Northwestern | INSTITUTE FOR POLICY RESEARCH

Working Paper Series

WP-16-11

Simplifying Teaching: A Field Experiment with "Off-the-Shelf" Lessons

Kirabo Jackson
Associate Professor of Human Development and Social Policy and IPR Fellow
Northwestern University

Alexey Makarin
Graduate Student
Northwestern University

Version: July 2016

DRAFT

ABSTRACT

We analyze an experiment in which middle-school math teachers were randomly given access to "off-the-shelf" lessons designed to develop students' deep understanding. These lessons were provided online, but are designed to be taught by teachers in a traditional classroom setting. Teaching involves multiple complementary tasks, but we model two: imparting knowledge and developing understanding. In our model, lessons designed to develop understanding substitute for teacher effort on this task so that teachers who may only excel at imparting knowledge can be effective overall – simplifying the job of teaching. Providing teachers with online access to the lessons with supports to promote their use increased students' math achievement by about 0.08 of a standard deviation. These effects appear to be mediated by the lessons promoting deep understanding, and teachers therefore being able to provide more individualized attention. Benefits were much larger for weaker teachers, suggesting that weaker teachers compensated for skill deficiencies by substituting the lessons for their own efforts. The intervention is highly scalable and is more cost effective than most policies aimed at improving teacher quality.

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, little is known about how to effectively improve teacher effectiveness (Jackson et al., 2014). While many policies have been found to be ineffective, the most promising policies aimed at improving teacher effectiveness involve professional development to increase teacher skills, incentives to induce greater teacher effort, or the removal of ineffective teachers. Such policies tend to be costly, politically infeasible, or difficult to scale (Yoon et al., 2007; Rothstein, 2014; Gonser, 2016). These aforementioned approaches seek to alter teacher skill or effort, but take the production technology (i.e. how teachers turn skills and effort into performance) as given. We take the opposite approach. We propose an intervention that takes teacher skill and effort as given, but alters the production technology. Teachers in our intervention were granted access to a technology that simplified the complex multitask job of teaching and allowed for greater specialization. The intervention was designed to improve teacher effectiveness at low cost, and in a manner that is scalable and easily replicable.

The job-simplifying technology we study is high-quality off-the-shelf lessons. The lessons we examine are known as "anchor" lessons, and are designed to complement traditional classroom instruction. These anchor lessons were designed to build students' math intuition through real-life experiences that can be leveraged by the teacher throughout the year when teaching formal math concepts by saying "*remember when we did ...*" (Harvey and Goudvis, 2007). A single anchor lesson (which may be taught over several class periods) is designed to facilitate deep understanding of concepts covered over several weeks during subsequent class meetings. These anchor lessons are not typical of traditional math instruction which has been described as "learning terms and practicing procedures". 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, during these anchor lessons (described in Section II), students are guided by the teacher to think creatively and critically about the real world, and then expected to come up with their own mathematical models to better understand the world around them. This exploratory and inquiry-based approach to learning is hypothesized by education theorists to better develop students' deep understanding (see Dostál, 2015). However, experimental evidence of this using standardized achievement tests is limited.

In our intervention, teachers were randomly assigned to one of three treatment conditions. In the "license only" condition, teachers were given free access to these online lessons. These lessons were provided online, but 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 anchor lessons, received email reminders to use the lessons, and were invited to an online social media group focused on implementing the lessons. Finally, teachers randomly assigned to the control condition continued "business-as-usual."

To highlight the economics at play, we conceptualize teaching as being a complex job that involves multiple tasks (Holmstrom and Milgrom, 1991). For parsimony, we model two complementary tasks: imparting knowledge and developing understanding. We model student achievement as a function of both their knowledge of, and their understanding of, math content. As such, the best teachers are those that can perform both tasks simultaneously. Off-the-shelf lessons are a technology that guarantees a minimum quality of instruction to develop understanding. In our model, by allowing teachers who use this technology to focus all their effort on imparting knowledge (i.e. specialize), the technology simplifies the job of teaching (turning a two-task job into a single task). Even though technology is complementary to worker skill in many settings (e.g. Katz et al., 1999, Akerman et al., 2015), our model predicts that the benefits of lesson use are larger for the weaker teachers. We test this, and other, predictions empirically.

Because the treatments were assigned randomly, we identify causal effects using multiple regression. Teachers who were only granted free access to the lessons had low levels of take-up. However, on average, fully-treated teachers (access plus supports) looked at 4.56 more lessons, and taught 2.17 more lessons than control teachers. This represents looking at lessons that relate to about three-quarter's of a years' worth of material, and teaching lessons that relate to about one-third of a years' worth of material. Consistent with the take-up effects, students of teachers in the license only group experienced modest statistically insignificant increases in math test scores, while full treatment teachers increased average student test scores by about 0.08σ relative to those in the control condition. Though this may seem modest, this is 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 (Jackson et al., 2014). 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. The per-teacher average cost of the intervention was about $431, and each teacher has about 90 students on average. Back-of-the envelope calculations suggest that the test score effect of about 0.08σ would generate about $360,000 in present value of student future earnings. This implies a benefit-cost ratio of 835, and an internal rate of return far greater than that of well-known educational 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).

To test whether off-the-shelf lessons substitute for teacher skills, as implied by the model, we test for heterogeneous effects at different points in the distribution of teacher effectiveness. Using conditional quantile regression (Koenker and Bassett, 1978), the benefits of lesson use are the largest for the least effective teachers, and decrease monotonically with effectiveness (as measured by classroom value added). This is consistent with off-the-shelf lessons substituting for teacher ability to promote deep understanding. Our results suggest that allowing teachers to substitute for deficiencies in their skills with high-quality off-the-shelf instructional materials is a viable alternative to policies that remove the least effective teachers, incentive pay, or interventions that involve teacher training for remediation purposes.

To explore mechanisms, we analyze effects on student and teacher surveys. Consistent with 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 deep understanding (as predicted by the model), treated teachers give students more individual attention. Looking at the teacher surveys, we find little evidence that the intervention had an effect on teachers' pedagogical practices. This is consistent with teachers substituting on-line lessons for their own efforts at promoting deep understanding rather than the effects being due to teacher learning or increases in teacher skill. The fact that there were no other changes in teachers' pedagogical practices 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 that the benefits are driven by lesson use, 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 monotonically 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 greater. Since lesson use is voluntary, regular reminders and additional supports to use the lessons may be crucial. Indeed, we uncover patterns in the data that suggest that the relatively low levels of lesson use can be explained by 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). Our conclusion 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 of 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 others. As such, this paper adds to a nascent literature exploring the potential productivity benefits of teacher specialization in schools (e.g. Fryer, 2016). 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 et al., 2015). 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). In contrast, this paper examines whether technology can help teachers enhance their traditional teaching practices in a scalable and cost effective way.[1] Moreover, 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). Finally, we add to

[1]Similarly, Comi et al. (2016) find that effectiveness of technology at school depends on teachers' ability to incorporate it into their teaching practices.

a growing literature on how overcoming behavioral biases can improve interventions.

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

II The Intervention

The job simplifying technology at the heart of the intervention are off-the-shelf lessons. These lessons are from the Mathalicious curriculum.[2] 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.[3]

For example, in one of the more simple lessons titled "*New-Tritional Info*" (see Appendix J), 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 two class periods. Because most secondary school children are familiar with McDonalds and Le-Bron 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 K) is a more complex 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

[2] http://www.mathalicious.com/about

[3] Mathalicious 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 involves showing student 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 is 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.

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, becuase 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 these lessons are memorable lessons that 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 topic covered for 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, 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).[4] 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

[4] http://www.edmodo.com/

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.1 The Experiment

Three Virginia school districts participated in this study: Chesterfield, Henrico, and Hanover. In total, 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 randomly assigned to one of the three conditions described below:

Treatment Arm 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 reviewed 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.[5] Importantly, participation in all components of the treatment was entirely voluntary.

Treatment Arm 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, 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 Arm 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

[5]The 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).

and would not have been using it.

Randomization

We randomly assigned teachers to one of the three treatment conditions. Randomization ensured that 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). 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. Once students were assigned to classroom rosters, for two of the three districts we received student data for each teacher, including demographics and assessment data from the previous year. To facilitate district participation in the study, two of the districts had pre-selected teachers that they wished to receive access to the Mathalicious lessons. All teachers who were requested to receive licenses received one. Randomization was conducted within each district, conditional on whether the district had requested a license for the teacher. As such, treatment status was random conditional on both "requested" status and district. Accordingly, all models condition on district and "requested" status.[6] 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 yield a p-value below 0.05. This is consistent with sampling variability and indicates that the randomization was successful.

III Model

To inform our empirical work and facilitate interpretation of our results, we lay out a stylized model of teacher behavior that yields some basic predictions regarding which teachers may use the lessons under different conditions, and the effect of the treatments on student outcomes. For ease of exposition, we provide an intuitive graphical presentation of the model below, and provide formal mathematical arguments and proofs of all claims in Appendix B.

We model teachers as being akin to firms choosing the 'profit' maximizing mix of inputs given a fixed total cost and fixed input prices. Teachers produce student test scores (y_i), where i is a student in the teacher's class. Test scores depend on the teacher's allocation of time (T) between imparting knowledge (n) and developing deep understanding (d). Because teachers are involved in several tasks, n captures all tasks complementary to developing understanding. Teacher ability to introduce new concepts and develop understanding are modeled as input "prices" (p_n) and (p_d). These prices denote the amount of time needed to impart one unit of knowledge and develop one unit of understanding, respectively. All teachers have the same time allocation (T), but higher ability teachers have lower p's such

[6] Table A1 summarizes teacher participation by district, requested status, and treatment condition.

that they can produce more learning per unit of time.

