

# Can Online Off-The-Shelf Lessons Improve Student Outcomes? Evidence from A Field Experiment\*

C. Kirabo Jackson  
Northwestern University and NBER

Alexey Makarin  
Northwestern University

August 20, 2017

## Abstract

There has been a proliferation of websites that warehouse instructional materials designed to be taught by teachers in a traditional classroom. While many teachers now use these online materials, instructional materials can vary considerable in quality and the potential benefits of this innovation are unknown. We analyze an experiment in which high-quality online “off-the-shelf” lessons were identified, and then middle-school math teachers were randomly given access to these lessons. Only providing teachers with online access to the lessons increased students’ math achievement by 0.06 of a standard deviation, but providing teachers with on-line access to the lessons along with supports to promote their use increased students’ math achievement by 0.09 of a standard deviation. Benefits were much larger for weaker teachers, suggesting that weaker teachers compensated for skill deficiencies by substituting the lessons for their own efforts. The intervention is more scalable and cost effective than most policies aimed at improving teacher quality, suggesting that if search costs can be overcome, there is a real benefit to making high-quality instructional materials available to teachers on the Internet.

---

\*Jackson email: [kirabo-jackson@northwestern.edu](mailto:kirabo-jackson@northwestern.edu); Makarin email: [alexey.makarin@u.northwestern.edu](mailto:alexey.makarin@u.northwestern.edu). A previous version of this paper was circulated under the title “Simplifying Teaching: A Field Experiment with Online ‘Off-the-Shelf’ Lessons” as NBER Working Paper No. 22398. This paper was made possible by a grant from the Carnegie Corporation of New York through 100Kin10. We’re extremely grateful to Ginny Stuckey and Kate Novak at Mathalicious, and Sarah Emmons of the University of Chicago Education Lab, and Tracy Dell’Angela at the University of Chicago Urban Education Institute. We also thank math coordinators and the data management persons in Hanover, Henrico, and Chesterfield school districts. We thank Amy Wagner, Jenni Heissel, Hao Hu, and Mathew Steinberg for excellent research assistance. This paper benefited from comments from Ivan Canay, Jon Guryan, Eric Mbakop, Irma Perez-Johnson, Sergey V. Popov, Egor Starkov, and participants of APPAM 2015. The statements made and views expressed are solely the responsibility of the authors.

# I Introduction

Teachers have sizable effects on student test scores (Kane and Staiger 2008; Rivkin, Hanushek and Kain, 2005) and longer-run outcomes (Chetty, Friedman and Rockoff, 2014; Jackson, 2017). Yet, relatively little is known about how to improve teacher quality (Jackson, Rockoff and Staiger, 2014). Teaching is a complex job that involves multiple tasks (Holmstrom and Milgrom, 1991) such as designing lessons, delivering them, managing the classroom, etc. However, much research on teacher effectiveness has focused on how teachers deliver lessons (e.g. Pianta, 2011; Taylor and Tyler, 2012; Araujo et al., 2016) and stayed largely silent on the potentially important task of improving the lessons that teachers deliver. To help fill this space, in this paper we examine an intervention aimed at increasing the quality of the lessons used by teachers in the classroom. Specifically, we study the student achievement effects of providing teachers with free access to high-quality, off-the-shelf lessons on the Internet.

There has been a recent proliferation of lesson plans and instructional materials designed to be taught by teachers in a traditional classroom that can be accessed online. One early site called **Teachers Pay Teachers** was launched in 2006 and allowed teachers to sell their lesson plans and instructional materials to other teachers. As of 2016, this site is estimated to have an active membership of approximately 4 million (this is more than all primary and secondary teachers in the United States). Other major players in this product space provide mostly free and openly licensed instructional materials such as **LearnZillion**, **Pinterest**, and **Amazon Inspire** (Madda, 2016). There is considerable demand among teachers for these online resources. Opfer, Kaufman and Thompson (2016) and Purcell et al. (2013) found that over 90 percent of middle and high school teachers use the Internet to source instructional materials when planning lessons. Though this innovation may have little visible effect on how teachers deliver lessons, it has altered how teachers plan and create their lesson content.<sup>1</sup>

Lesson sharing websites create a positive information externality such that all teachers, irrespective of geography or experience, may have access to high-quality lesson plans. These lesson plans may be designed by expert educators and may embody years of teaching knowledge and skills that most individual teachers do not possess themselves. In principle, through these websites, the creation of one high-quality lesson has the potential to improve the outcomes of millions of students. However, many teachers may not use high quality lessons because (a) there are search costs associated with identifying high quality lessons and (b) there may be direct user fee costs for high quality lessons. Accordingly, removing these costs by identifying high quality lessons and providing them to teachers free of charge could yield considerable achievement gains for students. Despite these potential benefits, the extent to which providing teachers access to free high-quality

---

<sup>1</sup>Indeed, this change in how teachers create instructional materials has led to recent popular press headlines such as “*How the Internet is complicating the art of teaching*” and “*How did we teach before the Internet?*”

online instructional materials improves their student’s performance is unknown. We present the first rigorous examination of this question. Specifically, we implemented a randomized field experiment in which middle-school math teachers in three school districts were randomly provided access to high-quality off-the-shelf lessons, and we examine the effects on their students’ subsequent academic achievement.

At the heart of our intervention are high-quality, off-the-shelf lessons. These lessons differ from those in traditional math classrooms. In the typical US math class, teachers present definitions and show students procedures for solving specific problems. Students are then expected to memorize the definitions and practice the procedures (Stigler et al., 1999). In contrast, informed by education theory on inquiry-based instruction (Dostál, 2015), embedded learning (Lave and Wenger, 1991; Brown, Collins and Duguid, 1989), classroom discussion (Bonwell and Eison, 1991), and scaffolding (Sawyer, 2005), the off-the-shelf lessons used in this study were designed to promote deep understanding, improve student engagement, and promote retention of knowledge.<sup>2</sup> These lessons were designed to be taught over two to five class periods, and each lesson lays the foundation for between 3 and 8 weeks of course material. Under our experiment, teachers were randomly assigned to one of three treatment conditions. In the “license only” condition, teachers were informed that these lessons were high quality and given free access to these online lessons. While these lessons were provided online, they are designed to be taught by teachers in a traditional classroom setting. To promote lesson adoption, some teachers were randomly assigned to the “full treatment” condition. In the full treatment, teachers were granted free access to the online lessons, received email reminders to use them, and were invited to an online social media group focused on lesson implementation. Finally, teachers randomly assigned to the control condition continued “business-as-usual.”

Because the treatments were assigned randomly, we identify causal effects using multiple regression. Students of teachers in the license only group and the full treatment group experienced a  $0.06\sigma$  and  $0.09\sigma$  test score increase relative to those in the control condition, respectively. The full treatment has a similarly sized effect as that of moving from an average teacher to one at the 80th percentile of quality, or reducing class size by 15 percent.<sup>3</sup> Because the lessons and supports were provided online, the marginal cost of this intervention is low. Moreover, the intervention can be deployed to teachers in remote areas where coaching and training personnel may be scarce, and there is no limit to how many teachers can benefit from it. Back-of-the-envelope calculations suggest a benefit-cost ratio above 900, and an internal rate of return greater than that of the Perry Pre-School

---

<sup>2</sup>While there is much observational evidence that teachers who engage in these best practices have better student outcomes (e.g., Pianta, 2011, Mihaly et al., 2013, Araujo et al., 2016), there is very little experimental evidence on how promoting best practices among existing teachers impacts achievement tests.

<sup>3</sup>This is based on estimates from a variety of studies on teacher quality summarized in Jackson, Rockoff and Staiger (2014). Our evidence on the effects of class size comes from Krueger (1999) and Chetty et al. (2011).

Program (Heckman and Masterov, 2007), Head Start (Deming, 2009), class size reduction (Chetty et al., 2011) or increases in per-pupil school spending (Jackson, Johnson and Persico, 2016).

We demonstrate that, even with a single year of achievement data, one can test for heterogeneous treatment effects by teacher/classroom quality using conditional quantile regression models (Koenker and Bassett, 1978). We follow this methodological innovation. Even though information technology is complementary to worker skill in many settings (e.g. Katz, 1999, Akerman, Gaarder and Mogstad, 2015), the benefits of online lesson use are the largest for the least effective teachers, and decrease with effectiveness (as measured by teacher/classroom value added). We theorize that this is due largely to lesson quality improvements being largest for weaker teachers. We also find suggestive evidence that lesson provision had larger effects for first-year teachers, implying that the off-the-shelf lessons may have provided some time savings for these teachers.

Looking to mechanisms, teachers who were only granted free access to the lessons looked at 1.59 more lessons, and taught 0.65 more lessons than control teachers, while, on average, fully-treated teachers (access plus supports) looked at 4.4 more lessons, and taught 1.9 more lessons than control teachers. The level of lesson use in the full treatment relates to about one-third of a year's worth of material. Consistent with improved lesson quality and the aims of the intervention, treated students are more likely to report that teachers emphasize deep learning, and to report feeling that math has real life applications. The meaningful test score gains in the license only condition suggest that the improved outcomes in the full treatment condition are not driven by the additional supports but by the increased lesson use. To provide evidence of this, we show that (a) the treatment arms with the most lesson use also had the largest test score improvements, (b) on average, the test score effects increase with lesson use, and (c) conditional on lesson use, receiving the extra supports was unrelated to test scores. Given the large documented benefits to lesson use, we explore why take-up was not more robust. We speculate that teachers may have some behavioral biases such that the regular reminders and additional supports to use the lessons may have been important drivers of success in the full treatment condition. Overall, our findings suggest that if districts can identify high-quality lessons, make them freely available to teachers, and promote their use, the benefits could be as large, if not larger, than the positive effects we document here. 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, 2014).

Our findings contribute to a few related literatures in different ways. First, the approach to improving instructional quality we study is a form of division of labor; classroom teachers focus on some tasks, while creating instructional content is (partially) performed by experts with particular skills in that domain. This paper adds to a nascent literature exploring the potential productivity benefits of teacher specialization in schools (e.g. Fryer, 2016; Jacob and Rockoff, 2012). Second,

while certain kinds of instructional materials are *associated* with better student outcomes (Bhatt and Koedel, 2012; Chingos and Whitehurst, 2012; Kane et al., 2016; Koedel et al., 2017), we show that exogenously introducing high-quality instructional materials into existing classrooms has a sizable causal effect on student outcomes. Third, 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, Markman and Rouse, 2009, Taylor, 2015). In contrast, we add to this literature by examining whether technology can help teachers enhance their traditional teaching practices through the dissemination of teaching knowledge.<sup>4</sup> Finally, this study relates to the personnel economics and management literatures by presenting a context in which one can improve worker productivity by simplifying the jobs workers perform (Bloom et al., 2012; Jackson and Schneider, 2015).

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

## II The Intervention

### II.1 The Off-the-Shelf Lessons

The job simplifying technology at the heart of the intervention is off-the-shelf lessons. These lessons are from the "Mathalicious" curriculum. 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.<sup>5</sup> Lessons range from the simple to the more complex.

The lesson titled "*Xbox Xponential*" (see Appendix P) is a typical 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

---

<sup>4</sup>In related work, Comi et al. (2016) find that effectiveness of technology at school depends on teachers' ability to incorporate it into their teaching practices.

<sup>5</sup>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 involve showing students how to perform certain mathematical procedures, such as how to do long division. Teaching concepts would involve simple word problems that make the mathematical concept clear. Teaching applications are where students use math to explore multiple facets of some real-world question. In teaching applications, students would develop their own models (Lesh and Doerr, 2003), test and refine their thinking, and talk about it with each other.

the evolution of football video games over time. Students are asked to “*sketch a rough graph of how football games have changed over time*” and then asked to describe what they are measuring (realism, speed, complexity, etc). They are then guided by the teacher to realize that “*while a subjective element like ‘realism’ is difficult to quantify, it is possible to measure speed (in MHz) of a console’s processor.*” In the second part of the lesson, students are introduced to Moore’s 1965 prediction that computer processor speeds would double every two years. They are then provided with data on the processor speeds of game consoles over time (starting with the Atari 2600 in 1977 through to the XBOX 360 in 2005). Students are instructed to explain Moore’s law in real world terms and to use this law to predict the console speeds during different years. In the third part of the lesson, students are asked to sketch graphs of how game consoles speeds have actually evolved over time, come up with mathematical representations of the patterns in the data, and compare the predictions from Moore’s Law to the actual evolution of processor speeds over time. During this lesson, students gain an intuitive understanding of measurement, exponential functions, extrapolation, and regression through a topic that is very familiar to them - video games.<sup>6</sup>

Teachers during these lessons do not serve as instructors to present facts (as 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, Collins and Duguid, 1989; Dostál, 2015). However, because the existing empirical studies on this topic are observational, this paper presents some of the first experimental evidence of a causal link between inquiry-based situated math instruction and student achievement outcomes.

