior as reported by teachers themselves in an end-of-year survey.

As with the student surveys, we created factors based on several questions. The first four factors measure teachers' classroom practices: the first is based on a single question is how much homework teachers assign; the second one measures how much time teachers spend practicing for standardized exams; the third factor measures inquiry-based teaching practices, and the fourth factor measures how much teacher engage in individual or group work. We also asked questions regarding teacher attitudes to create three factors. The first factor we construct represents teacher's loyalty to the school. The second factor is measuring the level of support coming from schools. The third factor measures whether teachers enjoy teaching students. Similar to the classroom practices, we find no systematic changes on these measures. Finally, we also construct a measure of teachers' perceptions of student attitudes. The first such factor measures whether teachers consider their students disciplined, and the other factor measures teachers' perception of the classroom climate among students.

[Table J1](#) summarizes our regression results. Unfortunately, there are large difference in survey response rates across the treatment arms for teachers. The fully treated teachers were 12 percentage points more likely to response to the surveys than control teachers. As such, one should interpret the teacher survey results with caution. Having presented the limitation of the teacher surveys, the data provide little evidence that either the full treatment or the license only treatment has any effect on teacher satisfaction, teacher classroom practices, or their perception of the classroom dynamics among students. The only practice for which the effect is on the borderline of being statistically significant is treatment teachers assigning more homework. Taken at face value, these patterns suggest that teacher in the full treatment condition simply substituted the off-the-shelf lessons for their own lessons and may have assigned more homework as a results. However, treated teachers did not appear to make many any other changes to their classroom practices or teaching style. This implies that the positive observed effects simply reflect off-the-shelf substituting for low teacher skills rather than any learning of change in teacher teaching style.

Table I1. Teacher Post-Treatment Survey Analysis.

	Missing survey	Teaching practices			Teacher attitude			Student attitude		
		Homeworks assigned (hours)	Time spent practicing standardized exams (%)	Teaching practices (factor)	Student-teacher interactions (factor)	Would like to stay in this school (factor)	Supportive school (factor)	Enjoy teaching (factor)	Students are disciplined (factor)	Student group dynamics (factor)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
License Only	0.071 [0.065]	-0.033 [0.093]	0.007 [0.241]	0.077 [0.188]	-0.024 [0.201]	-0.142 [0.164]	0.010 [0.203]	0.065 [0.222]	0.056 [0.190]	0.064 [0.177]
Full Treatment	0.079 [0.076]	0.117 [0.093]	-0.042 [0.266]	0.003 [0.195]	-0.100 [0.207]	-0.090 [0.176]	-0.019 [0.217]	-0.193 [0.201]	0.173 [0.207]	0.004 [0.209]
District FE x Requested	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
All controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	363	209	209	205	203	207	206	204	205	205

Notes: *** - significance at less than 1%; ** - significance at 5%; * - significance at 10%. Robust standard errors are reported in square brackets. Factors are obtained through factor analysis of related survey questions. For details, see exact factor loadings in [Table I2](#). All specifications include controls for the requested indicator, average teacher-level 2013 math and reading test scores, and teacher-level shares of students with missing 2013 math and reading test scores - all interacted with district fixed effects. Other controls include teachers' education level, years of experience, sex, race, grade fixed effects, as well as the percentage of male, black, white, Asian, and Hispanic students in their class.

Table I2. Teacher Post-Treatment Survey. Factor Loadings.