Teachers maximize their students' average test scores by choosing how much time to spend imparting knowledge ($n \geq 0$) and how much time to spend on deep understanding ($d \geq 0$), subject to the time allocation (T) and the prices they face (p_n and p_d).[7] 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 average test scores are represented by the indifference curves. Higher average test scores are on higher indifference curves. 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 B, the optimal level of test scores increases monotonically with d and n, so that 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 model off-the-shelf lessons as a technology that guarantees a minimum level of understanding ($\underline{d}$). This is depicted in Panel B of Figure 1. The new technology leaves the isocost curve unchanged for all $d > \underline{d}$. However, it creates a flat portion of the isocost curve at $\underline{d}$ for all levels of n where $d < \underline{d}$.[8] The flat portion of the isocost curve reflects the fact that to do better than $\underline{d}$ a teacher must exert the effort to create $d > \underline{d}$ on her own. The basic idea is that even though time can be added, quality cannot. That is, no number of low-quality lessons will develop more deep understanding than a single high-quality lesson.[9]

With lesson access, teachers can either produce test scores as they would without the lessons, or they can delegate developing understanding to off-the-shelf lessons (so that $d = \underline{d}$) and focus all of their time imparting knowledge (so that $n = T/p_n$). The teacher adopts the technology if average test scores under this technology are greater than that without. All else equal, there is some threshold price $\hat{p}_d$ such that a teacher is indifferent between relying on herself (an optimal point on the original budget line) and adopting off-the-shelf lessons (point $\{\underline{d}, T/p_2\}$) - this situation is reflected by the isoquant curve (IC_1) that is both tangent to the original isocost and goes through the allocation of inputs under lesson adoption. Whenever price p_d is above $\hat{p}_d$ (a flatter isocost curve), adopting off-the-shelf lessons is strictly preferred to not. This is shown in Panel C of Figure 1. IC_2 represents the average test scores for a teacher with a high p_d under the no lesson allocation. This teacher can attain $IC_1 > IC_2$ with lesson use and will therefore choose to adopt them. The distance between IC_1 and

[7] This discussion corresponds to a special case analyzed in Section B2, where we study the teacher objective function with a specific functional form, i.e. a weighted average of student test scores. All of our predictions, except for one, were proved to hold for any monotonically increasing function of students' test scores (see discussion in Section B3).

[8] How off-the-shelf lessons affect the choice of how to allocate time (T) between (d) and (n) is identical to how the basic welfare system alters the decision of how to allocate one's endowment of time between leisure and labor (e.g. Ehrenberg and Smith, 2009, Chapter 6).

[9] For example, if a student has already attended a high quality lesson that develops $\underline{d}$ of understanding of fractions, attending another lecture on fractions with d of understanding will not add to their understanding of fractions unless $d > \underline{d}$. This example makes clear that, to do better than $\underline{d}$, a teacher must exert the effort to create $d > \underline{d}$ on her own.

IC_2 denotes the teacher's gain from off-the-shelf lesson use. The higher is p_d, the greater these gains are. As shown in Appendix B, if higher levels of p_d and p_n are not negatively correlated in the population (i.e. teachers who are good at n are not bad at d, and vice versa), lower quality teachers (i.e. those that produce lower average test scores) gain the most from off-the-shelf lessons and may be most likely to adopt them.

Figure 1 reveals additional predictions from the model. First, all teachers who adopt the lessons will spend more time doing complementary tasks (n). Because the point $\{\underline{d}, T/p_2\}$ entails the maximum feasible amount of time on n, and the optimal level of d without lessons is greater than zero, it follows that $n^* < T/p_2$ so that teachers who use the lessons spend more time imparting knowledge. The second prediction is that the effect on deep understanding is ambiguous. From Figure 1, all teachers for whom $d^* \leq \underline{d}$ lesson use will increase d. However, some teachers with $d^* > \underline{d}$ will use the lessons (and use the increase in n to compensate for the reduction in d such that test scores do not decrease). As such, the overall effect on d is ambiguous in sign. However, whether d increases overall will depend on lesson quality in predicable ways. Specifically, with a smooth distribution of prices, the higher quality the lessons (higher $\underline{d}$), the higher the fraction of teachers for whom d increases, the larger the increase in d for these teachers, and the lower the possible reduction in d for those who reduce d. As such, for sufficiently high $\underline{d}$, we expect deep understanding to increase with lesson use.

The model yields the following five predictions that we test empirically:

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 (imparting knowledge/all tasks complementary to deep understanding) should increase.

Prediction 3: Among the students of teachers who chose to adopt lessons, the effect on d (deep understanding) is ambiguous, but is positive if lessons quality is sufficiently high.

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

Prediction 5: Absent external nudges or reminders (i.e. the license only condition), if teachers know their ability, less effective teachers (as measured by ability to raise average test scores) will be more likely to use the lessons.

IV 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 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. Only one of the 17 pre-treatment student or teacher characteristics is statistically significantly different across the three treatment conditions at the 5 percent level. This is consistent with a successful randomization.

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] This raw mean difference of 0.163σ between the full treatment and the control group is not statistically significant. However, it telegraphs the more precise multiple regression estimates we present in Section VI.

We supplement administrative records with data from other sources to measure lesson use, and to uncover underlying mechanisms. Our two measures of Mathalicious lesson use are the number of lessons looked at, and the number of lessons taught. Using teacher survey data, we observe the self-reported lessons they taught and read. Because these data are from surveys, using these data 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 VI. To improve the quality of our looked at measure, we also use the more objective measure of lessons downloaded. 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.[12] The lessons taught measure comes exclusively from survey reports.

To explore causal mechanisms, surveys were given to both teachers and their students. 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). Teacher surveys were administered in the middle, and at the end, of the

[10] In Hanover district, the exam codes were not provided so that the test scores are standardized by grade and year only. All results are robust to interacting incoming test score with district.

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

[12] We use multiple imputation (described in Section VI) to fill in missing self-reported data and then add in the tracking data to each multiple imputation replication

intervention year in all three school districts. We will focus on the end of year surveys. The teacher surveys were designed to measure teacher job satisfaction and classroom practices. 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.

V 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 difference in outcomes between the control and the license only groups, and between the control and the full treatment groups, respectively. The 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 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

[13] Intuitively, 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).

(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:

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

In (2), X_{idt} includes student race, student gender, teacher level averages of the student-level covariates (including lagged 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. The fact that the group mean differences in Table 1 and the precisely estimated regression estimates from (2) are of similar magnitude underscores this point. 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.

VI Results

Before presenting effects on teacher and student outcomes, we first analyze the effects of the randomized treatment conditions on Mathalicious lesson use. We then present effects on students' math test scores and on survey responses to uncover underlying mechanisms.

VI.1 Effects on Mathalicious lesson use

We have two sources of data to measure Mathalicious use. First, we rely on measures of which Mathalicious lessons were taught or read. This information was self-reported during the mid-of-year and end-of-year surveys. Second, we use the data received from Mathalicious site logs on whether a teacher downloaded a certain lesson or not. With this information, we construct two measures: the number of Mathalicious lessons taught by the teacher and the number of Mathalicious lessons the teacher looked at (either taught, read, or downloaded).

Because our measures of lesson use are (partially) obtained from survey data, we only have lesson use for individuals who completed the surveys during both waves. Because lesson use is zero in the control condition, imputing zero Mathalicious 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. To deal with the missing data problem more effectively, we use multiple imputation (Rubin, 2004; Schafer, 1997) to impute lesson use for those individuals who did not complete the surveys.[14] 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 condition from a Poisson

[14] We present lower bound estimates assuming zero adoption for non-responding teachers in Appendix D.

regression (note that these are count data). For the lessons looked at, we conduct multiple imputation for the survey responses before combining it with the tracker data. To see how much this matters, we also present naïve results using only those with survey data. Such results are similar to, but less precise than, those using multiple imputation.

The regression results are presented in Table 2. Panel A presents results for the subsample of teachers with complete data on Mathalicious use, while Panel D reports results that use multiple imputation for missing data. As the results are largely similar across the two panels, we focus our discussion on Panel D as it makes the appropriate adjustment for missing data. Models 13 and 15 present results that only control for district fixed effects interacted with an indicator for whether a teacher was requested to receive a license. Columns 14 and 16 include district fixed effects interacted with a 'requested' indicator, as well as other teacher-level and student-level covariates mentioned in Section V.

Across these models there is a clear pattern. 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. In the preferred model with the full set of controls (columns 14 and 16), average full treatment teachers taught 2.179 more Mathalicious lessons (p-value<0.01) and looked at 4.563 more lessons (p-value<0.01) than teachers in the comparison group. License only teachers looked at 1.604 more lessons on average than control teachers (p-value<0.01) and taught 0.595 more Mathalicious lessons on average (p-value<0.01). Note that the differences between the two treatment arms are statistically significant at the 5 percent level. As one can see in Panel A, the results using only teachers with non-missing data are similar (and indistinguishable in a statistical sense). To show the importance of using only teachers with complete data, Panels B and C use data for any teacher who filled in at least one survey (but not necessarily both). Though this increases sample size, it predictably increases attenuation bias. In models that use multiple imputation (Panel C) and those that do not (Panel B), the estimated effects on lesson use are smaller. 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 lager and more credible estimates in Panels A and D that rely on teachers with complete data on lesson use.

To put these effect 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 Mathalicious 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.