Because the Mathalicious lessons are memorable and develop mathematical intuition through experience, they serve as “anchor lessons” that teachers can build upon during the year when introducing formal math ideas. For example, after teaching “*Xbox Xponential*”, teachers who are introducing the idea of an exponential function formally would say, “*Remember how we figured out the speed of videogame consoles over time? This was an exponential function!*” and students would use the intuition built up during the anchor lesson to help them understand the more formal lesson about exponential functions (which may occur days or weeks later). Each of these “anchor lessons” touches on several topics and serves as an intuitive anchor for as much as two months of math classes. When the Mathalicious curriculum is purchased by a school district, each Mathalicious lesson lists the grade and specific topics covered in that lesson, and proposed dates when each lesson might be taught. Full fidelity with the curriculum entailed teaching 5 to 7 lessons each year.

One treatment arm of the intervention involved an additional component to facilitate lesson use, called Project Groundswell. Project Groundswell allowed teachers to interact with other teach-

---

<sup>6</sup>See the lesson titled “*New-Tritional Info*” in [Appendix Q](#) for a less complex lesson.

ers using Mathalicious lessons online through "Edmodo" (a social networking platform designed to facilitate collaboration among teachers, parents, and students). Project Groundswell provided a private online space to have asynchronous discussions with Mathalicious developers and other Mathalicious teachers concerning lesson implementation. Project Groundswell also included webinars (about 7 per year) created by Mathalicious developers. During these webinars, Mathalicious personnel would walk teachers through the narrative flow of a lesson, highlight key understandings that should result from each portion of the lesson, anticipate student responses and misconceptions, and model helpful language to discuss the math concepts at the heart of the lesson.

## II.2 The Experiment

During the Spring of 2012, Mathalicious and the research team decided to conduct an evaluation of the Mathalicious curriculum and Project Groundswell. Mathematics coordinators in three Virginia districts sought to purchase Mathalicious licenses for some of their teachers. These districts were offered additional licenses free of charge and free access to Project Groundswell in exchange for participation in the evaluation. Participation in the study entailed sharing the public school email addresses of eligible participant teachers, allowing the research team to assign teachers to different treatment conditions (described below), and providing administrative data to the research team. No school leaders were involved in the running of the intervention or had access to any non-administrative data created by the research team. All three Virginia districts agreed to participate: Chesterfield, Henrico, and Hanover. Across all grade levels, 59,186 students were enrolled in 62 Chesterfield public schools, 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 8 math teachers in these districts were part of the study. Teachers were placed into one of the three conditions described below:

*Treatment Condition 1: Full Treatment (Mathalicious subscription and Project Groundswell).* Full treatment teachers were granted access to both the Mathalicious lessons and Project Groundswell. They were invited to an in-person kickoff event where Mathalicious personnel reviewed the online 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, and 8). 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.<sup>7</sup> Importantly, participation in all components of the treatment was entirely voluntary.

---

<sup>7</sup>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).

*Treatment Condition 2: License Only Treatment (Mathalicious subscription only).* Teachers who were assigned to the license only treatment were only provided with a subscription to the Mathalicious curriculum. These teachers received the same basic technical supports available to all Mathalicious subscribers. However, they were not invited to participate in Project Groundswell (i.e. they were not invited to join an Edmodo group and did not receive email reminders). In sum, at the start of the school year, these teachers were provided access to the lessons, given their login information, and left to their own devices.

*Treatment Condition 3: Control Condition (business-as-usual).* Teachers who were randomly assigned to the control condition continued “business-as-usual.” They were not offered the Mathalicious lessons, nor were they invited to participate in Project Groundswell. Even though control teachers were not *prevented* from using the Mathalicious lessons, the overwhelming majority of control teachers continued to use the non-Mathalicious curriculum of their choice. While we do not observe how control teachers planned their lessons, existing studies provide some guidance. According to [Opfer, Kaufman and Thompson \(2016\)](#), over ninety percent of teachers use district developed materials. Virginia publishes a Curriculum Framework and simple lesson guides for the topics that teachers are expected to cover in each grade.<sup>8</sup> The Virginia Curriculum Framework lesson plans are not as inquiry-ordinated or project-based as the Mathalicious lessons. It is reasonable to expect that these were used heavily by teachers in the control condition. [Opfer, Kaufman and Thompson \(2016\)](#) also found that most teachers use district materials in conjunction with material developed on their own. Teachers’ own efforts likely included a considerable amount of Internet-sourced content ([Opfer, Kaufman and Thompson, 2016](#); [Purcell et al., 2013](#)). If there are non-trivial search costs associated with identifying high-quality lessons, the lessons that teachers would have used in the control condition may not have been as high quality as the Mathalicious lessons. We present evidence on this in Section VII.2. Because these school districts had not been offered Mathalicious lessons before the intervention, control teachers would not have been familiar with the curriculum and would not have been using it. Insofar as any spillovers did occur (through treatment teachers sharing materials with colleagues in the control group), they would attenuate our estimated effects toward zero.<sup>9</sup>

*Assignment of Teachers to Treatment Conditions.* Prior to conducting the study, the research team and Mathalicious decided on a predetermined number of licenses that could be allocated to teachers in each district. In summer 2013 (the summer before the intervention), the research team received a list of all math teachers eligible for this study from each district. To facilitate district participation in the study, two of the districts were allowed to pre-select certain regular classroom teachers that they wished to receive access to the Mathalicious licenses (i.e., receive either Treat-

---

<sup>8</sup><http://bit.ly/2u97XaZ>

<sup>9</sup>We show in [Appendix F](#) that the impacts of any spillovers on our estimates, if they exist, are negligible.

ment Condition 1 or Treatment Condition 2). We refer to these teachers as “requested” teachers. All requested teachers were identified and removed from the control condition. All of the remaining unrequested licenses in each district were allocated randomly to the remaining teachers.<sup>10</sup> As such, among those that were not requested teachers, whether a teacher received a license was random. In a second stage, among all teachers who had licenses (i.e. both those who were pre-selected and those who received the license by random chance) we randomly assigned half to receive the full treatment (i.e. Treatment Condition 1). Among non-requested teachers, treatment status is random conditional on district, and among requested teachers, assignment to the full treatment is random conditional on district. As such, treatment assignment was random conditional on *both* requested status and district, and the interaction between the two.<sup>11</sup> Accordingly, all models condition on district and “requested” status and their interaction.<sup>12</sup> Moreover, our main results are robust to excluding the requested teachers.<sup>13</sup> The use of randomization ensured that conditional on requested status and district, teachers (and their students) had no control over their treatment condition and therefore reduced the plausibility of alternative explanations for any observed *ex post* differences in outcomes across treatment groups.

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

### III Data

Our data 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 2013-14 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

---

<sup>10</sup>Because the number of unrequested licenses varied across districts, the probability of being assigned to the license condition varied by district. All empirical models include district fixed effects to account for such differences.

<sup>11</sup>Table A1 of Appendix A summarizes teacher participation by district, requested status, and treatment condition.

<sup>12</sup>This set-up is analogous to covariate-adaptive randomization procedures in which randomization occurs within certain strata of baseline covariates. Bugni, Canay and Shaikh (2017) show that in the case of multiple treatments, i.e. our set-up, a regression with strata fixed effects and robust standard errors is also a valid specification.

<sup>13</sup>We present these results later in Section VI. In Appendix B, we present evidence that requested teachers are not that different from the rest of the participants. Furthermore, there is no evidence that the treatment effect on test scores varies by the “requested” status.

teachers. These pre-treatment student and teacher attributes are shown in [Table 1](#).

The key outcome for this study is student math achievement (as measured by test scores). We obtained student results on the math portion of the Virginia Standards of Learning (SoL) assessment for each district for the academic years 2012-13 and 2013-14. These tests comprise the math content that Virginia students were expected to learn 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.<sup>14</sup> 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  $0.04\sigma$ ) but in 2014 are  $0.163\sigma$  higher in the full treatment condition relative to the control condition.<sup>15</sup> The relative improvement in math scores over time is  $0.163-0.04=0.123\sigma$  between the full treatment and the control group. By comparison, the relative improvement in English scores over time (where there should be no effect) between the full treatment and the control group is  $0.003\sigma$ . These simple comparisons telegraph the more precise multiple regression estimates we present in [Section VI](#).

We use data from other sources to measure lesson use and to uncover underlying mechanisms. Each teacher was invited to answer two surveys: 22% and 61% of teachers completed a mid-year and an end-of-year survey, respectively.<sup>16</sup> Using teacher survey data, we observe the self-reported lessons they taught and read. Because these data are from surveys, using them will automatically have zeros for those individuals who do not complete the surveys - leading to an underestimate of the effect of the treatments on lesson use. We describe how we address this problem in [Section VII](#). We supplement these data with the more objective measure of lessons downloaded. Specifically, for each lesson, we record whether it was downloaded for each teacher's account using tracker data from the Mathalicious website. Based on both these data sources, our three measures of Mathalicious lesson use are (a) the number of lessons looked at, (b) the number of lessons taught, and (c) the number of lessons downloaded. For each lesson, we code up a lesson as having been looked at if either the tracker indicated that it was downloaded or if the teacher reported reading or teaching that lesson. The lessons taught measure comes exclusively from survey reports.

To explore mechanisms, surveys were given to students.<sup>17</sup> Survey questions were designed by the research team and Mathalicious to measure changes in factors hypothesized to be affected by the intervention (see [Appendix C](#) for survey items). The student surveys were administered in the

---

<sup>14</sup>In Hanover district, the exam codes were not provided so that the test scores are standardized by grade and year only. In our preferred specification, we control for the interaction between incoming test scores and district indicators.

<sup>15</sup>Students with missing 2013 math scores are given an imputed standardized score of zero. To account for this in regression models we also include an indicator denoting these individuals in all specifications.

<sup>16</sup>20% of teachers completed both surveys and 61% of teachers completed either of them.

<sup>17</sup>We also administered teacher surveys for this study. However, due to high differential attrition rates the results are inconclusive and we do not discuss effects on these data in the main text. Teacher surveys were designed to measure teacher job satisfaction and classroom practices. Results on the teacher surveys are presented in [Appendix I](#).

middle and at the end of the intervention year in two of the districts. The surveys were designed to measure student attitudes toward mathematics and academic engagement. The student survey items are linked to individual teachers, but were anonymous. The survey items are discussed in greater detail in Section VII.

## IV Theoretical Framework

We lay out a theoretical framework to help organize our thinking about the effect of off-the-shelf lessons. Teaching is a multitask job (Holmstrom and Milgrom, 1991) involving complementary tasks: planning lessons and all other teaching activities (lesson delivery, classroom management, etc.). Teachers allocate their time toward lesson planning, other teaching tasks, and leisure. The off-the-shelf lessons (a) guarantee a minimum level of lesson quality, and (b) free up teacher time that would have been spent planning lessons, but (c) require some implementation time cost. Within this framework, teachers (and their students) may benefit from using the lessons in two ways.

The first way a teacher could benefit from the off-the-shelf lessons is through the Lesson Quality Mechanism. Specifically, *all else equal*, if a teacher substitutes the Mathalicious lessons for her own lessons, then lesson quality may improve for those teachers who would have had poor lesson quality if left to their own efforts. The second way to benefit is via the Time Savings Mechanism. Holding lesson quality fixed, if the time saved on lesson planning through using the off-the-shelf lessons is larger than the implementation time costs, adopting teachers will have more time to allocate to *all* tasks (i.e. lesson planning, other teaching tasks, and leisure), some of which may go toward increasing test scores. However, because teachers could use any time savings (or potential benefits to test scores) as a way to increase leisure, in theory, lesson use could *reduce* student achievement.