Factor 1: Teaching practices	Factor 2: Student-teacher interactions	Factor 3: Would like to stay in this school	Factor 4: Supportive school	Factor 5: Enjoy teaching	Factor 6: Students are disciplined	Factor 7: Student group dynamics
<i>How often do you ask your students to:</i>	<i>How often do students do the following?</i>				<i>How many of your students do the following?</i>	
... explain the reasoning behind an idea? (0.464)	Work individually without assistance from the teacher (0.585)	I usually look forward to each working day at this school (0.754)	My school encourages me to come up with new and better ways of doing things. (0.705)	Teaching offers me an opportunity to continually grow as a professional. (0.329)	Come to class on time. (0.20)	Students build on each other's ideas during discussion. (0.734)
... analyze relationships using tables, charts, or graphs? (0.608)	Work individually with assistance from the teacher (0.713)	I feel loyal to this school. (0.705)	I am satisfied with the recognition I receive for doing my job. (0.679)	I find teaching to be intellectually stimulating. (0.47)	Attend class regularly. (0.226)	Students show each other respect. (0.51)
... work on problems for which there are no obvious methods of solution? (0.626)	Work together as a class with the teacher teaching the whole class (0.635)	I would recommend this school to parents seeking a place for their child (0.675)	The people I work with at my school cooperate to get the job done. (0.496)	I enjoy sharing things I'm interested in with my students (0.692)	Come to class prepared with the appropriate supplies and books. (0.516)	Most students participate in the discussion at some point. (0.60)
... use computers to complete exercises or solve problems? (0.277)	Work together as a class with students responding to one another (0.355)	I would recommend this school district as a great place to work for my friends (0.414)	I have access to the resources (materials, equipment, etc.) I need (0.424)	I enjoy teaching others. (0.731)	Regularly pay attention in class. (0.733)	Students generate topics for class discussions. (0.636)
... write equations to represent relationships? (0.395)	Work in pairs or small groups without assistance from each other (0.221)	If I were offered a comparable teaching position at another district, I would stay. (0.502)		I find teaching interesting. (0.713)	Actively participate in class activities. (0.747)	
... practice procedural fluency? (0.206)	Work in pairs or small groups with assistance from each other (0.182)			Teaching is challenging. (0.194)	Always turn in their homework. (0.685)	
				Teaching is dull. (-0.435)		
				I have fun teaching (0.673)		
				Teaching is inspiring. (0.59)		

Notes: Each factor is represented in a different column. The individual questions used to create each factor are presented. The rotated factor loadings are presented in parentheses under each question.

Appendix J. Instrumental Variables Estimation.

As an additional test of whether lesson use is indeed responsible for an increase in math scores, we estimate instrumental variables regressions of test scores against lesson use using indicators for the six treatments as instruments. Note that we impute lesson use for those with missing or incomplete use data. The results are presented in [Table J1](#). Looking at the student level regression (Column 2), the instrumental variable coefficient on lessons taught is 0.033σ and is statistically significant at the 5 percent level. The effects are similar at the teacher level (Column 4). Note that in both these models the first stage F-statistic is above 10. In our placebo tests, the effects for English scores are very close to zero and are not statistically significant (Columns 8). To directly test for the possibility that the additional supports may have a positive effect irrespective of lesson use, we estimate the same instrumental variables regression while controlling for receiving the full treatment. In such models (Column 3 and 6), conditional on lesson use, the coefficient on the full treatment dummy is negative and not statistically significant, while the coefficient on lesson use is slightly larger (albeit no longer statistically significant due to larger standard errors). This is very similar to the results based on comparisons across the different treatments. Overall the patterns presented are inconsistent with the benefits being due to the extra supports, and provide compelling evidence that all of our effects are driven by the increased lesson use itself.

Table J1. Instrumental Variables (IV) Estimation with Lessons Taught as an Endogenous Variable.

	Mathematics						Falsification: English	
	2014 Standardized Score	2014 Standardized Score	2014 Standardized Score	2014 Standardized Score	2014 Standardized Score	2014 Standardized Score	2014 Standardized Score	2014 Standardized Score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lessons Taught	0.038** [0.018]	0.033** [0.015]	0.039 [0.033]	0.044** [0.018]	0.039** [0.016]	0.032 [0.031]	0.002 [0.010]	0.004 [0.008]
Full Treatment			-0.014 [0.076]			0.018 [0.071]		
District FE x Requested	Y	Y	Y	Y	Y	Y	Y	Y
District FE x Teacher-Level Lagged Test Scores	Y	Y	Y	Y	Y	Y	Y	Y
District FE x Individual Lagged Test Scores	Y	Y	Y	N	N	N	Y	Y
All controls	Y	Y	Y	Y	Y	Y	Y	Y
Observations	27,613	27,613	27,613	363	363	363	25,038	25,038
First Stage F-stat	23.84	41.87	4.607	15.51	16.69	3.252	20.94	46.52
Unit of Observation	Student	Student	Student	Teacher	Teacher	Teacher	Student	Student
Instruments	Treatment	Treatment X District	Treatment X District	Treatment	Treatment X District	Treatment X District	Treatment	Treatment X District