VI.2 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 3. While the effect of providing licenses only is positive and not statistically significantly different from zero, the full treatment had a positive significant effect on students' math test scores. The first model (columns 1 and 3) includes the key conditioning variables (district fixed effects and requested status) and the average lagged math scores in the classroom. Looking at raw test scores, this parsimonious model (Column 1) shows that scores were 7.618 points higher among teachers in the full treatment condition than the control condition (p-value<0.05). 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. Using standardized scores in this model (column 3), teachers who only had access to the lessons had test scores that were 2% of a standard deviation higher than those in the control condition. Looking at the full treatment condition (where we observed a more robust increase in lesson use), teachers with access to both Mathalicious lessons and extra supports increased their students' test scores by 10.2% of a standard deviation relative to those in the control condition (p-value<0.1). 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 25 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 5.3% of a standard deviation higher than those in the control condition. This modest positive effect is consistent with the positive effect on lessons taught. Looking at the full treatment condition, teachers with access to both Mathalicious lessons and extra supports increased their students' test scores by 7.8% of a standard deviation relative to those in the control condition (p-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 is 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 4.7% of a standard deviation relative to those in the control condition (not statistically significant), and full treatment condition increased their students' test scores by 9.1% of a standard deviation relative to those in the control condition (p-value<0.05). In sum, across all the models, there is a robust positive effect of the full treatment (relative to the control condition) on student test scores of between 7.7 and 10.2 percent of a standard deviation.

To assuage any concerns that the estimated effects are spurious, we report a falsifica-

tion exercise with English test scores as a 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 or Hawthorne effects.

VII Mechanisms

Section VI established that the treatment lead teachers to use Mathalicious lessons as intended, and that student achievement in mathematics improved in the full treatment condition where lesson use was most robust. In this section, we provide evidence on the mechanisms, and empirically test the key predictions from the model.

VII.1 Effects on Student Perceptions and Attitudes

By changing the nature of teacher instruction, Mathalicious lessons could alter student attitudes toward mathematics. To test this, we analyze effects on student responses in an anonymous survey given at the end of the experiment. We asked several questions on a Likert scale and used factor analysis to extract common variation from similar questions. When grouping questions measuring the same construct, each group of questions is explained by only one factor. After grouping similar questions, we ended up with 6 distinct factors.[15] Each factor is standardized to be mean zero, unit variance. We discuss each of them in turn.

Using the specification in (1), we estimate the effect on student responses to the survey items. As with test scores, we analyze the student surveys at the student level. Table 4 presents the results. We present results with no controls in Panel A, and results from models that include the full set of controls in Panel B – the results are almost identical. 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 (Specifications 1 and 8) are models where the dependent variable is the survey nonresponse rate.[16] Overall, the survey response rate was

[15]Factor loadings for each individual question are presented in Appendix C.

[16]For 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 missing data for each teacher and weight the

58 percent. Importantly, the survey nonresponse rates are similar across the three treatment arms. In models with no controls, the estimated effect on survey non-response is small for both treatment arms and neither is statistically significantly different from zero. In the models with all controls, the coefficient on the full treatment is larger, but is not statistically significantly different from zero. Reassuringly, the survey results are very similar across both models, so that one can be reasonably confident that any differences in response to questions are not driven by differential non-response.

Given the similarity in the results with and without controls, we focus our discussion on the results with all controls in Panel B. The first factor measures whether students believe that math has real life applications. The results in Specification 9 of Table 4 show that 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 0.183σ more than those of control teachers (p-value$<$0.01). This is consistent with the substance and stated aims of the Mathalicious lessons. This result implies that the content of the Mathalicious lessons 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 economically meaningfully higher levels of interest in math (0.086σ). However, this effect is not statistically significant at traditional levels. The estimated coefficient on effort in math class is a precisely estimated zero, while there is a small positive effect on the general motivation to study. None of the effects on these factors are statistically significant, but the magnitude 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 model and the specific aims of the Mathalicious lessons. Although our model does not yield a clear prediction on whether adoption of the off-the-shelf lessons increases the level of d for all teachers,[17] the model does predict that if the lessons are of high enough quality, there would be higher levels of d among the treated teachers, on average. 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 promoting deep understanding (though we refer to this as imparting knowledge in our model). There are many such tasks, but we hypothesize that providing one-on-one time is one such complementary task. 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.[18] The results support the premise of our model that Mathalicious lessons promote

regressions by the total number of students with the teacher.

[17] One can see from Figure 1 that d could actually fall after adopting the technology for some levels of p_d.

[18] Jackson (2016a) also uses more one-on-one time as a measure of teacher time. He finds that in more

deep understanding. Students from the full treatment group are 0.182σ (p-value<0.05) more likely to agree that their math teacher promotes deep understanding. As the model predicts, this indicates that the lessons were of sufficiently high quality. 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 0.159σ higher in the full treatment condition than in the control condition (p-value<0.05). For both these factors in the lesson only condition, consistent with the modest increase in lessons taught, the effect is positive and modest but not statistically significant.

The survey evidence shows that, among students whose teachers used the Mathalicious lessons most robustly, student perceptions regarding math and their math teachers changed in the expected directions. Students 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. As an additional test, we report results on the mid-year student survey that was conducted two months after the experiment began in Appendix E. Because students were exposed to the treatment, but for a shorter amount of time, one might expect patterns similar to those in Table 4, but effects that are smaller in magnitude. This is precisely what one observes. This provides further evidence that the survey results presented are not driven by differential survey response, operate through the hypothesized channels, and reflect causal effects.

We 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 slightly higher than on 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 teacher survey results in Appendix F.

VII.2 Effect Heterogeneity by Teacher Quality

One of our key theoretical predictions is that the gains from off-the-lessons increase with teacher effectiveness. That is, weaker teachers who are relatively ineffective at improving student performance may benefit greatly from the provision of off-the-shelf lessons. To test for this, we see if the marginal effect of the treatment is larger 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.

homogeneous classrooms, teachers spend more on-on-one time with students likely due to time savings.

(2014b), 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 different effects for classrooms at different 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 show in Appendix G, when average test scores at the teacher level is the dependent variable, the teacher-level residual from (1) is precicely the standard value-added measure of classroom quality. 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. To assuage any concerns that the teacher-level model yields different results from the student-level model, Appendix Table H1 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.

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 5th through 95th 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.

If our hypothesized mechanisms are in effect, one would expect large positive effects for the low quantiles of classroom/teacher quality that then decline for higher quantiles. This is exactly what we observe. Indeed, 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 -0.00085 (p-value=0.000) 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*0.00085 = 0.043σ. 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. However, the model indicates that if teachers act in the best interest of their students, and they know their own ability, the effect of adopting the lessons (among those who voluntarily chose to adopt them) should be positive. As such, the marginal effect of the full treatment should be nonnegative for all teachers. This prediction is consistent with the data. Specifically, even at the 95th and 99th percentile of classroom quality, the point estimates are positive (albeit not statistically different from zero). As a falsification exercise, we estimate the same quantile regression model for English test scores (see Appendix I). As one would expect, there is no systematic relationship for English scores, and the estimated point estimates for English are never statistically significantly different from zero. This provides further evidence that the estimated effects on math scores are real.

The last remaining prediction from the 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 could not test this convincingly due to limitations in our data. Specifically, testing for heterogeneous effects on lessons taught by teacher quality within the lesson only condition requires that there is sufficient data on lessons taught within each treatment arm teacher quality cell. While the data we have were sufficient to analyze average effects across the three treatment arms, there was not enough data to credibly detect effects by teacher quality within each treatment arm.[19] As such, we are unable to provide any firm evidence on this last prediction of the model. Given that all the other theoretical predictions are consistent with the data, we see the weight of the evidence as supportive of the model.

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 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 in each of the three districts. This results in six separate treatments.[20] 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 0.035 (p-value$<$0.05). This suggests that for every additional lesson taught test scores increase by 0.035σ. 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 because they had larger effect on lesson use.

[19] For suggestive evidence, we attempted to use the partially complete lesson taught data. However, the patterns of the results were sensitive to small changes in specification and therefore inconclusive.

[20] When estimating effects on lessons taught, we use multiple imputation as outlined in Section VI.

As an additional test, 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 5. 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.

VII.4 Patterns of Lesson Use

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 of other interesting patterns. They document 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.

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 maybe due to individuals underestimating the odds that they will be impatient in the future and then procrastinating (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 much 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 much more rapid. If the reason for the drop-off was low lesson quality, drop-off should have been very rapid for both groups. The second piece of evidence is that the most rapid attrition from lesson use in the full treatment condition occurs after February when Mathalicious ceased sending out email reminders to teachers. 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% state that is 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 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 empirical patterns and the survey evidence is that, without the reminders and extra supports, teachers were unable to hold themselves to make the time to implement the lessons. This suggests that overcoming behavioral biases played a key role in increasing lesson use over the control condition. 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 be reduce the incentive to procrastinate in 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. We designed an intervention that allowed teachers to "delegate" some of their tasks using off-the-shelf lessons so that they could focus their efforts on other duties. This intervention simplified the complex job of teaching by eliminating the need to perform one of the many tasks entailed in quality teaching. 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 others. As such, this paper provides evidence on

the productivity benefits of teacher specialization in schools.

The online "off-the-shelf" lessons provided 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.

Simply offering the lessons for free had modest effects on lesson use and therefore small effects on test scores. However, fully-treated teachers (who also received on-line supports to promote lesson use) used the lessons and improved their students' test scores by about 0.08σ relative to the 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 efforts to develop deep understanding, 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 necessary 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 are relatively low. An upper bound estimate for the cost of the program is \$431 per teacher.[21] Chetty et al. (2014b) 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 0.08σ would generate roughly \$4,000 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 \$360,000 in present value of students' future earnings. This implies a benefit-cost ratio of 835. Because of the low marginal cost of the intervention, it is extraordinarily cost effective. Furthermore, because the lessons and supports are provided on the Internet, the intervention is highly scalable, and can be implemented in remote locations where other policy approaches would be infeasible.