To shed further light on this we model the teacher's problem (see Appendix E for the full model and formal proofs). We assume that teachers care about the test scores of their students and leisure, and that these two goods are complementary. For analytical tractability, we assume that both teacher utility and test-scores are Cobb-Douglas. This model yields four non-obvious and testable results:

Result 1: *The gains in average test scores from using the off-the-shelf lessons are non-negative.* This is because (a) teachers adopt lessons *iff* the time savings are large enough to allow test scores to weakly increase, and (b) teachers will use some of the time savings to increase tests scores.

Result 2: *The relationship between the benefits of lesson use and teacher quality is ambiguous in sign.* If weaker teachers are more likely to have low lesson quality, the benefits of lesson use may be higher for weaker teachers. However, if lesson implementation costs are higher for weaker teachers, lesson use will have larger benefits for stronger teachers.

Result 3: *The effect of lesson adoption on lesson quality is non-negative.* Under the assumptions, lesson quality is a normal good. As a result, lesson quality will not decrease with lesson use.

Result 4: *The effect of lesson adoption on time spent on other teaching tasks is ambiguous. It may be optimal for adopting teachers to increase or decrease time spent on other teaching tasks depending on the curvature of the test score production function and the quality of the lessons.*<sup>18</sup>

## V Empirical Strategy

We aim to identify the effect of treatment status on various teacher and student outcomes. We compare outcomes across treatment categories using a multiple regression framework. 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,  $\beta_1$  and  $\beta_2$  represent the differences in outcomes between the control and the license only groups, and between the control and the full treatment groups, respectively. The treatment assignment was random within districts and after accounting for whether the teacher was requested for a Mathalicious license. Consequently, following [Bugni, Canay and Shaikh \(2017\)](#), all models include a separate dummy variable for each district to absorb the district effects,  $\alpha_d$ , an indicator variable  $Req_{dt}$  denoting whether teacher  $t$  requested a license in district  $d$ , and the interaction between the two, denoted by the fact that coefficient  $\pi$  varies by district  $d$ . To improve precision,<sup>19</sup> we also include  $X_{dt}$ , a vector of teacher covariates (these include teacher experience, gender, ethnicity, and grade level taught) and student covariates averaged at the teacher level (average incoming student math and English test scores, the proportion of males, and the proportion of black, white, Hispanic, and Asian students).

Our main outcome of interest is student math test scores. For this outcome, we estimate models at the individual student level and employ a standard value added model ([Jackson, Rockoff and Staiger, 2014](#)) that includes individual lagged test scores as a covariate. Specifically, where students are denoted with the subscript  $i$ , in our test score models, we estimate the following regression

---

<sup>18</sup>The online lessons produce a kink in the teacher’s budget constraint. If teachers locate at the kink, time on other tasks will decrease. For teachers who do not locate at the kink, time on other teaching tasks will increase.

<sup>19</sup>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).

equation using OLS:

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

In (2),  $X_{idt}$  includes student race and gender, and classroom averages of all the student-level covariates (including lagged math and English test scores), as well as all of the teacher-level covariates from (1). Standard errors are adjusted for clustering at the teacher level in all student-level models.

## VI Main Results

### VI.1 Effects on Student Achievement in Mathematics

The first result from the theoretical section is that the intervention effect on math scores should be non-negative. To test this, we focus on test scores standardized by exam. However, effects on raw test scores (measured on a 0-600 scale) are also presented. Test scores are analyzed at the individual student level and standard errors are adjusted for clustering at the teacher level in Panel A of [Table 2](#). The results reveal positive effects on math test scores from simply providing licenses, and even larger positive and statistically significant effects for the full treatment. The first model (columns 1 and 3) includes the key conditioning variables (district fixed effects interacted with requested status) and the average lagged math scores in the classroom interacted with the district. In this model (column 3), teachers who only had access to the lessons had test scores that were 5% of a standard deviation higher than those in the control condition (p-value>0.1), and teachers with access to both Mathalicious lessons and extra supports increased their students' test scores by 10.5% of a standard deviation relative to those in the control condition (p-value<0.05). One cannot reject that the full treatment teachers have outcomes different from those in the license only group, but one can reject that they have the same outcomes as teachers in the control group.

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

students' test scores by 5.5% of a standard deviation relative to those in the control condition ( $p$ -value $<0.1$ ), and full treatment condition increased their students' test scores by 9.3% of a standard deviation relative to those in the control condition ( $p$ -value $<0.01$ ). In sum, across all the models, there is a robust positive effect of both the license only treatment and the full treatment (relative to the control condition) on student test scores of roughly 6 and 9 percent of a standard deviation, respectively. Also, across all models, the full treatment is associated with larger and more precisely estimated math test score gains than the license only treatment.

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, or that there was something else that could drive the positive math test score effects. To assuage such concerns, we report a falsification exercise with end-of-year English test scores as the outcome in Columns 7 and 8. Because the Mathalicious website provided lessons only for math curriculum, English test scores 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. This reinforces the notion that the improved math scores are due to increased lesson use and are not driven by student selection, Hawthorne effects, or John Henry effects. As an additional robustness check on the experimental design, we also estimate models without requested teachers and the pattern of results are the same (see Panel B of [Table 2](#)).

Because control teachers were not prevented from using the lessons, we test for possible spillovers on control teachers (see [Appendix F](#)). We do this in two ways. First, we include the fraction of other math teachers at the school in each treatment condition. Second, we include school fixed effects so that we only compare teachers at the same school. In neither the teacher nor the student-level analysis, can one reject that our results are the same as those in [Table 2](#). However, the pattern of the results does indicate that there may have been some positive spillovers to control teachers such that the results we present in [Table 2](#) may slightly understate the true effect.

## **VI.2 Effect Heterogeneity by Teacher Quality**

The second theoretical result is that the treatment effect may be larger or smaller for less effective teachers. Weaker teachers who are relatively ineffective at improving student performance may benefit greatly from the provision of the lessons. However, less effective teachers may not have the requisite skills to properly implement or support the lessons so that they benefit less from lesson use. To test which scenario holds empirically, we see if the marginal effect of the treatment is larger or smaller for teachers lower down in the quality distribution. Following the teacher quality literature, we conceptualize teacher quality as the ability to raise test scores. 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. Because we only have a single year of data, we cannot distinguish between classroom quality and teacher quality *per se*; however, prior research indicates that the two are closely related. Following [Chetty et al. \(2011\)](#), we proxy for teacher quality with classroom quality.

To test for effects by teacher effectiveness, one would typically estimate teacher effectiveness using some pre-experimental data, and then interact the randomized treatment with the teacher’s pre-treatment effectiveness. Because we only have a single year of achievement data for each teacher, we take a different, but closely related, approach. To test for different effects for classrooms at various points in the distribution of classroom quality, we employ conditional quantile regression. Conditional quantile regression models provide marginal effect estimates at particular quantiles of the residual distribution ([Koenker and Bassett, 1978](#)). As we formally show in [Appendix G](#), when average test scores at the teacher level is the dependent variable, the teacher-level residual from (1) is precisely the standard value-added measure of classroom quality. Accordingly, the marginal effect of the treatment at the  $p$ -th percentile from the conditional quantile regression of equation (1) is the marginal effect of the treatment for teachers at the  $p$ -th percentile of effectiveness. To verify this claim computationally, we implement a Monte Carlo simulation (see [Appendix G](#)) and are able to consistently uncover treatment effects at different percentiles of the teacher quality distribution.<sup>20</sup>

To estimate the marginal effect of the full treatment for different percentiles of the classroom quality distribution, we aggregate math test scores to the teacher level and estimate conditional quantile regressions for the 10th through 90th percentiles in intervals of 5 percentile points. We plot the marginal effects of the full treatment against the corresponding quantiles along with the 90 percent confidence interval for each regression estimate in [Figure 1](#). There is a clear declining pattern indicating larger benefits for low-quality classrooms than for high-quality classrooms. To model the non-linear relationship, we fit a piece-wise linear function with a structural break at the 60th percentile. At and below the 60th quantile the slope is 0.0003 and not statistically significant, while above the 60th quantile the slope is -0.00314 (p-value=0.001).<sup>21</sup> In sum, for the bottom 60 percent of teachers, the marginal effect of the full treatment is roughly  $0.11\sigma$ , and the full

---

<sup>20</sup>[Appendix H](#) 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.

<sup>21</sup>Because we have an estimated dependent variable, the standard errors need to be adjusted for heteroskedasticity. As pointed out in [Lewis and Linzer \(2005\)](#), heteroskedasticity correction by Huber-White is sufficient. We follow this approach. As an alternative approach, we follow the adjustment outlined in [Hanushek \(1974\)](#) to account for estimation error in the dependent variable (also used in [Card and Krueger \(1992\)](#) and [Eichholtz, Kok and Quigley \(2010\)](#)). Models with this adjustment yield standard errors which are virtually identical to the Huber-White standard errors. To further assuage any concerns, we ran a Monte Carlo simulation where we assigned random placebo treatments (using the same distribution as the actual treatments), estimated quantile regressions based on the placebo treatment, then estimated the piecewise linear model. Based on 1,000 placebo replications, our actual estimated slope above the 60th percentile (-0.00314) lies below the 5th percentile of the distribution of placebo slopes.

treatment is only ineffective for the most able teachers in the top ten percent of the effectiveness distribution. Given the decline in treatment effectiveness by teacher “quality”, one may worry that the intervention reduced achievement for high-quality classrooms. Such patterns were observed for computer-aided instruction in [Taylor \(2015\)](#). However, even at the 99th percentile of classroom quality, the semi-parametric point estimate is positive (albeit not statistically different from zero).<sup>22</sup> 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.

One may wonder if this pattern is driven by larger effects for less experienced teachers, for whom both the time savings mechanism and the lesson quality mechanisms may be at play. We test this formally by interacting the treatment with teacher experience. [Table I1](#) in [Appendix I](#) presents the results both at the teacher and student levels. Columns (1) and (3) suggest that, on average, there is no linear relationship between the effect of the intervention and teacher experience. However, in a model that interacts the treatment with an indicator for being a first- or a second-year teacher (Columns (2) and (4)), the point estimate on the interaction with the license only treatment is positive and one can reject that it is zero at the 5 percent level. Given that we find positive effects for more than 60 percent of teachers and only 5.5 percent of all teachers in our sample are first- or second-year teachers, this cannot explain the full pattern of results. Moreover, only license-only first- or second-year teachers exhibit such differential response, while fully treated first- or second-year teachers are indistinguishable from their peers. However, these results are broadly consistent with a model in which the benefits of the off-the-shelf lessons are larger for those teachers who are (a) less effective and/or (b) more likely to spend a lot of time planning lessons.

## VII Mechanisms

### VII.1 Effects on Mathalicious lesson use

We now explore the extent to which the test score effects are driven by increased Mathalicious lessons use. We have two sources of data to measure Mathalicious use, both of which are imperfect. First, we rely on self-reported measures of which Mathalicious lessons were taught or read. This information was reported by teachers during the mid-year and end-of-year surveys and may suffer from bias due to survey non-response. Second, we use the data received from Mathalicious site logs on whether a teacher downloaded a certain lesson or not (based on login email). Unfortunately, the download tracker may understate lessons downloaded for two reasons. First, the download tracker was not available for the first month of the experiment. Second, the tracker only tracked

---

<sup>22</sup>To ensure that these patterns are real, as a falsification exercise, we estimate the same quantile regression model for English test scores (see [Figure J1](#) in [Appendix J](#)). 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 at the ten percent level. This provides further evidence that the estimated effects on math scores are causal, and that the pattern of larger treatment effect for the less able teachers is real.

downloads for official public school email address, and there was nothing preventing teachers from using their personal email accounts. With these imperfect sources of information on lesson use, we construct three measures: the number of Mathalicious lessons taught (as reported by the teacher), the number of lessons the teacher looked at (either reported as taught, reported as read, or tracked as downloaded), and the number of lessons tracked as downloaded. We also employ data on webinars attended in real-time. While the webinars were designed to facilitate real-time interaction among teachers and Mathalicious facilitators, they were recorded and made available for asynchronous viewing. As such, this measure may not capture the extent to which teachers *viewed* webinars, and may understate teacher use of these additional supports.