Notes: *** - significance at less than 1%; ** - significance at 5%, * - significance at 10%. Standard errors clustered at the teacher level are reported in square brackets. All specifications include controls for the requested indicator, average teacher-level 2013 math and reading test scores, and a teacher-level shares of students with missing 2013 math and reading test scores - all interacted with district fixed effects. Additional controls include teachers' education level, years of experience, sex, race, grade fixed effects, as well as the percentage of male, black, white, Asian, and Hispanic students in their class. In addition, the student-level specifications in Columns (1)-(3) and (7)-(8) control for individual-level math and reading test scores and all student level demographics. Standardized test scores refer to the raw test scores standardized by exam type. In the absence of exam type data for Hanover, test scores for that district were standardized by grade.

Appendix K. Sample Mathalicious Lesson #1.

This appendix includes the first 3 out of 7 pages extracted from the lesson guide for teachers.

licensed under CC-BY-NC

NEW-TRITIONAL INFO

How long does it take to burn off food from McDonald's?

lesson
guide



Many restaurants are required to post nutritional information for their foods, including the number of calories. But what does "550 calories" really mean? Instead of calories, what if McDonald's rewrote its menu in terms of exercise?

In this lesson, students will use unit rates and proportional reasoning to determine how long they'd have to exercise to burn off different McDonald's menu items. For instance, a 160-pound person would have to run for 50 minutes to burn off a Big Mac. So...want fries with that?!

Primary Objectives

- Calculate the number of calories burned per minute for different types of exercise and body weights
- Correctly write units (e.g. calories, cal/min, etc.) and simplify equations using them
- Calculate how long it would take to burn off menu items from McDonald's
- Discuss effects of posting calorie counts, and what might happen if exercise information were posted instead

Content Standards (CCSS)	Mathematical Practices (CCMP)	Materials
Grade 6 RP.3d, NS.3	MP.3, MP.6	<ul style="list-style-type: none">• Student handout• LCD projector• Computer speakers

Before Beginning...

Students should understand what a unit rate is; if they have experience calculating and using unit rates to solve problems, even better.

Preview & Guiding Questions

Students watch a McDonald's commercial in which NBA superstars LeBron James and Dwight Howard play one-on-one to determine who will win a Big Mac Extra Value Meal. When it's done, ask students, "How long do you think LeBron James would have to play basketball to burn off all the calories in a Big Mac?"

The goal isn't for students to come up with an exact answer. Instead, it's to get them thinking about the various factors that determine how many calories someone burns when he/she exercises. People burn calories at a faster rate when they do more strenuous exercise. Also, larger people burn more calories doing the same activity than smaller people. We don't expect students to know these things for sure, but they might conjecture that easier activities burn fewer calories, and that different people doing the same activity burn calories at a different rate.

- *How long do you think LeBron James would have to play basketball to burn off the calories in a Big Mac?*
- *What are some factors that might determine how long it would take someone to burn off calories?*
- *Do you think everyone burns the same number of calories when they exercise? Why or why not?*

Act One

After students have discussed some possible factors affecting how quickly someone burns calories, they will learn in Act One that there are three essential things to consider: their body, the type of exercise, and the duration of exercise. Students will first calculate how many calories people with different body types (including LeBron) will burn per minute while performing a variety of activities. Based on this, they'll be able to answer the question in the preview: LeBron would have to play basketball for about 86 minutes in order to burn off a Big Mac Extra Value Meal. Even if he played for an entire game, he wouldn't be able to burn off his lunch!

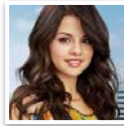
Act Two

Act Two broadens the scope even further by considering a wider assortment of exercises and different McDonald's items. Students will determine how long someone would have to do different activities to burn off each menu item. Then, they will listen to an NPR clip about the fact that McDonald's now posts calorie information for all of its items on the menu. Students will discuss whether or not this seems like an effective way to change people's behavior. We end with the following question: what might happen if McDonald's rewrote its menu in terms of *exercise*?