More generally, our findings show that simplifying the complex job of teaching 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 et al., 2015). They also suggest that policies aiming to change the production technology (such as changes in curriculum design, innovative instructional materials, and others) may be fruitful avenues for

[21] The price of an annual Mathalicious subscription is \$320. The cost of providing the additional supports (extra time for Mathalicious staff time to run Project Groundswell etc.) 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.

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.

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.

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 II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9): 2633–2679, 2014a.

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, 2014b.

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*. Palackỳ 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. Ehrenberg and R. Smith. *Modern labor economics: Theory and public policy*. 2009.

R. G. Fryer. The 'pupil' factory: Specialization and the production of human capital in schools. *NBER Working Paper w22205*, 2016.

S. Gonser. This may be the best way to train teachers, but can we afford it? *Huffington Post*, May 2016.

S. Harvey and A. Goudvis. *Strategies that work: Teaching comprehension for understanding and engagement*. Stenhouse Publishers, 2007.

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.

T. J. Kane and D. O. Staiger. Estimating teacher impacts on student achievement: An experimental evaluation. *NBER Working Paper w14607*, December 2008.

L. F. Katz et al. 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. Ö. 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.

K. Muralidharan and V. Sundararaman. Contract teachers: Experimental evidence from India. *NBER Working Paper w19440*, September 2013.

T. O'Donoghue and M. Rabin. Doing it now or later. *American Economic Review*, pages 103–124, 1999.

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. Revisiting the impacts of teachers. *UC-Berkeley Working Paper*, 2014.

J. Rothstein et al. 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.

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.

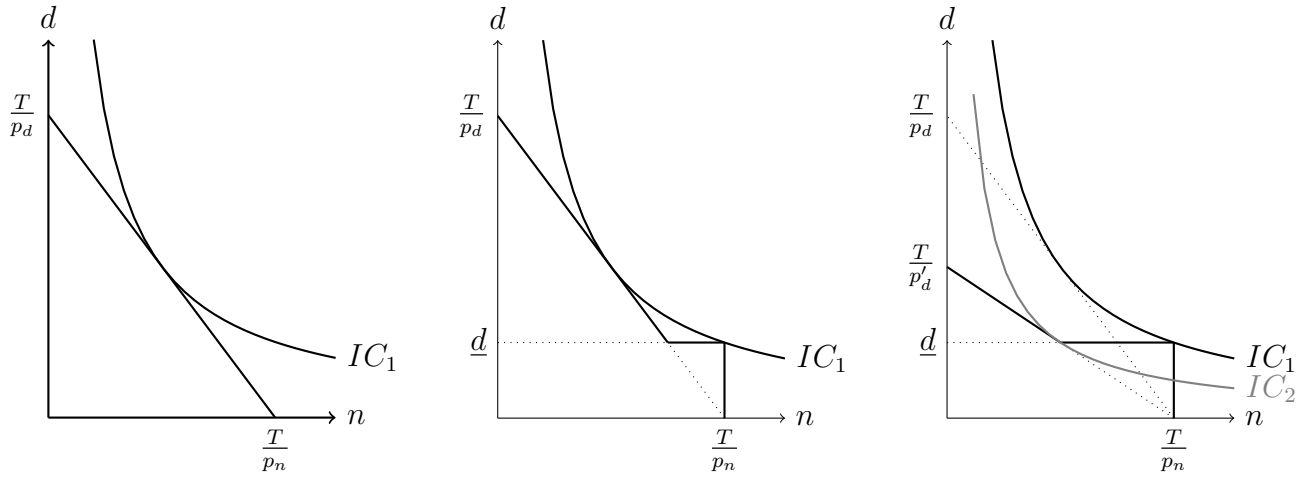
E. S. Taylor and J. H. Tyler. The effect of evaluation on teacher performance. *The American Economic Review*, 102(7):3628–3651, 2012.

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.

K. S. Yoon, T. Duncan, S. W.-Y. Lee, B. Scarloss, and K. L. Shapley. Reviewing the evidence on how teacher professional development affects student achievement. Issues & answers. *Regional Educational Laboratory Southwest (NJ1)*, 2007.

Tables and Figures

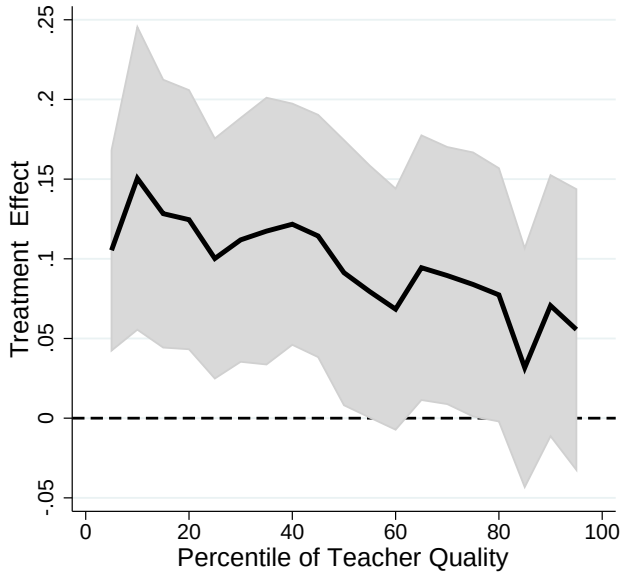
Figure 1: Illustration of the Model.



(a) No off-the-shelf lessons (b) With off-the-shelf lessons (c) Gains from adoption

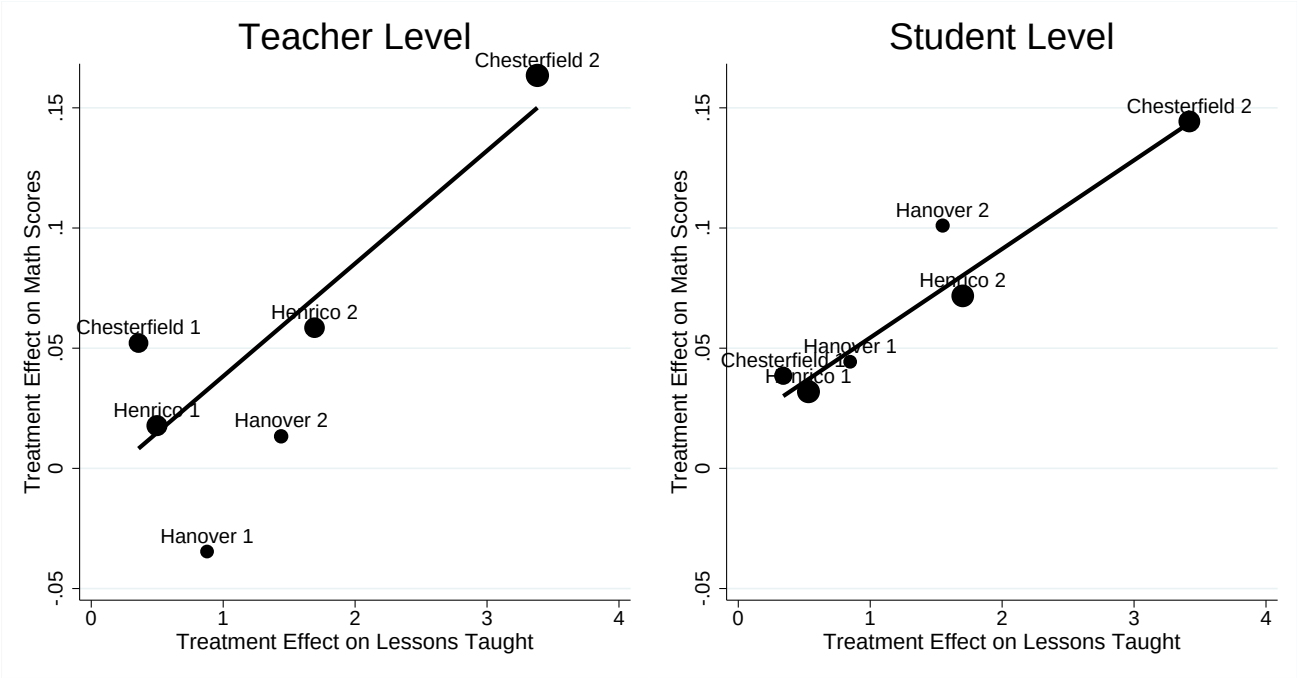
Notes: This is an Illustration of a stylized version of the model presented in Section B2.