We analyze the effect of the treatment on these measures of use in [Table 3](#). Because our measures of lessons taught and viewed are (partially) obtained from survey data, we only have complete lesson use for the twenty percent of teachers who completed the surveys during both waves. We address this by using multiple imputation ([Rubin, 2004](#); [Schafer, 1997](#)) to impute lesson use for those individuals who did not complete the surveys.<sup>23</sup> Using teachers with complete survey data to impute lesson use for those with missing data may introduce upward bias if teachers who complete surveys tend to have higher levels of use than those who do not. We test for this formally in [Appendix K](#), where we show that conditional on treatment status, survey participation is unrelated to lessons downloaded so that the imputation method is likely valid. For the lessons looked at, we conduct multiple imputation for the survey responses before combining it with the tracker data.

The regression results based on imputed use (for missing data) are presented in Panel A of [Table 3](#). Note that standard errors are corrected for multiple imputation as in [Rubin \(2004\)](#) and all models include the full set of controls. Teachers in the license only condition looked at 1.586 more lessons and taught 0.657 more lessons than teachers in the control condition, while teachers in the full treatment condition looked at 4.4 more lessons and taught 1.925 more lessons than teachers in the control condition. The results are very similar to, but more precise than, those that only use teachers with complete survey data (see Panel A of [Table L1](#) in [Appendix L](#)). To assuage concerns that teachers with and without requested licenses are systematically different in their behavior, Columns 3 and 4 of Panel A of [Table 3](#) show the same models excluding requested teachers. The effects on lessons looked at are similar, and the effects on lessons taught are virtually identical.<sup>24</sup>

---

<sup>23</sup>Within each multiple imputation sample, we impute the missing numbers of lessons looked at and lessons taught using predicted values for other teachers with complete data in the same treatment condition from a Poisson regression (note that these are count data).

<sup>24</sup>As an additional check on our method, we compute lesson use based on the 60 percent of teachers that completed either the mid-year survey or the end of year survey. Because teachers who do not complete one of the surveys are automatically assigned zero use for *that survey wave*, these results are biased toward zero. As such, these estimates are likely to be lower in magnitude than the real effects. Panel B of [Table L1](#) in [Appendix L](#) presents the estimated effects among the 60 percent of teachers with at least partially complete survey data (i.e. survey data in at least one of the two waves). While the point estimates are smaller than results using the 20 percent of teachers with full data (as expected), all the marginal effects are meaningful and significant at the 5 percent level for the full treatment condition.

We also present lower bound estimates for the full sample in Panel B where lesson use from either survey was used (even if the teacher did not complete both surveys) and all missing values are assumed to be zero. While the point estimates are smaller, as expected, the general pattern of results holds.<sup>25</sup> To assess the potential role of the additional supports, we examine the effects on the number of live webinars attended (Column (8) in Panel B). While this is an imperfect measure of webinar use (because teachers could have watched the recordings asynchronously), the point estimates indicate that teachers in the full treatment watched only 0.05 more webinars.

Each Mathalicious lesson provides intuition for topics that span between 3 and 8 weeks. As such, teachers in the full treatment looked at Mathalicious lessons that could impact about one-half of the school year and report teaching lessons that could impact about two-thirds of the school year. Accordingly, while the full treatment group never reached full fidelity with the Mathalicious model (which is between 5 and 7 lessons per year), the increased lesson use likely translated into changes in instruction for a sizable proportion of the school year. Another noteworthy result is that the attendance at webinars was very low in the full treatment condition even though lesson use was higher. This suggests that the increased use in the full treatment condition was not driven by the additional supports *per se*, but may have been driven by the regular reminders to use the lessons.

## VII.2 Effects on Student Perceptions and Attitudes

The aims of the Mathalicious lessons were to promote deep student understanding, make math seem relevant to the real world, and develop greater student interest and engagement in the subject. As such, by changing the lessons teachers deliver, the intervention lessons could alter student attitudes toward mathematics. To test this, we analyze effects on student responses to an anonymous survey given at the end of the Fall semester (December) and also at the end of the experiment (May). These survey responses cannot be linked to individual students, but are linked to the math teacher. Due to permission restrictions, these survey data were collected for Chesterfield and Hanover only. On the surveys, we asked several questions on a Likert scale and used factor analysis to extract common variation from similar items. After grouping similar questions, we ended up with 6 distinct factors.<sup>26</sup> Each factor is standardized to be mean zero, unit variance.

---

<sup>25</sup>Because lesson use is essentially zero in the control condition and greater than zero in the treatment conditions, imputing zero lesson use for those who did not fill in both the mid-year and the end-of-year surveys will mechanically lead to a downward bias for those in the partial or full treatment conditions. Teachers in the license only condition looked at *at least* 1.115 more lessons and downloaded *at least* 0.916 more lessons than those in the control condition. Both effects are significant at the 5 percent level. Teachers in the full treatment condition looked at *at least* 2.236 more lessons and downloaded *at least* 1.9 more lessons than those in the control condition. Importantly, both of these differences is statistically significant at the 1 percent level. As expected, the lower bound estimate for lessons taught is smaller than among those with complete data (Panel A of Table L1 in Appendix L) or the multiple imputation results (Panel A of Table 3). These estimates indicate that teachers in the license only condition taught *at least* 0.262 more lessons, and those in the full treatment taught *at least* 0.573 more lessons than those in the control condition.

<sup>26</sup>To avoid any contamination associated with the treatments, we only used data for the control group in forming the factors. When grouping questions measuring the same construct, each group is explained by only one underlying

Teachers are only partially treated at the time of the mid-year survey, while responses at the end of the year reflect exposure to the intervention for the full duration. To account for this, among those in the license only treatment, we code the variable  $License_{dt}$  to be 1 during the end-of-year survey and 1/3 in the mid-year survey. Similarly, among those in the full treatment, we code the variable  $Full_{dt}$  to be 1 during the end-of-year survey and 1/3 in the mid-year survey.<sup>27</sup> Using data from both surveys simultaneously, we estimate the effect on student responses to the survey items using the following equation, where all variables are defined as in (1) and  $Post_{idt}$  is an indicator that is equal to 1 for the end-of-year survey and zero otherwise.

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

As with test scores, we analyze the student surveys at the student level. Table 4 presents results from models that include the full set of controls.

Credible estimation of effects on survey responses requires that survey response rates are similar across treatment arms. The first column is a model where the dependent variable is the survey response rate computed at the teacher level.<sup>28</sup> The analytic sample in this model is all students in the testing file (irrespective of whether they completed a survey) in the two participating districts. Overall, the survey response rate was 66 percent. Importantly, while the point estimates are non-trivial for the license only group, there are no statistically significant differences in survey response rates across the three treatment arms and the point estimate for the full treatment is small.<sup>29</sup>

The first factor measures whether students believe that math has real life applications. The results in Column 9 of Table 4 show that, while there is no effect for the license only condition, students of the full treatment teachers agree that math has real world applications  $0.162\sigma$  more than those of control teachers (p-value<0.05). This is consistent with the substance and stated aims of the Mathalicious lessons and confirms our priors that their content was more heavily grounded in relevant real-world examples than what teachers would have been teaching otherwise. This result also implies that identifying a high quality curriculum is a non trivial part of this intervention, and that the benefits may not be as large if lessons were of middling quality.

The next three factors measure student interest in math class, effort in math class, and motiva-

---

factor. Factor loadings for each individual question are presented in Appendix C.

<sup>27</sup>Note that our results are robust to using fractions of similar magnitude, e.g., 1/2 or 1/4.

<sup>28</sup>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 completed surveys for each teacher and weight the regressions by the total number of students with the teacher.

<sup>29</sup>In fact, in the model with no controls (Table MI in Appendix M), the survey response rate is slightly *lower* in the treatment arms than in the control group, while in the model with full controls the survey response rate is slightly *higher* in the treatment arms than in the control group. Despite this, the estimated treatment effects on the survey questions are similar in models with and without controls (for which the direction of the response rates are opposite in sign), so that any differences in response to questions are not likely driven by differential non-response.

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

The next two factors relate to student perceptions of their math teacher. They allow us to test, albeit imperfectly, Results 3 and 4 from the theoretical framework. The fifth factor measures whether students believe their math teacher emphasizes deep understanding of concepts. This relates directly to the specific aims of the Mathalicious lessons. The model predicts that the optimal lesson quality would likely increase under the treatment so that we should see increases in agreement with statements regarding the teacher promoting deeper understanding. The sixth factor measures whether students feel that their math teacher gives them individual attention. Our model predicts that off-the-shelf lessons may free up teacher time toward other teaching tasks that are complementary to lesson planning. Given that teachers do not typically plan lessons during class time, such complementary tasks would be other kinds of class preparation that may impact classroom activities. Such tasks may include deciding which students should work together, choosing homework problems, or reading student work in order to better differentiate instruction to each student in the classroom. We hypothesize that the additional class preparation time afforded by the lessons may allow teachers to better provide students with one-on-one instruction inside the classroom.<sup>30</sup> The results support the premise of our model that teachers who used the lessons improved lesson quality. Students from the full treatment group are  $0.175\sigma$  ( $p\text{-value}<0.05$ ) more likely to agree that their math teacher promotes deep understanding. Also, consistent with off-the-shelf lessons freeing up teacher time to exert more effort in complementary teaching tasks, student agreement with statements indicating that their math teacher spends more one-on-one time with them is  $0.144\sigma$  higher in the full treatment condition than in the control condition ( $p\text{-value}<0.05$ ). While the results are consistent with the time savings hypothesis, we cannot rule out that the increases in one-on-one time are due to changes in classroom practices due to using the new lessons.

In sum, we do not find strong evidence of effects on these survey measures among students in the license only condition. This may either reflect no movement on these survey measures in the license only condition or that effects of the license only condition that are too small to detect. However, students of teachers in the full treatment (for whom lesson use was more robust) say

---

<sup>30</sup>Jackson (2016) also uses more one-on-one time as a measure of teacher time.

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.<sup>31</sup>

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

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

Because randomization was within districts, one can consider each district as having its own experiment. If the benefits of the intervention were driven by lesson use, then those treatments that generated the largest increases in lesson use should also have generated the largest test score increases. To test for this, using our preferred student level models, we estimate the effects of each treatment arm (license only or full) in each of the three districts (i.e. six separate treatments) relative to the control group in each district. [Figure 2](#) 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 full treatment and 2 denotes the license only 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. We estimate a regression line through the 7 data points (including the control group located at the origin) predicting the estimated test score effect using the estimated effect on lessons taught and the treatment indicators. Conditional on treatment type, the estimated slope for lessons taught on test scores is 0.051 (p-value<0.01). To use this variation more formally, we estimate instrumental variables models predicting student math test scores and using the individual treatment arms as instruments for lessons taught (detailed in [Appendix N](#)). The preferred instrumental variables regression model yields a coefficient on lessons taught of 0.033, suggesting that for every additional lesson taught test scores increase by  $0.033\sigma$ .

---

<sup>31</sup>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 similar to that of the student surveys (61.43 percent), the response rates were substantially higher among teachers in the full treatment condition. As such, the results on the teacher surveys are inconclusive. Moreover, we do not find any systematic effects on any of the factors based on the teacher survey items. We present a detailed discussion of the teacher survey results in [Appendix D](#).

Importantly, one cannot reject the null hypothesis that the marginal effect of the full treatment is zero conditional on the lessons taught effect. These patterns indicate that (a) those treatments with larger effects on lesson use had larger test score gains and (b) the reason the full treatments had a larger effect on test scores is that they had a larger effect on lesson use.