Act One: Burn It

- 1 When you exercise, the number of calories you burn depends on two things: the type of exercise and your weight. Playing basketball for one minute, for example, burns 0.063 calories for every pound of body weight.

Complete the table below to find out how many calories each celebrity will burn in **one minute of exercise**.



cal. burned in one min.	Selena Gomez 125 lb	Justin Timberlake 160 lb	Abby Wambach 178 lb	LeBron James 250 lb
Basketball 0.063 cal/lb	<i>7.88 calories per minute</i>	<i>10.08 calories per minute</i>	<i>11.21 calories per minute</i>	<i>15.75 calories per minute</i>
Soccer 0.076 cal/lb	<i>9.50 calories per minute</i>	<i>12.16 calories per minute</i>	<i>13.53 calories per minute</i>	<i>19.00 calories per minute</i>
Walking 0.019 cal/lb	<i>2.38 calories per minute</i>	<i>3.04 calories per minute</i>	<i>3.38 calories per minute</i>	<i>4.75 calories per minute</i>

Explanation & Guiding Questions

The math in this question is fairly straightforward. However, students might get confused by all the different units, and it may be worth demonstrating how they simplify. For instance, when LeBron James plays basketball, he burns 0.063 calories for every pound of body weight *each minute*. Since he weighs 250 pounds, he will burn

$$\left(\frac{0.063 \text{ cal}}{1 \text{ lb}} \times 250 \text{ lb} \right) \text{ per minute} = \frac{0.063 \text{ cal}}{1 \text{ lb}} \times \frac{250 \text{ lb}}{1} \text{ per minute} = 15.75 \text{ calories in one minute.}$$

Of course, not all students will be this intentional with their units, and it would be cumbersome to repeat this process for all twelve boxes. Still, it may be worth pointing out how the units simplify, lest “calories per minute” seem to come out of left field. However students calculate their unit rates, they should be able to explain what they mean in their own words, e.g. “Every minute that LeBron plays basketball, he burns 15.75 calories.”

- For a given exercise, who do you think will burn more calories in a minute – LeBron or Selena – and why?
- What does the unit rate, “0.063 calories per pound,” mean?
- What does the unit rate, “15.75 calories per minute,” mean?

Deeper Understanding

- Why do you think Selena Gomez burns so many fewer calories than LeBron does? (All your cells consume energy, i.e. burn calories, and LeBron, being so much heavier, has many more cells.)
- Why does playing soccer burn so many more calories per minute than walking does? (In soccer, a player runs, jumps, and kicks. These require more energy than walking. A calorie is a measure of energy.)
- How long would someone have to walk to burn the same number of calories as a minute of soccer? (Since walking burns 1/4 the calories of soccer, a person would have to walk 4 times as long, or 4 minutes.)

Appendix L. Sample Mathalicious Lesson #2.

This appendix includes the first 3 out of 8 pages extracted from the lesson guide for teachers.

licensed under CC-BY-NC

XBOX XPONENTIAL

How have video game console speeds changed over time?

lesson
guide



In 1965 Gordon Moore, computer scientist and Intel co-founder, predicted that computer processor speeds would double every two years. Twelve years later the first modern video game console, the Atari 2600, was released.

In this lesson, students write an exponential function based on the Atari 2600 and Moore's Law and research other consoles to determine whether they've followed Moore's Law.

Primary Objectives

- Apply an exponential growth model, stated verbally, to various inputs
- Generalize with an exponential function to model processor speed for a given year
- Research actual processor speeds, and compare them to the model's prediction
- Calculate the *annual* growth rate of the model (given biannual growth rate)
- Use technology to model the actual processor speeds with an exponential function
- Interpret the components of the regression function in this context, and compare them to the model

Content Standards (CCSS)		Mathematical Practices (CCMP)	Materials
Functions	IF.8b, BF.1a, LE.2, LE.5	MP.4, MP.7	<ul style="list-style-type: none">• Student handout• LCD projector• Computer speakers• Graphing calculators• Computers with Internet access
Statistics	ID.6a		

Before Beginning...