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



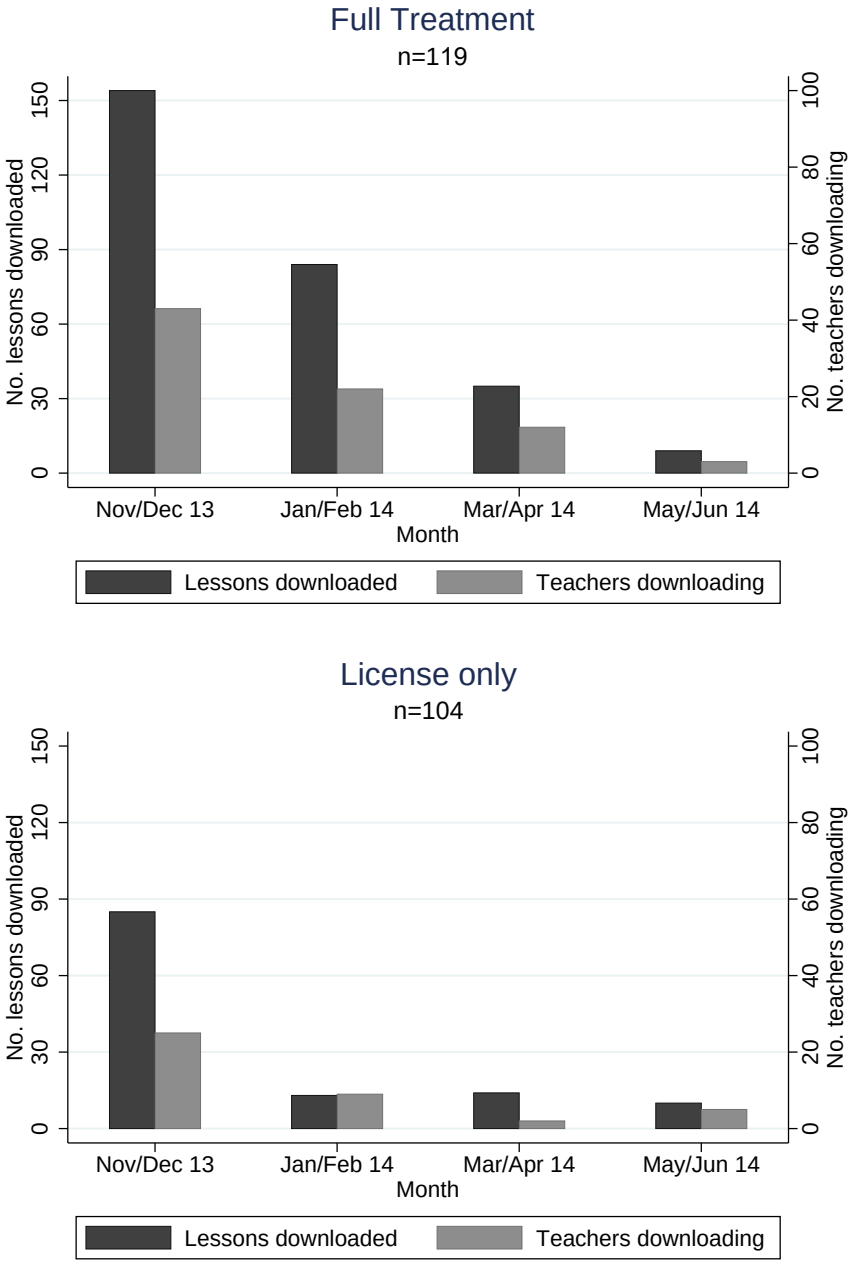
Notes: The solid black line represents the treatment effect estimates from estimating equation (1) using conditional quantile regression. The dependent variable is the teacher-level average standardized 2014 math test scores. The shaded area depicts the 90% confidence interval for each conditional quantile regression estimate. For a formal discussion of the method, see Appendix G.

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



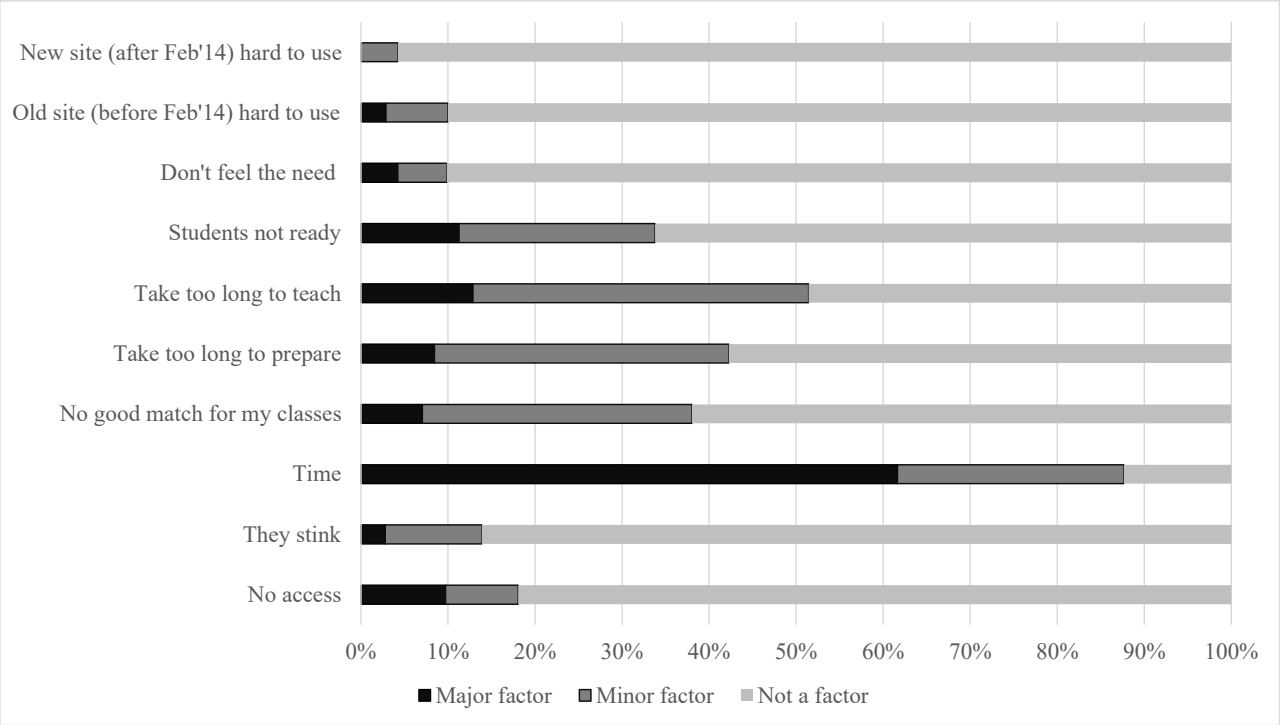
Notes: This figure plots average treatment effects on lesson use and standardized math scores, separately by district and by treatment. Chesterfield, Hanover, and Henrico are the school districts in Virginia where the intervention took place. The 'License only' treatment is denoted by the number 1, and the 'Full Treatment' is denoted by the number 2. In each panel, the Y-axes displays coefficients for specifications identical to those estimated in Columns (5) and (6) of Table 3 for student and teacher level models, respectively. The X-axes displays coefficients for specifications similar to those estimated in Panel D Column (16) of Table 2. 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 observations (students or teachers).

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 comes 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[c]	0.008	0.997	-0.021	-0.027	0.068	0.969

*** - 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 a detailed 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).

Table 2. Effects on Lesson Use.

Panel A: Baseline results. Analysis on teachers with non-missing outcomes (20%).

	Lesson Looked	Lesson Looked	Lessons Taught	Lessons Taught
	(1)	(2)	(3)	(4)
License Only	3.310	4.682	0.400	0.651
	[2.609]	[5.234]	[0.653]	[1.589]
Full Treatment	4.635*	6.672	1.522*	2.424
	[2.440]	[5.029]	[0.878]	[1.848]
District FE x Requested	Y	Y	Y	Y
All controls	N	Y	N	Y
Observations	69	69	69	69

Panel B: Baseline results. Analysis on teachers with partially-missing and non-missing outcomes (60%).

	Lesson Looked	Lesson Looked	Lessons Taught	Lessons Taught
	(5)	(6)	(7)	(8)
License Only	1.552***	1.729***	0.537	0.502
	[0.587]	[0.582]	[0.401]	[0.380]
Full Treatment	2.984***	3.255***	1.001***	1.021***
	[0.660]	[0.686]	[0.366]	[0.388]
District FE x Requested	Y	Y	Y	Y
All controls	N	Y	N	Y
Observations	256	256	236	236

Panel C: Multiple imputation estimates. Missing outcome data imputed using multiple imputation.

	Lesson Looked	Lesson Looked	Lessons Taught	Lessons Taught
	(9)	(10)	(11)	(12)
License Only	1.396***	1.558***	0.489*	0.491**
	[0.490]	[0.484]	[0.256]	[0.235]
Full Treatment	3.216***	3.438***	1.054***	1.049***
	[0.607]	[0.629]	[0.260]	[0.262]
District FE x Requested	Y	Y	Y	Y
All controls	N	Y	N	Y
Observations	363	363	363	363

Panel D: Multiple Imputation estimates. Missing and partially-missing outcome data imputed using multiple imputation.

	Lesson Looked	Lesson Looked	Lessons Taught	Lessons Taught
	(13)	(14)	(15)	(16)
License Only	1.422***	1.604***	0.534***	0.595***
	[0.360]	[0.391]	[0.157]	[0.187]
Full Treatment	4.326***	4.563***	2.138***	2.179***
	[0.585]	[0.606]	[0.528]	[0.524]
District FE x Requested	Y	Y	Y	Y
All controls	N	Y	N	Y
Observations	363	363	363	363

*** - significance at less than 1%; ** - significance at 5%; * - significance at 10%

Robust standard errors are reported in square brackets. Standard errors in Panels C and D are corrected for multiple imputation according to Rubin (2004). All specifications include controls for district fixed effects interacted with the requested indicator. 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. Outcomes were missing for some teachers because of survey non-response: 69 teachers completed both mid-year and end-of-year surveys (examined in Panel A), 256 teachers completed either of the two (see Panel A). Panels C and D use data from these 256 and 69 teachers, respectively, 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.