## VIII Discussion and Conclusions

Teaching is a complex job that requires that teachers perform several complementary tasks. One important task is planning lessons. In the past few years, the availability of lesson plans and instructional material for use in the traditional classroom that can be downloaded from the Internet has increased rapidly. Today over 90 percent of secondary teachers look to the Internet for instructional materials when planning lessons [Opfer, Kaufman and Thompson \(2016\)](#) and lesson warehouse sites such as [Teachers Pay Teachers](#) have more active user accounts than teachers in the United States. Teacher use of these online lessons is a high-tech form of division of labor; classroom teachers focus on some tasks while creating instructional content is (partially) performed by others. If this technological change now provides all teachers access to high-quality lessons, the social benefits could be large. However, because there may be information barriers regarding identifying quality lessons, such benefits may not be realized. To shed light on whether providing teachers access to high-quality online instructional materials improves their student's performance, we implemented a randomized field experiment in which middle-school math teachers in three school districts were randomly provided access to high-quality, off-the-shelf lessons, and we examine the effects on their students' subsequent academic achievement.

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

Offering the lessons for free had modest effects on lesson use and modest (but economically meaningful) effects on test scores ( $0.06\sigma$ ). However, fully-treated teachers (who also received online supports to promote lesson use) used the lessons more and improved their students' test scores by about  $0.09\sigma$  relative to teachers in the control condition. These positive effects appear to have been mediated by students feeling that math had more real life applications, and having deeper levels of understanding. There is also evidence that as teachers substituted the lessons for their own lesson planning efforts, they were able engage in other tasks that facilitated spending more one-on-one time with students. The positive test score effects are largest for the weaker teachers indicating that, on average, the online lessons and teacher quality are substitutes.

Given the sizable benefits to using the off-the-shelf lessons, one may wonder why lesson use

was not even more widespread. In Appendix O, we document that lesson use was moderate during the first couple months of the intervention in both treatment arms. Lesson use decayed in both treatment arms, but did so more rapidly in the license only group. Based on survey evidence, only two percent of treated teachers mentioned that low quality this was a major factor in their lack of use, and the main reason cited for not using more lessons was a lack of time. Based on these patterns, we speculate that without the reminders and extra supports (i.e. Edmodo groups), teachers who initially were enthusiastic about using the lessons, were unable to hold themselves to make the time to implement the lessons as the school year progressed (i.e. there was a commitment problem).

Because the lessons and supports were all provided online, the per-teacher costs of the intervention are low. An upper bound estimate of the cost of the program is \$431 per teacher.<sup>32</sup> Chetty, Friedman and Rockoff (2014) estimate that a teacher who raises test scores by  $0.14\sigma$  generates marginal gains of about \$7,000 per student in present value future earnings. Using this estimate, the test score effect of about  $0.09\sigma$  would generate roughly \$4,500 in present value of future earnings per student. While this may seem like a modest benefit, consider that each teacher has about 90 students in a given year so that each teacher would generate \$405,000 in present value of students' future earnings. This implies a benefit-cost ratio of 939. Because of the low marginal cost of the intervention, it is extraordinarily cost effective. Furthermore, because the lessons and supports are provided on the Internet, the intervention is highly scalable and can be implemented in remote locations where other policy approaches would be infeasible.

As in any experimental study, one must address concerns of generalizability and the broader policy implications. The experiment was conducted in school districts at which school leaders had some pre-existing interest in the Mathalicious curriculum. As such, our estimates are most applicable to districts seeking to adopt new lessons, rather than those in which leaders are not supportive of such efforts. Also, the experiment was based on a particular curriculum. Given that lesson quality appears to be a key driver of the intervention's success, our estimates may not apply to *all* online off-the-shelf lessons, but to lessons of similarly high quality. This does not diminish the importance of the results, but rather highlights the importance of first identifying high-quality lessons prior to adopting them in districts. Given that search costs may be high relative to the private benefits for an individual teacher, this underscores the potentially important role districts can play in identifying high-quality instructional content on the Internet.

Taken as a whole, our findings show that providing teachers with access to high-quality, off-the-shelf lessons on the Internet is a viable and cost-effective alternative to the typical policies that seek to improve the skills of the existing stock of teachers through training, selection, or changes

---

<sup>32</sup>The price of an annual Mathalicious subscription is \$320. The cost of providing the additional supports (e.g., extra time for Mathalicious staff time to run Project Groundswell) was \$25,000. With 225 treated teachers, this implies an average per teacher cost of \$431. Because the subscription partly recovers fixed costs, the marginal cost is lower than this. One can treat this as an upper bound of the marginal cost.

in incentives. Our findings also suggest that policies aiming to modify the production technology of teaching (such as changes in curriculum design, innovative instructional materials, and others) may be fruitful avenues for policymakers to consider.

## References

- Akerman, Anders, Ingvil Gaarder, and Magne Mogstad.** 2015. “The Skill Complementarity of Broadband Internet.” *The Quarterly Journal of Economics*, 130(4): 1781–1824.
- Angrist, Joshua, and Victor Lavy.** 2002. “New evidence on classroom computers and pupil learning.” *The Economic Journal*, 112(482): 735–765.
- Araujo, M. Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady.** 2016. “Teacher Quality and Learning Outcomes in Kindergarten.” *The Quarterly Journal of Economics*, 131(3): 1415–1453.
- Banerjee, Abhijit V, Shawn Cole, Esther Duflo, and Leigh Linden.** 2007. “Remedying Education: Evidence from Two Randomized Experiments in India.” *The Quarterly Journal of Economics*, 122(3): 1235–1264.
- Barrow, Lisa, Lisa Markman, and Cecilia Elena Rouse.** 2009. “Technology’s Edge: The Educational Benefits of Computer-Aided Instruction.” *American Economic Journal: Economic Policy*, 52–74.
- Beuermann, Diether W, Julian Cristia, Santiago Cueto, Ofer Malamud, and Yyannu Cruz-Aguayo.** 2015. “One Laptop per Child at Home: Short-Term Impacts from a Randomized Experiment in Peru.” *American Economic Journal: Applied Economics*, 7(2): 53–80.
- Bhatt, Rachana, and Cory Koedel.** 2012. “Large-scale evaluations of curricular effectiveness: The case of elementary mathematics in Indiana.” *Educational Evaluation and Policy Analysis*, 34(4): 391–412.
- Bloom, Nicholas, Christos Genakos, Raffaella Sadun, and John Van Reenen.** 2012. “Management Practices Across Firms and Countries.” *The Academy of Management Perspectives*, 26(1): 12–33.
- Bonwell, Charles C, and James A Eison.** 1991. *Active Learning: Creating Excitement in the Classroom*. 1991 ASHE-ERIC Higher Education Reports. ERIC.
- Brown, John Seely, Allan Collins, and Paul Duguid.** 1989. “Situated cognition and the culture of learning.” *Educational Researcher*, 18(1): 32–42.
- Bugni, Federico, Ivan Canay, and Azeem Shaikh.** 2017. “Inference under Covariate Adaptive Randomization with Multiple Treatments.”
- Bulman, George, and Robert W. Fairlie.** 2016. “Technology and Education: Computers, Software, and the Internet.” *NBER Working Paper w22237*.
- Card, David, and Alan B Krueger.** 1992. “Does school quality matter? Returns to education and the characteristics of public schools in the United States.” *Journal of political Economy*, 100(1): 1–40.

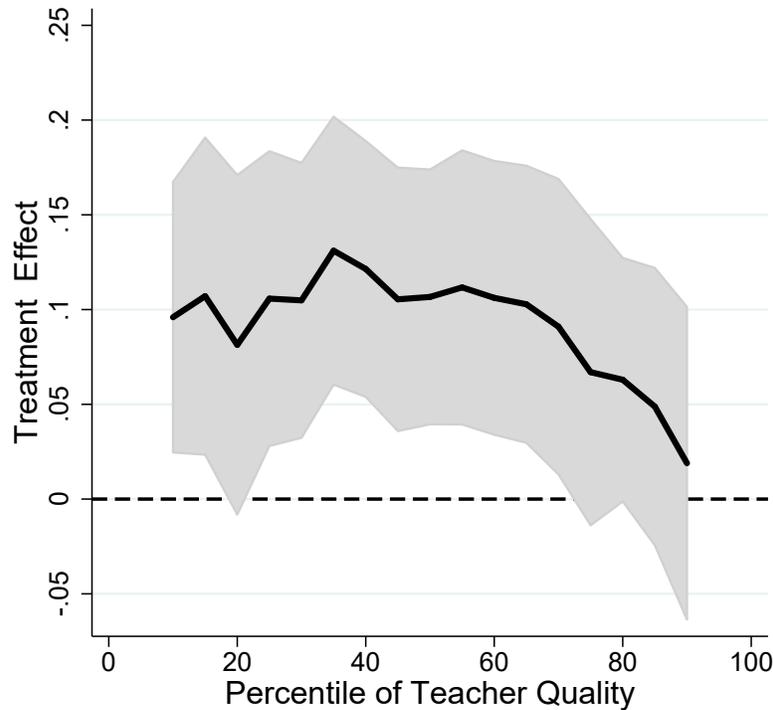
- Chetty, Raj, John N Friedman, and Jonah E Rockoff.** 2014. “Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood.” *American Economic Review*, 104(9): 2633–2679.
- Chetty, Raj, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan.** 2011. “How does your kindergarten classroom affect your earnings? Evidence from Project STAR.” *The Quarterly Journal of Economics*, 126(4): 1593–1660.
- Chingos, Matthew M, and Grover J Whitehurst.** 2012. “Choosing Blindly: Instructional Materials, Teacher Effectiveness, and the Common Core.” *Brookings Institution*.
- Comi, Simona, Marco Gui, Federica Origo, Laura Pagani, and Gianluca Argentin.** 2016. “Is it the way they use it? Teachers, ICT and student achievement.” *DEMS Working Paper No. 341*.
- Darling-Hammond, Linda, Ruth Chung Wei, Alethea Andree, Nikole Richardson, and Stelios Orphanos.** 2009. “Professional learning in the learning profession.”
- Deming, David.** 2009. “Early childhood intervention and life-cycle skill development: Evidence from Head Start.” *American Economic Journal: Applied Economics*, 111–134.
- Dostál, Jiri.** 2015. *Inquiry-based instruction : Concept, essence, importance and contribution*. Palacky University Olomouc.
- Eichholtz, Piet, Nils Kok, and John M Quigley.** 2010. “Doing well by doing good? Green office buildings.” *The American Economic Review*, 100(5): 2492–2509.
- Fryer, Roland G.** 2016. “The ‘Pupil’ Factory: Specialization and the Production of Human Capital in Schools.” *NBER Working Paper w22205*.
- Hanushek, Eric A.** 1974. “Efficient estimators for regressing regression coefficients.” *The American Statistician*, 28(2): 66–67.
- Heckman, James J, and Dimitriy V Masterov.** 2007. “The productivity argument for investing in young children.” *Applied Economic Perspectives and Policy*, 29(3): 446–493.
- Holmstrom, Bengt, and Paul Milgrom.** 1991. “Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design.” *Journal of Law, Economics, & Organization*, 24–52.
- Jackson, C. Kirabo.** 2016. “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*.
- Jackson, C. Kirabo.** 2017. “What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes.” *Journal of Political Economy*, forthcoming.
- Jackson, C Kirabo, and Henry S Schneider.** 2015. “Checklists and Worker Behavior: A Field Experiment.” *American Economic Journal: Applied Economics*, 7(4): 136–168.
- Jackson, C Kirabo, Jonah E Rockoff, and Douglas O Staiger.** 2014. “Teacher Effects and Teacher-Related Policies.” *Annual Review of Economics*, 6(1): 801–825.
- Jackson, C. Kirabo, Rucker C. Johnson, and Claudia Persico.** 2016. “The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms.” *The Quarterly Journal of Economics*, 131(1): 157–218.