Students should be familiar with the meaning of and notation for exponents, square roots, percent growth and the basics of exponential functions of the general form $y = ab^x$. Students will need to enter data in calculator lists and perform an exponential regression, so if they're inexperienced with this process, you will need time to demonstrate.

Preview & Guiding Questions

We'll begin by watching a short video showing the evolution of football video games.



Ask students to sketch a rough graph of how football games have changed over time. Some will come up with a graph that increases linearly, perhaps some increasing at an accelerating rate. Some students may show great leaps in technology with new inventions, while others may show the quality leveling off in the more recent past.

Then, ask them to label the axes. The horizontal axis will be time in years, but what about the vertical axis? Ask students to describe what they are measuring, exactly, when they express the quality of a video game. They might suggest realism, speed or power. Students should try to explain how they would measure these (or others they come up with), and realize that while a subjective element like "realism" is difficult to quantify, it is possible to measure speed (in MHz) of a console's processor.

- Sketch a graph of how you think video games have changed over time.
- What was the reasoning behind the shape of the graph you sketched?
- What does your horizontal axis represent?
- What label did you assign to the vertical axis? Which of these are measurable?

Act One

In 1965 Gordon Moore, computer scientist and Intel co-founder, predicted that computer processor speeds would double every two years. Starting with the 1.2 MHz Atari 2600 in 1977 (the first console with an internal microprocessor), students apply the rule "doubles every two years" to predict the speed of consoles released in several different years. By extending the rule far into the future, they are motivated to write a function to model processor speed in terms of release year: $1.2 \cdot 2^{t/2}$. They will understand that 1.2 represents the speed of the initial processor, the base of 2 is due to doubling, and the exponent $t/2$ represents the number of doublings.







Act Two

How does the prediction compare to what has actually happened? Students research the actual processor speed of several consoles released over the years. By comparing predicted vs. actual processor speeds in a table, we see that they were slower than Moore's Law predicted. How different are the models, though? Students first algebraically manipulate the "doubling every two years" model to create one that expresses the growth rate each year. Then, they use the list and regression functionality of their graphing calculators to create an exponential function that models the actual data. By comparing the two functions, they conclude that while the actual annual growth rate (30%) was slower than the predicted annual growth rate based on Moore's Law (41%), the Atari 2600 was also ahead of its time.

Act One: Moore Fast

- 1 In 1965, computer scientist Gordon Moore predicted that computer processor speeds would double every two years. Twelve years later, Atari released the 2600 with a processor speed of 1.2 MHz.

Based on **Moore's Law**, how fast would you expect the processors to be in each of the consoles below?

					
Atari 2600 1977	Intellivision 1979	N.E.S. 1983	Atari Jaguar 1993	GameCube 2001	XBOX 360 2005
1.2 MHz	2.4 MHz	9.6 MHz	307.2 MHz	4,915 MHz	19,661 MHz
	×2	×2×2	×2×2×2×2×2	×2×2×2×2×2	×2×2

Explanation & Guiding Questions

Before turning students loose on this question, make sure they can articulate the rule "doubles every two years".

It is common for students to correctly double 1.2MHz and get 2.4 MHz in 1979, but then to continue adding 1.2 at a constant rate every two years. Most will self-correct as they check in with their neighbors, but be on the lookout for that misunderstanding of the pattern.

Once students have finished the table, and some have started to think about the next question, you can display the answers and prompt students to explain their reasoning.

- Restate Moore's Law in your own words.
- How many times should the processor speed have doubled between the release of the Intellivision and the release of the N.E.S.?
- What operation did you keep doing over and over again?
- Where did that 307.2 come from? How did you calculate that?

Deeper Understanding

- What's an easier way to write $\times 2 \times 2 \times 2 \times 2 \times 2$? ($\times 2^5$)
- In what year would Gordon Moore say a 76.8 MHz processor would be released? (1989, since $76.8 = 9.6 \times 2^3$, so 6 years after 1983.)