Table 3. 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	0.976	3.068	0.020	0.052	0.053	0.047	0.502	0.014
	[2.578]	[2.028]	[0.046]	[0.035]	[0.036]	[0.033]	[1.626]	[0.029]
Full Treatment	7.618**	6.293**	0.102*	0.087**	0.078**	0.091**	-0.050	-0.002
	[3.214]	[2.543]	[0.054]	[0.040]	[0.0396]	[0.038]	[2.044]	[0.037]
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	24,409	24,409
Unit of Observation	Student	Student	Student	Student	Student	Teacher	Student	Student

*** - 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 test scores, and a teacher-level share of students with missing 2013 math 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 math scores are given an imputed score of zero. To account for this in regression models, we also include an indicator denoting these individuals in all specifications. The student level controls include individual-level math test scores, race, and gender. 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 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 4. Students' Post-Treatment Survey Analysis (Chesterfield and Hanover only).

	Share of Missing 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	-0.038	-0.019	-0.031	0.013	-0.028	0.043	0.088
	[0.071]	[0.065]	[0.069]	[0.042]	[0.050]	[0.070]	[0.070]
Full Treatment	0.014	0.150**	0.053	0.026	0.052	0.192**	0.174**
	[0.066]	[0.071]	[0.071]	[0.042]	[0.047]	[0.075]	[0.067]
District FE x Requested	N	N	N	N	N	N	N
All controls	N	N	N	N	N	N	N
Observations	15,844	10,034	10,111	9,994	10,057	10,389	10,849
Panel B. With All Controls.							
	(8)	(9)	(10)	(11)	(12)	(13)	(14)
License Only	-0.015	0.008	0.012	0.041	-0.028	0.026	0.079
	[0.066]	[0.056]	[0.061]	[0.035]	[0.036]	[0.064]	[0.062]
Full Treatment	0.113	0.183***	0.086	-0.002	0.027	0.182**	0.159**
	[0.069]	[0.065]	[0.076]	[0.048]	[0.045]	[0.075]	[0.075]
District FE x Requested	Y	Y	Y	Y	Y	Y	Y
All controls	Y	Y	Y	Y	Y	Y	Y
Observations	15,844	10,034	10,111	9,994	10,057	10,389	10,849

*** - significance at less than 1%; ** - significance at 5%, * - significance at 10%.
Standard errors clustered at the teacher level are reported in square brackets. Each outcome, except for the share of missing 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 indicators. The specifications in Panel B add controls for district fixed effects 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 teacher level where each teacher is weighted by the total number of her students. The share of missing 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. When the number of students with completed surveys exceeded the number of students with complete data, negative shares of missing values were replaced with zeros.

Table 5. 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	0.035*	0.033**	0.045	0.042**	0.041**	0.046	0.000	0.008
	[0.019]	[0.017]	[0.037]	[0.019]	[0.018]	[0.037]	[0.016]	[0.015]
Full Treatment			-0.029			-0.013		
			[0.083]			[0.076]		
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	24.39	37.56	4.492	15.51	16.69	3.252	21.72	43.05
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

*** - 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 test scores, and a teacher-level share of students with missing 2013 math 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 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 A. Treatment Allocation.

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

Treatment By District					
	Control	License Only	Full Treatment	Total	Requested
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. Stylized Model of Teacher Multitasking.

B1. General set-up

Let us consider the general optimization problem for a teacher. 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 imparting knowledge (n) and developing deep understanding (d). Teacher's (in)ability to impart knowledge is modeled as a 'price' p_n that amplifies the time needed to achieve n units of knowledge. Similarly, 'price' p_d denotes the teacher's ability to develop deep understanding. Note that the higher teacher abilities are, the lower are her corresponding p's. Formally, we write:

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

B2. Special Case with Functional Form Assumptions

To illustrate the main ideas behind the model, let us consider a special case 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:[22]

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

B2.1. Solving the model without off-the-shelf lessons

Using standard arguments for a Cobb-Douglas utility function, we get the following optimal allocation:

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

[22] $A_i > 0$ 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.

The optimal level of test scores is:

$$U(n^*, d^*) = \sum_{i=1}^{s} \frac{w_i A_i T}{s} \left[\frac{\alpha^\alpha (1-\alpha)^{1-\alpha}}{p_n^\alpha p_d^{1-\alpha}} \right] \tag{3}$$

Note that average test scores decrease both in p_n and p_d, i.e. increases 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.

B2.2. Solving the model with off-the-shelf lessons

We model off-the-shelf lessons as a technology that guarantees a minimum quality of instruction to provide a minimum level of understanding ($\underline{d}$). Teachers can now either stick to their own efforts or they can delegate developing deep understanding to off-the-shelf lessons and spend the rest of their time imparting knowledge, i.e. $n = T/p_n$. 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:

$$\sum_{i=1}^{s} w_i A_i T \left[\frac{\alpha^\alpha (1-\alpha)^{1-\alpha}}{\hat{p}_n^\alpha \hat{p}_d^{1-\alpha}} \right] = \sum_{i=1}^{s} w_i A_i \left(\frac{T}{\hat{p}_n} \right)^\alpha \underline{d}^{1-\alpha} \tag{4}$$

, with the maximum utility level with no off-the-shelf lessons on the left-hand side and the maximum utility after adopting off-the-shelf lessons on the right-hand side. Since $\hat{p}_n$ can be canceled from both sides, the threshold set now only consists of $\hat{p}_d$. Solving for $\hat{p}_d$ leads to:

$$\hat{p}_d = \left(\frac{T}{\underline{d}} \right)^\alpha \left[\alpha^\alpha (1-\alpha)^{1-\alpha} \right]$$

Thus, for the case of the Cobb-Douglas utility function, the adoption decision fully depends on the ability to develop deep understanding. That is, teachers with $p_d \geq \hat{p}_d$ adopt off-the-shelf lessons according to our model, while teachers with $p_d < \hat{p}_d$ stick with their own efforts. This convenient property informed the construction of Figure 1.

B2.3. Changes in d as a result of adopting off-the-shelf lessons

This section explores *Prediction 3* from Section III within this special case. First, let us compute d^* that a teacher with a threshold skill $\hat{p}_d$ was getting without off-the-shelf lessons:

$$d^*(\hat{p}_d) = \left(\frac{1-\alpha}{\alpha} \right)^\alpha T^{1-\alpha} \underline{d}^\alpha$$

As d^* is strictly decreasing in p_d, all teachers with $d^* < d^*(\hat{p}_d)$ will adopt off-the-shelf lessons. However, it is not always true that $d^*(\hat{p}_d)$ is greater than $\underline{d}$, meaning that not everyone who adopts off-the-shelf lessons was producing a lower level of d than the one guaranteed by off-the-shelf lessons. Specifically, whenever $\underline{d} < ((1-\alpha)/\alpha)^{\alpha/(1-\alpha)}T$, i.e. when quality of off-the-shelf lessons is relatively low, there will be some teachers who adopt the lessons and, as a result, decrease their production of d. As such, the overall effect of lesson use on d will depend on how many teachers experience an increase in d, how many experience a decrease in d, and the relative magnitudes of those increases and decreases in d. However, if $\underline{d} \geq ((1-\alpha)/\alpha)^{\alpha/(1-\alpha)}T$, i.e. when the quality of off-the-shelf lessons is sufficiently high, every teacher who adopts the lessons increases the level of deep understanding produced (d).

B3. General model: testable predictions

We now turn to a more general setting, in which we do not make parametric assumptions regarding the form of the teachers objective function. Instead, throughout the section, we will make use of the following assumptions:

(A1) $U(\cdot)$ and $y_i(n,d)$ are continuous and differentiable on $\mathbb{R}^s$ and $\mathbb{R}^2_+$ respectively, $\forall i \in S$

(A2) $U_i'(\cdot) > 0$, $\partial y_i/\partial n > 0$, and $\partial y_i/\partial d > 0$, $\forall (n,d) \in \mathbb{R}^2_+$ and $\forall i \in S$

This assumption states that teachers derive positive utility from increasing the test scores of their students. In turn, students' test scores are monotone with respect to the quality of teaching.

(A3) $\lim_{k\to\infty} \partial y_i/\partial k = 0$, $\lim_{k\to 0} \partial y_i/\partial k = \infty$, where $k \in \{n,d\}$, $\forall i \in S$

This assumption guarantees that all y_i's satisfy Inada conditions. Inada conditions on test scores functions guarantee that it will never be optimal for a teacher to set either n or d to zero.

(A4) $y_i(0,d) = y_i(0,d')$, $\forall d, d' \in \mathbb{R}_+$

This assumption states that, when a teacher imparts zero knowledge, increasing deep understanding has no effect on test scores.

(A5) Teacher ability parameters p_n and p_d are distributed among the universe of teachers according to the continuous cumulative distribution functions $F_{P_n}(p_n)$ and $F_{P_d}(p_d)$

(A6) $F_{P_n|P_d}(\cdot|p_d) > F_{P_n|P_d}(\cdot|p_d')$, $\forall p_d' > p_d$

This assumption assumes that conditional distribution of p_n given p_d' first order stochastically dominates conditional distribution of p_n for any smaller p_d. Intuitively, this implies a positive relationship between different sets of skills in the population of teachers, i.e. teachers with lower p_d's will on average have lower p_s's, and vice versa.