- Jacob, Brian A, and Jonah E Rockoff.** 2012. “Organizing schools to improve student achievement: Start times, grade configurations, and teacher assignments.” *The Education Digest*, 77(8): 28.
- Kane, Thomas J, and Douglas O Staiger.** 2008. “Estimating teacher impacts on student achievement: An experimental evaluation.” *NBER Working Paper w14607*.
- Kane, TJ, AM Owens, WH Marinell, DRC Thal, and DO Staiger.** 2016. “Teaching higher: Educators’ perspectives on Common Core implementation.” *Center for Education Policy Research*. Retrieved from: <http://cepr.harvard.edu/files/cepr/files/teaching-higher-report.pdf>.
- Katz, Lawrence F.** 1999. “Changes in the wage structure and earnings inequality.” *Handbook of Labor Economics*, 3: 1463–1555.
- Koedel, Cory, Diyi Li, Morgan S. Polikoff, Tenice Hardaway, and Stephani L. Wrabel.** 2017. “Mathematics Curriculum Effects on Student Achievement in California.” *AERA Open*, 3(1).
- Koenker, Roger, and Gilbert Bassett.** 1978. “Regression quantiles.” *Econometrica*, 33–50.
- Krueger, Alan B.** 1999. “Experimental estimates of education production functions.” *The quarterly journal of economics*, 114(2): 497–532.
- Lave, Jean, and Etienne Wenger.** 1991. *Situated learning: Legitimate peripheral participation*. Cambridge University Press.
- Lesh, Richard, and Helen Doerr.** 2003. “Foundations of a model and modeling perspective on mathematics teaching, learning, and problem solving.”
- Lewis, Jeffrey B, and Drew A Linzer.** 2005. “Estimating regression models in which the dependent variable is based on estimates.” *Political analysis*, 13(4): 345–364.
- Madda, Mary.** 2016. “Amazon Launches ‘Inspire,’ a Free Education Resource Search Platform for Educators.” *EdSurge*.
- Mihaly, Kata, Daniel F McCaffrey, Douglas O Staiger, and JR Lockwood.** 2013. “A composite estimator of effective teaching.” *Seattle, WA: Bill & Melinda Gates Foundation*.
- Muralidharan, Karthik, and Venkatesh Sundararaman.** 2013. “Contract teachers: Experimental evidence from India.” *NBER Working Paper w19440*.
- Opfer, V Darleen, Julia H Kaufman, and Lindsey E Thompson.** 2016. “Implementation of K–12 State Standards for Mathematics and English Language Arts and Literacy.” RAND Corporation.
- Pianta, Robert C.** 2011. “Teaching Children Well: New Evidence-Based Approaches to Teacher Professional Development and Training.” *Center for American Progress*.
- Purcell, Kristen, Alan Heaps, Judy Buchanan, and Linda Friedrich.** 2013. “How Teachers Are Using Technology at Home and in Their Classrooms.” *Pew Research Center*.
- Rivkin, Steven G, Eric A Hanushek, and John F Kain.** 2005. “Teachers, schools, and academic achievement.” *Econometrica*, 417–458.
- Rothstein, Jesse.** 2014. “Teacher quality policy when supply matters.” *The American Economic Review*, 105(1): 100–130.
- Rouse, Cecilia Elena, and Alan B Krueger.** 2004. “Putting computerized instruction to the test: a randomized evaluation of a “scientifically based” reading program.” *Economics of Education Review*, 23(4): 323–338.

- Rubin, Donald B.** 2004. *Multiple imputation for nonresponse in surveys*. Vol. 81, John Wiley & Sons.
- Sawyer, R Keith.** 2005. *The Cambridge handbook of the learning sciences*. Cambridge University Press.
- Schafer, Joseph L.** 1997. *Analysis of incomplete multivariate data*. CRC press.
- Stigler, James W, Patrick Gonzales, Takako Kwanaka, Steffen Knoll, and Ana Serrano.** 1999. "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."
- Taylor, Eric S.** 2015. "New Technology and Teacher Productivity."
- Taylor, Eric S, and John H Tyler.** 2012. "The effect of evaluation on teacher performance." *The American Economic Review*, 102(7): 3628–3651.

## Tables and Figures

Figure 1. Marginal Effect of the Full Treatment by Teacher Quality.  
Mathematics Test Scores.



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

Figure 2. 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 2, and the ‘Full Treatment’ is denoted by the number 1. The Y-axis displays coefficients for specifications identical to those estimated in Columns (5) of Table 2. The X-axis displays coefficients for specifications similar to those estimated in Panel C Column (2) of Table 3. However, all regressions are estimated based on a restricted sample within each district that compares each treatment group to the control group in the same district. For example, the ‘Chesterfield 2’ label means that the corresponding point displays the coefficients from the aforementioned regressions estimated within Chesterfield only and without the ‘Full Treatment’ teachers. The black line represents the best linear prediction based on six points displayed on each graph. The size of the dots corresponds to the relative size of the district-treatment groups in terms of the number of students.

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 <sup>a</sup>	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 <sup>b</sup>	0.818	2.123	0.275	0.750	1.519	0.053*
	Total no. Mathalicious lessons the teacher taught or read	236 <sup>b</sup>	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 <sup>c</sup>	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
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 <sup>d</sup>	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 <sup>d</sup>	0.015	0.997	-0.010	-0.025	0.077	0.690
Reading SOL scores, standardized by grade, 2014	24409 <sup>e</sup>	0.008	0.997	-0.021	-0.027	0.068	0.969	

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

Table 2. Effects on Student Test Scores.

	Mathematics						Falsification: English	
	2014 Raw Score	2014 Raw Score	2014 Standardized Score	2014 Standardized Score	2014 Standardized Score	2014 Standardized Score	2014 Raw Score	2014 Standardized Score
<i>Panel A: Baseline Results.</i>								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
License Only	2.653 [2.136]	3.583* [1.926]	0.050 [0.040]	0.061* [0.034]	0.060* [0.033]	0.055* [0.032]	1.105 [1.041]	0.025 [0.019]
Full Treatment	7.899*** [2.662]	7.057*** [2.308]	0.105*** [0.046]	0.094** [0.038]	0.086** [0.038]	0.093*** [0.035]	0.460 [1.223]	0.008 [0.022]
District FE x Requested	Y	Y	Y	Y	Y	Y	Y	Y
District FE x Teacher-Level Lagged Test Scores	Y	Y	Y	Y	Y	Y	Y	Y
District FE x Individual Lagged Test Scores	N	N	N	N	Y	N	Y	Y
All controls	N	Y	N	Y	Y	Y	Y	Y
Observations	27,613	27,613	27,613	27,613	27,613	363	25,038	25,038
Unit of Observation	Student	Student	Student	Student	Student	Teacher	Student	Student
<i>Panel B: Baseline Results without Requested Teachers.</i>								
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
License Only	2.125 [2.111]	2.684 [2.023]	0.039 [0.038]	0.042 [0.036]	0.043 [0.036]	0.048 [0.035]	-0.688 [1.050]	-0.012 [0.019]
Full Treatment	9.382*** [2.904]	8.714*** [2.692]	0.124*** [0.046]	0.108** [0.045]	0.101** [0.044]	0.117*** [0.043]	1.880 [1.450]	0.030 [0.026]
District FE x Requested	Y	Y	Y	Y	Y	Y	Y	Y
District FE x Teacher-Level Lagged Test Scores	Y	Y	Y	Y	Y	Y	Y	Y
District FE x Individual Lagged Test Scores	N	N	N	N	Y	N	Y	Y
All controls	N	Y	N	Y	Y	Y	Y	Y
Observations	16,883	16,883	16,883	16,883	16,883	241	14,427	14,427
Unit of Observation	Student	Student	Student	Student	Student	Teacher	Student	Student

Notes: \*\*\* - significance at less than 1%; \*\* - significance at 5%; \* - significance at 10%. Standard errors clustered at the teacher level are reported in square brackets. All specifications include controls for the requested indicator, average teacher-level 2013 math and reading test scores, and teacher-level shares of students with missing 2013 math and reading test scores - all interacted with district fixed effects. So that we can include all students with math scores in 2014 in regression models, students with missing 2013 standardized math and reading scores were given an imputed score of zero. To account for this in regression models, we also include indicators denoting these individuals in all specifications. Results are robust to restricting the sample to students with complete data. Columns (5), (7), and (8) control for individual-level 2013 math and reading test scores. Additional student-level controls include race, and gender. Additional teacher-level controls include teachers' educational attainment, years of experience, sex, race, grade fixed effects, as well as the percentage of male, black, white, Asian, and Hispanic students in the classroom. Standardized scores refer to the raw scores standardized by exam type. In the absence of exam type data for Hanover, test scores for that district were standardized by grade. Columns (9)-(16) replicate the baseline analysis on a subsample without Requested teachers.

Table 3. Effects on Lesson Use.

Panel A: Multiple Imputation Estimates. Missing Outcome Data Imputed Using Multiple Imputation.

	Lessons Looked	Lessons Taught	Lessons Looked	Lessons Taught
	(1)	(2)	(3)	(4)
License Only	1.586*** [0.418]	0.657*** [0.191]	1.726*** [0.499]	0.640*** [0.184]
Full Treatment	4.404*** [0.605]	1.925*** [0.282]	3.058*** [0.594]	2.017*** [0.467]
District FE x Requested	Y	Y	Y	Y
All controls	Y	Y	Y	Y
Sample of teachers	All	All	Non-requesters	Non-requesters
Observations	363	363	241	241

Panel B: Full Sample Estimates. Missing Data for Lessons Looked and Taught Replaced with Zero (Lower Bound).

	Lessons Looked	Lessons Taught	Lessons Downloaded	Webinars Viewed
	(5)	(6)	(7)	(8)
License Only	1.115*** [0.422]	0.262 [0.221]	0.916*** [0.328]	-0.013 [0.009]
Full Treatment	2.236*** [0.506]	0.573** [0.238]	1.900*** [0.457]	0.048** [0.022]
District FE x Requested	Y	Y	Y	Y
All controls	Y	Y	Y	Y
Observations	363	363	363	363

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

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

	Share of Completed Surveys	Standardized Factors					
		Math has Real Life Application	Increased Interest in Math Class	Increased Effort in Math Class	Increased Motivation for Studying in General	Math Teacher Promotes Deeper Understanding	Math Teacher Gives Individual Attention
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
License Only	0.100 [0.082]	-0.012 [0.060]	-0.018 [0.062]	0.045 [0.035]	-0.021 [0.035]	0.001 [0.065]	0.033 [0.063]
Full Treatment	0.012 [0.099]	0.162** [0.063]	0.087 [0.074]	0.003 [0.044]	0.039 [0.035]	0.175** [0.070]	0.144** [0.069]
End-of-Year Indicator	Y	Y	Y	Y	Y	Y	Y
District FE x Requested	Y	Y	Y	Y	Y	Y	Y
All controls	Y	Y	Y	Y	Y	Y	Y
Observations	27,450	17,959	17,799	17,954	17,768	17,843	18,443

Notes: \*\*\* - significance at less than 1%; \*\* - significance at 5%, \* - significance at 10%. Standard errors clustered at the teacher level are reported in square brackets. For details on the estimating strategy, see (3). Each outcome, except for the share of completed surveys, is a result of factor analysis and encompasses variation from several individual questions. For details on how the factors were formed, see Appendix C. All specifications include controls for district fixed effects, average teacher-level 2013 math and reading test scores, and teacher-level shares of students with missing 2013 math and reading test scores (all interacted with the requested indicator), as well as teachers' education level, years of experience, sex, race, grade fixed effects, and the percentage of male, black, white, Asian, and Hispanic students in their class. The fact that the survey was anonymous prevented us from including any student-level covariates. The regressions presented in Column (1) are estimated at the teacher level. The share of completed surveys for each teacher was calculated by comparing the number of completed student surveys with the number of students with complete data on math test scores.

**For Online Publication.**

**Appendix A. Treatment Allocation.**

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

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

## Appendix B. Auxiliary Results Regarding Requested Teachers.

Table B1. Summary Statistics: Requested vs. Non-Requested Teachers.

Variable	N	Mean	SD	Mean (Not Requested)	Mean (Requested)	P-value for balance hypothesis
	(1)	(2)	(3)	(4)	(5)	(7)
Has MA degree	363	0.424	0.495	0.419	0.434	0.781
Has PhD degree	363	0.008	0.091	0.008	0.008	0.992
Teacher is female	363	0.802	0.399	0.793	0.820	0.534
Years teaching <sup>a</sup>	363	11.730	8.628	12.591	10.029	0.003***
Teacher is white	363	0.884	0.320	0.871	0.910	0.256
Teacher is black	363	0.096	0.296	0.108	0.074	0.273
Grade 6	363	0.311	0.464	0.299	0.336	0.474
Grade 7	363	0.366	0.482	0.378	0.344	0.532
Grade 8	363	0.342	0.475	0.361	0.303	0.268

*Notes:* \*\*\* - significance at less than 1%; \*\* - significance at 5%, \* - significance at 10%. The test of equality of the group means is performed using a regression of each characteristic on the requested indicator and a constant. P-values for the significance of the requested indicator are reported in Column (7) and are calculated based on robust standard errors.

Table B2. Main Result by Requested Status.

	2014 Raw Math Score	2014 Raw Math Score	2014 Standardized Math Score	2014 Standardized Math Score	2014 Standardized Math Score	2014 Standardized Math Score
	(1)	(2)	(3)	(4)	(5)	(6)
License Only	2.337 [2.192]	3.124 [2.148]	0.043 [0.039]	0.049 [0.038]	0.047 [0.037]	0.053 [0.036]
Full Treatment	9.312*** [2.994]	8.391*** [2.823]	0.123*** [0.047]	0.100** [0.045]	0.089** [0.045]	0.115*** [0.042]
License Only x Requested	0.551 [5.923]	1.254 [4.965]	0.019 [0.115]	0.044 [0.087]	0.045 [0.085]	-0.003 [0.083]
Full Treatment x Requested	-2.912 [5.686]	-2.497 [4.829]	-0.034 [0.108]	0.003 [0.082]	0.009 [0.082]	-0.058 [0.077]
District FE x Requested	Y	Y	Y	Y	Y	Y
District FE x Teacher-Level Lagged Test Scores	Y	Y	Y	Y	Y	Y
District FE x Individual Lagged Test Scores	N	N	N	N	Y	N
All controls	N	Y	N	Y	Y	Y
Joint p-value for Treatment x Requested	0.265	0.416	0.627	0.799	0.902	0.547
Observations	27,613	27,613	27,613	27,613	27,613	363
Unit of Observation	Student	Student	Student	Student	Student	Teacher

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

## Appendix C. Construction of Factors for The Student Survey.

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

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

## Appendix D. 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 D1 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 D1. 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	0.071 [0.065]	-0.033 [0.093]	0.007 [0.241]	0.077 [0.188]	-0.024 [0.201]	-0.142 [0.164]	0.010 [0.203]	0.065 [0.222]	0.056 [0.190]	0.064 [0.177]
Full Treatment	0.079 [0.076]	0.117 [0.093]	-0.042 [0.266]	0.003 [0.195]	-0.100 [0.207]	-0.090 [0.176]	-0.019 [0.217]	-0.193 [0.201]	0.173 [0.207]	0.004 [0.209]
District FE x Requested	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
All controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	363	209	209	205	203	207	206	204	205	205

Notes: \*\*\* - significance at less than 1%; \*\* - significance at 5%; \* - significance at 10%. Robust standard errors are reported in square brackets. Factors are obtained through factor analysis of related survey questions. For details, see exact factor loadings in Table D2. All specifications include controls for the requested indicator, average teacher-level 2013 math and reading test scores, and teacher-level shares of students with missing 2013 math and reading test scores - all interacted with district fixed effects. Other controls include teachers' education level, years of experience, sex, race, grade fixed effects, as well as the percentage of male, black, white, Asian, and Hispanic students in their class.

Table D2. Teacher Post-Treatment Survey. Factor Loadings.

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

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

## Appendix E. Stylized Model of Teacher Multitasking.

### E1. 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  $s$ ) and leisure ( $l$ ). Student  $i$ 's test score depends on how much time the teacher spends planning lessons ( $d$ ) and other complementary teaching tasks ( $n$ ). A teacher's (in)ability to plan lessons is modeled as a 'price'  $p_d$  that amplifies the time needed to achieve  $d$  units of lesson quality. Similarly, 'price'  $p_n$  denotes the teacher's ability to achieve  $n$  units of other teaching tasks. Note that the higher teacher abilities are, the lower are her corresponding  $p$ 's. Each teacher chooses the allocation of her total time ( $T$ ) toward leisure ( $l$ ), lesson planning ( $d$ ) and other teaching tasks ( $n$ ) in order to maximize her utility. Formally, we write:

$$\begin{aligned} U\left(\left\{y_i(n,d)\right\}_{i=1}^s, l\right) &\rightarrow \max_{\{n,d,l\}} & (4) \\ \text{s.t. } p_n n + p_d d + l &\leq T \\ n \geq 0; d \geq 0; l &\geq 0 \end{aligned}$$

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

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

### E2. Special Case with Functional Form Assumptions

For the ease of exposition, we will consider a special case of the model with several functional form assumptions. First, let  $U$  be a weakly separable function where weighted average of students' test scores multiplied by a function of leisure:

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

Furthermore, let  $y_i$  be a Cobb-Douglas-type function with elasticities  $\alpha, \beta \in [0, 1]$ , but with a student-level heterogeneity parameter  $w_i$ :  $y_i(n,d) = w_i n^\alpha d^\beta$ .<sup>33</sup> We parametrize utility derived from leisure with a Cobb-Douglas-like function:  $g(l) = l^\gamma$ , where  $\gamma \in [0, 1]$ . For simplicity, we define

---

<sup>33</sup> $w_i > 0$  can be interpreted as student  $i$ 's ability, an individual shock parameter, the weight with which a teacher values student  $i$ 's test score, or some combination of these.

$b = \frac{1}{s} \sum_{i=1}^s w_i$ . Therefore, based on our assumptions, the utility function is:

$$U(n, d, l) = bn^\alpha d^\beta l^\gamma$$

## E2.1. Baseline Results

### No Off the shelf Lessons

With the above assumptions in mind, we solve for the case of *no off-the-shelf lessons*:

$$n^* = \frac{\alpha}{\alpha + \beta + \gamma} \frac{T}{p_n}; d^* = \frac{\beta}{\alpha + \beta + \gamma} \frac{T}{p_d}; l^* = \frac{\gamma}{\alpha + \beta + \gamma} T$$

$$U^* = \left[ b \frac{\alpha^\alpha \beta^\beta \gamma^\gamma}{(\alpha + \beta + \gamma)^{\alpha + \beta + \gamma}} \right] \left[ \frac{T^{\alpha + \beta + \gamma}}{p_n^\alpha p_d^\beta} \right]$$

### With Off the shelf Lessons

Next, we solve for the case *with off-the-shelf lessons*. As depicted in Figure E1, an adopting teacher may choose to locate along the new higher budget constraint or they may locate at the kink. We solve for each scenario below. Even though teachers cannot adopt lessons and locate below the kink, for analytical purpose it is helpful to describe teacher behaviors under this hypothetical situation.

#### Ignoring the $d \geq \underline{d}$ restriction

First, we solve for the interior solution with off-the-shelf lessons but ignoring the  $d \geq \underline{d}$  restriction and define some useful parameters under this condition:

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

$$\tilde{U} = \left[ b \frac{\alpha^\alpha \beta^\beta \gamma^\gamma}{(\alpha + \beta + \gamma)^{\alpha + \beta + \gamma}} \right] \left[ \frac{(T - F + p_d \underline{d})^{\alpha + \beta + \gamma}}{p_n^\alpha p_d^\beta} \right]$$

In this scenario (where we allow teachers to locate on the infeasible portion of the budget constraint under lesson use), the off-the-shelf lessons function like an increase in time that is allocated to all tasks.

#### Imposing the $d \geq \underline{d}$ restriction

However, it could be that  $\tilde{d} < \underline{d}$  such that the adopting teacher would locate at the kink of the budget line, as in case B in Figure E1. To solve for maximum teacher utility at the kink, one sets  $d^K = \underline{d}$  and maximizes teacher utility with respect to  $l$  and  $n$ . In this case, the values of the main variables are:

$$n^K = \frac{\alpha}{\alpha + \gamma} \frac{T - F}{p_n}; d^K = \underline{d}; l^K = \frac{\beta}{\alpha + \gamma} (T - F); U^K = \left[ b \frac{\alpha^\alpha \gamma^\gamma}{(\alpha + \gamma)^{\alpha + \gamma}} \right] \left[ \frac{\underline{d}^\beta (T - F)^{\alpha + \gamma}}{p_n^\alpha} \right]$$

## E2.2. Auxiliary Lemmas.

Before we move forward to proving the main results of the model, we state the following two lemmas.<sup>34</sup>

**Lemma 1.** The utility achieved with off-the-shelf lessons without the  $d \geq \underline{d}$  restriction is always weakly larger than the utility with off-the-shelf lessons when located at the kink. Simply put, allowing a teacher to locate below the kink cannot make them worse off than not allowing them to do so. Formally,  $\tilde{U} \geq U^K$ .

**Proof.** If  $\tilde{d} > \underline{d}$ , then the kink would not have been chosen with off-the-shelf lessons and  $\tilde{U} > U^K$  by construction. If  $\tilde{d} \leq \underline{d}$ , then the adopting teacher would locate at the kink. However, if one would remove the  $d \geq \underline{d}$  restriction, then, by a revealed preference argument, the adopting teacher would become at least weakly better off.

**Lemma 2.** A teacher adopts off-the-shelf lessons in one of two cases: (i) whenever  $\tilde{U} \geq U^* \geq U^K$  or  $\tilde{U} \geq U^K \geq U^*$  and  $\tilde{U}$  is attainable because  $\tilde{d} \geq \underline{d}$  (case A in Figure E1), and (ii) whenever  $\tilde{U} \geq U^K \geq U^*$  and  $\tilde{U}$  is unattainable because  $\tilde{d} < \underline{d}$  (case B in Figure E1).

**Proof.** Clearly, if  $U^* \geq \max\{\tilde{U}, U^K\}$ , then a teacher does not adopt because she would be better off without the lessons. Moreover, from Lemma 1, we know that  $\tilde{U} \geq U^K$ . Hence, the only two cases when a teacher adopts are either  $\tilde{U} \geq U^* \geq U^K$  or  $\tilde{U} \geq U^K \geq U^*$ . However, if  $\tilde{U} \geq U^* \geq U^K$  and  $\tilde{U}$  is unattainable because  $\tilde{d} > \underline{d}$ , then the teacher does not adopt because she is better off without the lessons ( $U^* \geq U^K$ ).

## E2.3. Predictions.

**Proposition 1.** *The effect of lesson adoption on lesson quality,  $d$ , is non-negative.*

**Proof.** Consider the first case (i) from Lemma 2.  $\tilde{U}$  is attainable ( $\tilde{d} \geq \underline{d}$ ) and  $\tilde{U} \geq U^*$ . By comparing the two functions, one gets that:  $\tilde{U} \geq U^* \iff F \leq p_d \underline{d}$ . Hence,  $\tilde{d} = (\beta/(\alpha + \beta + \gamma))(T - F + p_d \underline{d})/p_d \geq (\beta/(\alpha + \beta + \gamma))(T/p_d) = d^*$ . Thus, in this case, lesson quality does not decrease after lesson adoption.

Consider the second case (ii) from Lemma 2. In this case, since  $\tilde{U}$  is unattainable, it must be that  $\underline{d} > \tilde{d}$ . However, we still have that  $\tilde{U} \geq U^* \iff F \leq p_d \underline{d}$ . Therefore, we get that  $\underline{d} \geq \tilde{d} = (\beta/(\alpha + \beta + \gamma))(T - F + p_d \underline{d})/p_d \geq (\beta/(\alpha + \beta + \gamma))(T/p_d) = d^*$ . Hence, adoption of the off-the-shelf lessons does not lead to a decrease in lesson quality,  $d$ . ■<sup>35</sup>

**Proposition 2.** *The effect of lesson adoption on time spent on other teaching tasks,  $n$ , is ambiguous in sign.*

**Proof.** Consider the first case (i) from Lemma 2.  $\tilde{U}$  is attainable ( $\tilde{d} \geq \underline{d}$ ) and  $\tilde{U} \geq U^*$ . By comparing the two functions, one gets that:  $\tilde{U} \geq U^* \iff F \leq p_d \underline{d}$ . Hence,  $\tilde{n} = (\alpha/(\alpha + \beta + \gamma))(T - F + p_d \underline{d})/p_n \geq (\alpha/(\alpha + \beta + \gamma))(T/p_n) = n^*$ . Thus, in this case, time spent on other tasks does not decrease after lesson adoption.