In addition to assumptions (A1)-(A6), we will use simplified notation, in which teacher utility directly depends on n and d without explicit consideration of student test scores. That is, we will write $U(n,d)$ instead of $U(y_1(n,d),...,y_S(n,d))$. In this simplified notation, assumptions (A1)-(A3) naturally transform into assumptions (A'1)-(A'3):

(A'1) $U(\cdot)$ is continuous and differentiable on $\mathbb{R}^2$

(A'2) $\partial U/\partial n > 0$, and $\partial U/\partial d > 0$, $\forall (n,d) \in \mathbb{R}^2_+$

(A'3) $\lim_{k\to\infty} \partial U/\partial k = 0$, $\lim_{k\to 0} \partial U/\partial k = \infty$, where $k \in \{n,d\}$

In a general model, we derive three empirically testable predictions listed in Section III. One can see that *Prediction 1* follows directly from the voluntary nature of technology adoption in our experiment and the monotonicity assumption (A2). It would not be rational for the teacher in our model to use off-the-shelf lessons if it did not increase student test scores, as teacher utility depends solely on the test scores. *Prediction 2* is also a direct consequence of the way we model off-the-shelf lessons - since off-the-shelf lessons guarantee $\underline{d}$ at a zero time cost for the teacher, the teacher can now spend all the extra time on tasks complementary to deep understanding (n). *Prediction 3* says that d does not necessarily increase for teachers as a result of voluntary adoption of off-the-shelf lessons. However, if $\underline{d}$ is high enough, we expect more teachers to have their d increased. We were not able to derive a general proof of this fact; however, see Section A2.3. for its illustration under a specific utility function. *Prediction 4* and *Prediction 5* follow from the fact that the difference in utility from using and not using off-the-shelf lessons increases with teacher quality. *Prediction 4* will be an intensive margin consequence of this fact (i.e. gains from adopting off-the-shelf lessons decrease in teacher quality), while *Prediction 5* will represent an extensive margin (i.e. that the probability of adopting off-the-shelf lessons decreases in teacher quality).

The rest of the section is aimed at proving that *Predictions 4-5* hold in a general case. Specifically, *Claim 1* proves that the difference in utility from adopting and not adopting off-the-shelf lessons increases with teacher quality under assumptions (A'1)-(A'3), using parameter p_d to approximate for overall teacher quality. *Claim 2* shows that parameter p_d is indeed an adequate proxy for overall teacher quality by showing that teachers with lower p_d achieve higher test scores for their students, on average.

Claim 1. Consider a utility function of a general form, $U(n,d)$. Moreover, assume: (A'1) $U(n,d)$ is continuous and differentiable, (A'2) $U'_n > 0$ and $U'_d > 0$ for all (n,d), (A'3) $U(n,d)$ satisfies Inada conditions. Then $U(T/p_n, \underline{d}) - U(n^*,d^*)$ is strictly increasing in p_d. In other words, *(potential) teacher gains from off-the-shelf lessons are strictly decreasing in teacher ability to develop deep understanding.*

Proof. $U(n^*,d^*)$ is equivalent to an indirect utility function, $V(p_n,p_d,T)$. Under standard continuity and monotonicity assumptions, $V(p_n,p_d,T)$ is non-increasing in p_d.[23] Moreover, as

[23] See MWG, Proposition 3.D.3(ii).

we restrict ourselves to interior bundles, we get strict monotonicity, i.e. $V(p_n, p_d, T)$ is strictly decreasing in p_d.[24] Hence, $U(n^*, d^*)$ is strictly decreasing in p_d. Since $U(T/p_n, \underline{d})$ is constant with respect to p_d, the difference between adoption and non-adoption $U(T/p_n, \underline{d}) - U(n^*, d^*)$ is strictly increasing in p_d. ■

Claim 2. Consider a universe of teachers with identical utility functions who differ in terms of p_n and p_d. As teachers teach in identical classrooms, indirect utility function $V(p_n, p_d, T)$ has a convenient interpretation as the overall teacher quality. Assume (A'1)-(A'3) and (A4)-(A6). Then, if $V_{12}(p_n, p_d, T) > 0$ (that is, when there is skill complementarity between imparting knowledge and developing deep understanding),[25] we can prove that:

(i) On average, teachers with lower p_d are able to achieve greater test scores for their students, i.e.:

$$\frac{\partial}{\partial p_d}\mathbb{E}\left[V(p_n, p_d, T)|p_d\right] < 0$$

(ii) There is a positive correlation between overall teacher quality as measured by $V(p_n, p_d, T)$ and the ability to develop deep understanding p_d, i.e.:

$$cov\left[V(p_n, p_d, T), p_d\right] < 0$$

Proof.

(i)

$$\begin{aligned}
\mathbb{E}\left[V(p_n, p_d, T)|p_d\right] &= \int_0^\infty V(p_n, p_d, T) f_{P_n|P_d}(p_n|p_d) dp_n = \\
&= \int_0^\infty V(p_n, p_d, T) \frac{\partial}{\partial p_n}\left[F_{P_n|P_d}(p_n|p_d)\right] dp_n = \\
&= -\int_0^\infty F_{P_n|P_d}(p_n|p_d) V_1(p_n, p_d, T) dp_n + V(p_n, p_d, T) F_{P_n|P_d}(p_n|p_d)\Big|_0^\infty = \\
&= \int_0^\infty F_{P_n|P_d}(p_n|p_d)\left[-V_1(p_n, p_d, T)\right] dp_n
\end{aligned}$$

[24] More formally, envelope theorem implies that:

$$\frac{\partial V^*}{\partial p_d} = \frac{\partial L^*}{\partial p_d} = -\lambda d^* \ ; \ \frac{\partial V^*}{\partial T} = \frac{\partial L^*}{\partial T} = \lambda$$

Assumption (3) tells us that $d^* > 0$. Assumption (2) guarantees that we can always find a way in which an increase of T will result in a strict increase of utility, i.e. $\partial V^*/\partial T > 0$ and, as a result, $\lambda > 0$. Therefore, under our assumptions, $V(p_n, p_d, T)$ is strictly decreasing with p_d.

[25] Whenever skill of developing understanding is improved ($p_d \downarrow$), equilibrium returns from improving other skills go up as well ($V_{p_n} \downarrow$, i.e. $p_n \downarrow$ leads to a faster increase in indirect utility). In addition, from Roy identity (ignoring income effects) one can get that $V_{12} = -\partial n/\partial p_d > 0$, implying n and d are complements in consumption. Both of those interpretations are consistent with our intuition of the model, where n and d are complements by their nature. Moreover, this condition is satisfied in the special case considered in Section A2.

Here, integration by parts was used to proceed from the second to the third line, while assumption (A4) was invoked to eliminate the residual term in line three. Now, using a first-order stochastic dominance assumption (A6) and a sufficient condition $V_{12}(p_n, p_d, T) > 0$, and taking any $p'_d > p_d$:

$$F_{P_n|P_d}(\cdot|p_d) > F_{P_n|P_d}(\cdot|p'_d)$$

$$F_{P_n|P_d}(\cdot|p_d)\left[-V_1(p_n, p_d, T)\right] > F_{P_n|P_d}(\cdot|p'_d)\left[-V_1(p_n, p'_d, T)\right]$$

$$\int_0^\infty F_{P_n|P_d}(\cdot|p_d)\left[-V_1(p_n, p_d, T)\right] dp_n > \int_0^\infty F_{P_n|P_d}(\cdot|p'_d)\left[-V_1(p_n, p'_d, T)\right] dp_n$$

Thus, we were able to prove that conditional expectation $\mathbb{E}\left[V(p_n, p_d, T)|p_d\right]$ is decreasing in p_d.

(ii) To construct a proof of this fact, let us first prove the following lemma:

Lemma. $cov(p_d, f(p_d)) = 1/2E[p_d - p'_d][f(p_d) - f(p'_d)]$ for any function $f(p_d)$, where p_d and p'_d are i.i.d. random variables.

Proof.

$$\begin{aligned}\frac{1}{2}\mathbb{E}[p_d - p'_d][f(p_d) - f(p'_d)] &= \frac{1}{2}\left[\mathbb{E}p_d f(p_d) + \mathbb{E}p'_d f(p'_d) - \mathbb{E}p'_d f(p_d) - \mathbb{E}p_d f(p'_d)\right] = \\ &= \frac{1}{2}\left[2\mathbb{E}p_d f(p_d) - 2\mathbb{E}p_d \mathbb{E}f(p_d)\right] = \mathbb{E}p_d f(p_d) - \mathbb{E}p_d \mathbb{E}f(p_d) = \\ &= cov\left(p_d, f(p_d)\right) \quad \square\end{aligned}$$

Given the lemma shown above, one can show that any decreasing function of p_d has a negative covariance with p_d. That is, if for any $p_d > p'_d$ it follows that $f(p_d) < f(p'_d)$, the lemma claims that $cov(p_d, f(p_d)) < 0$. In our context, this means that:

$$\frac{\partial}{\partial p_d}\mathbb{E}\left[V(p_n, p_d, T)|p_d\right] < 0 \implies cov\left[\mathbb{E}\left[V(p_n, p_d, T)|p_d\right], p_d\right] = cov[V(p_n, p_d, T), p_d] < 0 \ \blacksquare$$

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	I usually look forward to this class	I work hard to do my best in this class	I set aside time to do my homework and study	My math teacher encourages students to share their ideas about things we study in class	My math teacher is willing to give extra help on schoolwork if I need it
(0.544)	(0.637)	(0.237)	(0.365)	(0.593)	(0.576)
In math how often do you apply math situations in life outside of school	Sometimes I get so interested in my work I don't want to stop	Lower bound hours per week studying/working on math outside class	I try to do well on my schoolwork even when it isn't interesting to me	My math teacher encourages us to consider different solutions or points of view	My math teacher notices if I have trouble learning something
(0.589)	(0.609)	(0.237)	(0.412)	(0.639)	(0.576)
In math how often do your assignments seem connected to the real world	The topics are interesting/challenging		I finish whatever I begin. Like you?	My math teacher wants us to become better thinkers, not just memorize things	
(0.622)	(0.546)		(0.637)	(0.558)	
Do you think math can help you understand questions or problems that pop up in your life?	Times per week you talk with your parents or friends about what you learn in math class		I am a hard worker. Like you?	In math how often do you talk about different solutions or points of view	
(0.505)	(0.339)		(0.707)	(0.475)	
	Number of students in math class who feel it is important to pay attention in class		I don't give up easily. Like you?	My math teacher explains things in a different way if I don't understand something in class	
	(0.286)		(0.62)	(0.557)	

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 on Lesson Use - Lower Bound.

In Section VI.1 we examined whether treatment assignment indeed affected usage of Mathalicious lessons. However, our usage measures are based on self-reported data from teacher surveys and, thus, may suffer from the problem of selective reporting (e.g. those who answered the survey were also those who actually responded to treatment). To show that our treatment would still have an effect even if such selection problem were severe, we replicate our results presented in Panel B Table 2 for the extremely conservative case in which all non-responses are interpreted as zero lessons read or taught. The results of such exercise are presented in Table D1 below. Although coefficients are now substantially smaller than those in Table 2, we can still reject the hypothesis that our treatment did not stimulate teachers to adopt the lessons. Even in the worst case scenario, being given a license to Mathalicious lessons and the corresponding supports (i.e. Full Treatment) encouraged teachers, on average, to read 1.8 lessons and teach 0.45 lessons. This represents looking at lessons that relate to about one third of a years' worth of material and teaching lessons that relate to about 9 percent of a years' worth of material.

Table D1. Lower Bound Estimates on Lesson Use:
Missing Outcome Data Replaced by Zeros.

	Lesson Looked	Lesson Looked	Lessons Taught	Lessons Taught
	(1)	(2)	(3)	(4)
License Only	1.101**	1.158***	0.292	0.300
	[0.430]	[0.416]	[0.259]	[0.236]
Full Treatment	1.776***	1.846***	0.468**	0.456**
	[0.431]	[0.445]	[0.217]	[0.218]
District FE x Requested	Y	Y	Y	Y
All controls	N	Y	N	Y
Observations	363	363	363	363

*** - significance at less than 1%; ** - significance at 5%; * - significance at 10%.
Robust standard errors are reported in square brackets. Missing values of lesson use are replaced by zeros. All specifications include controls for district fixed effects interacted with the requested indicator. 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.

Appendix E. Early Survey Results: Two Months After Treatment Assignment

Table E1. Student Survey Early in the Experiment (Chesterfield and Hanover Only).

		Standardized Factors					
	Share of Missing Surveys	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	0.011	0.004	-0.005	-0.009	0.018	0.058	0.029
	[0.073]	[0.074]	[0.071]	[0.051]	[0.060]	[0.079]	[0.088]
Full Treatment	-0.058	0.114	0.058	0.030	-0.018	0.147*	0.115
	[0.064]	[0.070]	[0.066]	[0.049]	[0.051]	[0.085]	[0.075]
District FE x Requested	N	N	N	N	N	N	N
All controls	N	N	N	N	N	N	N
Observations	15,844	7,755	7,704	7,780	7,668	7,722	7,987
Panel B. With All Controls.							
	(8)	(9)	(10)	(11)	(12)	(13)	(14)
License Only	-0.046	-0.039	-0.021	0.036	0.076*	0.027	0.000
	[0.062]	[0.065]	[0.064]	[0.044]	[0.042]	[0.071]	[0.073]
Full Treatment	-0.037	0.080	0.048	-0.003	-0.032	0.090	0.053
	[0.064]	[0.078]	[0.079]	[0.053]	[0.044]	[0.106]	[0.088]
District FE x Requested	Y	Y	Y	Y	Y	Y	Y
All controls	Y	Y	Y	Y	Y	Y	Y
Observations	15,844	7,755	7,704	7,780	7,668	7,722	7,987

*** - significance at less than 1%; ** - significance at 5%, * - significance at 10%.
Standard errors clustered at the teacher level are reported in square brackets. The survey was conducted two months after the experiment started, i.e. after the treatment conditions were assigned and the treatment was distributed. Each outcome, except for the share of missing 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 indicators. The specifications in Panel B add controls for district fixed effects 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 teacher level where each teacher is weighted by the total number of her students. The share of missing 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. When the number of students with completed surveys exceeded the number of students with complete data, negative shares of missing values were replaced with zeros.

Appendix F. 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 F1 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 F1. Teacher Post-Treatment Survey Analysis.

		Teaching practices				Teacher attitude			Student attitude	
	Missing survey	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	0.070	0.001	-0.017	0.142	0.030	-0.117	0.031	0.080	0.145	0.173
	[0.064]	[0.091]	[0.225]	[0.189]	[0.185]	[0.157]	[0.183]	[0.211]	[0.184]	[0.173]
Full Treatment	0.101	0.123	0.015	0.008	-0.086	-0.010	0.042	-0.164	0.129	-0.026
	[0.073]	[0.090]	[0.249]	[0.183]	[0.199]	[0.172]	[0.208]	[0.201]	[0.206]	[0.198]
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

*** - 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 F2. All specifications include controls for district fixed effects interacted with the requested indicator. 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 F2. 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
How often do you ask your students to:	*How often do students do the following?*				*How many of your students do the following?*	
… 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 G. Effect Heterogeneity by Teacher Quality.

One of the key predictions of the model is that the gains from off-the-shelf lessons increase with overall teacher quality. Although we cannot measure teacher quality directly, we can still test this prediction using average students' test scores (at the teacher level) as an overall quality measure.[26] As a start, we use the teacher value-added model as presented in Jackson et al. (2014).[27] 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 τth percentile of the teacher quality distribution.

The standard teacher effects model states that student test scores are determined by the following equation:

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

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's effect was affected by the treatment. That is, we posit that:

$$\mu_t = \beta_T T_t + \nu_t$$

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

$$\bar{Y}_t = \beta_T T_t + \bar{X}_t\beta_{-T} + \nu_t + \theta_c + \bar{\varepsilon}_t \tag{5}$$

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

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

[26] Claim 2 in Appendix B proves that the optimal scores level teachers can achieve is negatively related with p_d, a more direct measure of teacher skill.

[27] We suppress the time subscript, as there is no time dimension in our application.

Now, assuming $Q_\tau(\nu_t(\tau)+\theta_c(\tau)+\bar{\varepsilon}_t(\tau)|T,\bar{X})=0$ for each quantile τ,[28] the quantile regression coefficient $\hat{\beta}_T(\tau)$ is a consistent estimate of $\beta_T(\tau)$ in the model 5. 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.

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

Appendix H. Test Score Regressions - Teacher Level.

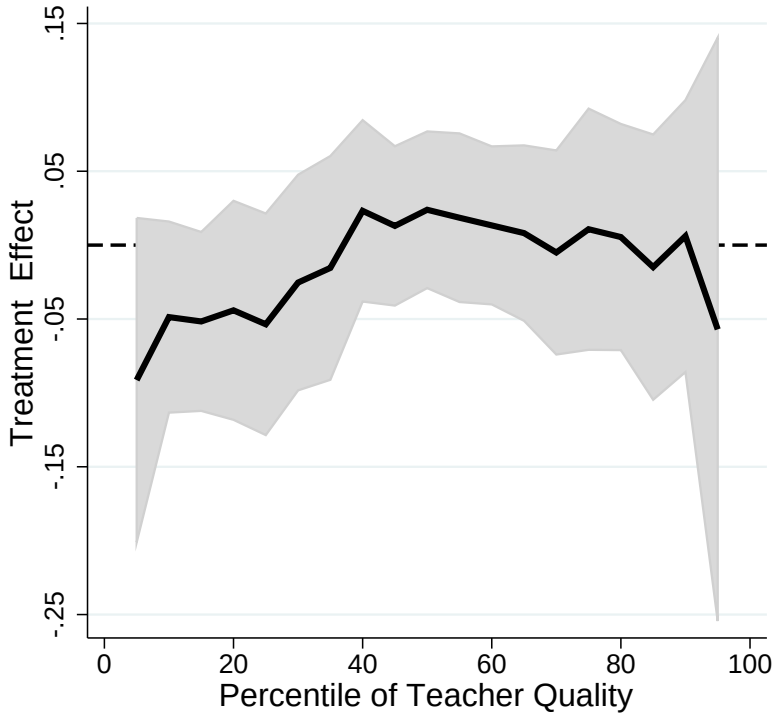
Table H1. Effect on Student Math Scores, Aggregated to the Teacher Level.

	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	1.423	4.119*	0.004	0.047	-0.006	0.004
	[2.511]	[2.129]	[0.039]	[0.033]	[0.042]	[0.039]
Full Treatment	9.144***	8.104***	0.100**	0.091**	0.033	0.016
	[3.119]	[2.704]	[0.045]	[0.038]	[0.041]	[0.037]
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

*** - significance at less than 1%; ** - significance at 5%; * - significance at 10%.
Robust standard errors are reported in square brackets. All specifications include controls for district fixed effects interacted with the requested indicator. Other controls include average lagged test scores, 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. Quantile Regression: English Test Scores.

Figure I1. 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 G.

Appendix J. Sample Mathalicious Lesson #1.

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

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	• 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	*7.88 calories per minute*	*10.08 calories per minute*	*11.21 calories per minute*	*15.75 calories per minute*
Soccer 0.076 cal/lb	*9.50 calories per minute*	*12.16 calories per minute*	*13.53 calories per minute*	*19.00 calories per minute*
Walking 0.019 cal/lb	*2.38 calories per minute*	*3.04 calories per minute*	*3.38 calories per minute*	*4.75 calories per minute*

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

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

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 K. Sample Mathalicious Lesson #2.

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

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	• 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 measureable?*

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*

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 ×2×2×2×2×2? ($\times 2^5$)*
- *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.)*