Consider the second case (ii) from Lemma 2. In this case, we get that whether  $n^K \geq n^*$  or not will depend on the parameters of the model. Specifically,  $n^K \geq n^*$  whenever  $F/T \leq \beta/(\alpha + \beta + \gamma)$ ,

<sup>34</sup>Note that these two lemmas do not require the Cobb-Douglas functional form assumptions and would hold under any other monotonic utility function.

<sup>35</sup>It is worth noting that this result holds regardless of the Cobb-Douglas functional form assumptions as long as  $n$ ,  $d$ , and  $l$  are all normal goods. For intuition, observe that  $d$  increases in both cases A and B in Figure E1 and that similar figures can be drawn for any monotonic and quasi-concave utility function.

while  $n^K < n^*$  whenever  $F/T > \beta/(\alpha + \beta + \gamma)$ . ■<sup>36</sup>

**Proposition 3.** *The gains in average test scores from using the off-the-shelf lessons are non-negative.*

**Proof.** Consider the first case (i) from Lemma 2. From Proposition 1 and Proposition 2, we know that in this case both  $n$  and  $d$  weakly go up with lesson adoption. Hence, average test scores,  $bn^\alpha d^\beta$ , weakly go up. That is, if a teacher adopts and does not locate at the kink, then test scores will weakly increase.

Consider the second case (ii) from Lemma 2. To start, let us derive the condition under which test scores are higher at the kink than without lesson use (i.e. test scores increase at the kink):

$$b(n^K)^\alpha (d^K)^\beta = b \underline{d}^\beta \left( \frac{\alpha}{\alpha + \gamma} \frac{T - F}{p_n} \right)^\alpha \geq b \left( \frac{\alpha^\alpha \beta^\beta}{(\alpha + \beta + \gamma)^{\alpha + \beta}} \right) \frac{T^{\alpha + \beta}}{p_n^\alpha p_d^\beta} = b(n^*)^\alpha (d^*)^\beta \quad (6)$$

$$(p_d \underline{d})^\beta \geq \frac{\beta^\beta (\alpha + \gamma)^\alpha}{(\alpha + \beta + \gamma)^{\alpha + \beta}} \frac{T^{\alpha + \beta}}{(T - F)^\alpha} = C \quad (7)$$

Intuitively, if adopting the lesson and locating at the kink increases tests scores, the time savings must be large enough for test scores to increase if the adopting teacher locates at the kink of the budget line.

We now turn to the adoption condition. In order for the teacher to adopt the lessons, it must be that  $U^K \geq U^*$ . We can write what this condition implies in terms of parameters of the model:

$$U^K = b \frac{\alpha^\alpha \gamma^\gamma}{(\alpha + \gamma)^{\alpha + \gamma}} \left( \frac{d^\beta (T - F)^{\alpha + \gamma}}{p_n^\alpha} \right) \geq b \frac{\alpha^\alpha \beta^\beta \gamma^\gamma}{(\alpha + \beta + \gamma)^{\alpha + \beta + \gamma}} \left( \frac{T^{\alpha + \beta + \gamma}}{p_n^\alpha p_d^\beta} \right) = U^* \quad (8)$$

$$(p_d \underline{d})^\beta \geq \frac{\beta^\beta (\alpha + \gamma)^{\alpha + \gamma}}{(\alpha + \beta + \gamma)^{\alpha + \beta + \gamma}} \frac{T^{\alpha + \beta + \gamma}}{(T - F)^{\alpha + \gamma}} \quad (9)$$

$$(p_d \underline{d})^\beta \geq C \left[ \frac{\alpha + \gamma}{\alpha + \beta + \gamma} \frac{T}{T - F} \right]^\gamma \quad (10)$$

Note that if  $\frac{\alpha + \gamma}{\alpha + \beta + \gamma} \frac{T}{T - F} = 1$  and (7) holds, then (4) follows immediately.

If instead  $\frac{\alpha + \gamma}{\alpha + \beta + \gamma} \frac{T}{T - F} > 1$  and (7) is true, then (4) also holds since for any  $D > 1 \implies (p_d \underline{d})^\beta \geq CD > C$ .

However, if  $\frac{\alpha + \gamma}{\alpha + \beta + \gamma} \frac{T}{T - F} < 1$ , condition (7) does not tell us anything about condition (4) as for any  $D < 1$  it could be that  $C > (p_d \underline{d})^\beta \geq CD$ . To make progress on this case, note that re-writing the first inequality leads to  $F/T < \beta/(\alpha + \beta + \gamma)$ . However, from Proposition 2, we know that in this case both  $n^K > n^*$  and  $d^K > d^*$ , meaning that test scores must go up as  $b(n^K)^\alpha (d^K)^\beta > b(n^*)^\alpha (d^*)^\beta$ . Hence, adoption of off-the-shelf lessons leads to a non-negative effect on test scores. ■<sup>37</sup>

<sup>36</sup>This result also does not require the Cobb-Douglas functional form assumptions. For intuition, observe that  $n$  increases in case A in Figure E1 but decreases in case B in Figure E1 and that similar figures can be drawn for any monotonic and quasi-concave utility function.

<sup>37</sup>Note that the first part of the proposition does not require the Cobb-Douglas functional form assumptions. In fact, it holds for any utility function for which  $n$ ,  $d$ , and  $l$  are all normal. For such functions, as long as the teachers who adopt the lessons put at least some extra effort into increasing lesson quality above the minimum,  $\tilde{d} \geq \underline{d}$ , test scores will not decrease.

Intuitively, if the fixed cost of adoption is low enough, then both  $n$  and  $d$  at the kink should always be higher than  $n$  and  $d$  without lessons. However, if the fixed cost is high enough,  $n$  can go down, as shown in Proposition 2. But then the utility at the kink must be bigger than the utility without lessons for a teacher to adopt. From this condition, one can derive that, if the fixed cost is high and teacher adopts and locates the kink, it must be that test scores do not decrease. Hence, either way, adoption will not occur at the cost of a decrease in test scores.

**Proposition 4.** *The relationship between the test score benefits of lesson use and teacher quality is ambiguous in sign as it depends on the definition of teacher quality. To see this, assume  $\alpha + \beta = 1$  and an interior (non-kink and non-corner) solution.*

**Case 1:** *In this scenario, better teachers are those that can produce better test scores with less time than weaker teachers. If teacher quality is defined as a set of  $(p_n, p_d)$ , an increment in test scores  $[b\tilde{n}^\alpha \tilde{d}^\beta - b(n^*)^\alpha (d^*)^\beta]$  goes down with a decrease in prices.*

**Case 2:** *Teacher quality is defined as ability to adopt new ways of teaching. As a simplification, consider that low teaching quality means that a teacher has a higher  $F$ . In this case, the difference  $[b\tilde{n}^\alpha \tilde{d}^\beta - b(n^*)^\alpha (d^*)^\beta]$  goes up with a decrease in  $F$ .*

**Proof.** Conditional on  $\alpha + \beta = 1$  and an interior (non-kink and non-corner) solution, an increment in utility from adopting the lessons takes the following form:

$$[b\tilde{n}^\alpha \tilde{d}^\beta - b(n^*)^\alpha (d^*)^\beta] = b \frac{\alpha^\alpha \beta^\beta \gamma^\gamma}{(\alpha + \beta + \gamma)^{\alpha + \beta + \gamma}} \left( \frac{p_d \underline{d} - F}{p_n^\alpha p_d^\beta} \right) \quad (11)$$

**Case 1:** First, one can show that the gains in test scores from adopting the lessons are strictly increasing in  $p_d$ :

$$\frac{\partial [b\tilde{n}^\alpha \tilde{d}^\beta - b(n^*)^\alpha (d^*)^\beta]}{\partial p_d} = b \frac{\alpha^\alpha \beta^\beta \gamma^\gamma}{(\alpha + \beta + \gamma)^{\alpha + \beta + \gamma}} \left[ (1 - \beta) p_n^{-\alpha} p_d^{-\beta} \underline{d} + \beta p_n^{-\alpha} p_d^{-\beta - 1} F \right] > 0$$

Moreover, one can prove that, when both  $p_d$  and  $p_n$  are increased *simultaneously by the same percentage*, the difference strictly increases. Specifically, after taking the exact differential of (11), we show that simultaneous increases of  $p_n$  and  $p_d$  by the same percentage (i.e. such that  $dp_n/p_n = dp_d/p_d = \varepsilon$ ) lead to an increase of the total difference:

$$\begin{aligned} d[b\tilde{n}^\alpha \tilde{d}^\beta - b(n^*)^\alpha (d^*)^\beta] &= b \frac{\alpha^\alpha \beta^\beta \gamma^\gamma}{(\alpha + \beta + \gamma)^{\alpha + \beta + \gamma}} \left[ -\alpha p_n^{-\alpha} (p_d^{1-\beta} \underline{d} - p_d^{-\beta} F) \frac{dp_n}{p_n} + \right. \\ &\quad \left. + (1 - \beta) p_d^{1-\beta} p_n^{-\alpha} \underline{d} \frac{dp_d}{p_d} + \beta p_n^{-\alpha} p_d^{-\beta} F \frac{dp_d}{p_d} \right] = \\ &= b \frac{\alpha^\alpha \beta^\beta \gamma^\gamma}{(\alpha + \beta + \gamma)^{\alpha + \beta + \gamma}} \left[ (\alpha + \beta) p_n^{-\alpha} p_d^{-\beta} F \right] \varepsilon > 0 \end{aligned}$$

Thus, if teacher quality is defined either via the ability to produce high-quality lessons,  $p_d$ , or as a vector of teacher abilities,  $(p_n, p_d)$ , the test score benefits of lesson use and teacher quality may decrease in teacher quality.

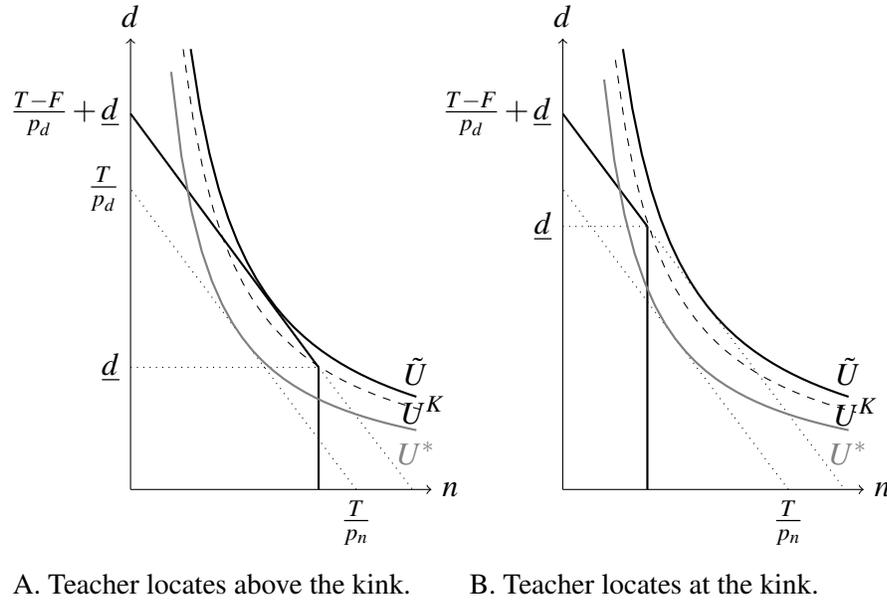
**Case 2:** An alternative definition of teacher quality is the ability to adopt new technology. In

principle, a more competent teacher should be able to adopt off-the-shelf lessons at a lower cost,  $F$ . Taking a derivative of (11) with respect to  $F$ , one gets:

$$\frac{\partial [b\tilde{n}^\alpha \tilde{d}^\beta - b(n^*)^\alpha (d^*)^\beta]}{\partial F} = -\frac{\alpha^\alpha \beta^\beta \gamma^\gamma}{(\alpha + \beta + \gamma)^{\alpha + \beta + \gamma}} \frac{b}{p_n^\alpha p_d^\beta} < 0$$

Here, more able teachers with lower  $F$  will be able to achieve a higher increase in test scores when adopting the technology. Therefore, if teacher quality is defined as the speed of adoption, as opposed to ‘prices’, then teacher quality may be associated with higher test score benefits from off-the-shelf lessons.

Figure E1. Illustration of the Model.



*Notes:* This figure depicts the effect of off-the-shelf lessons on test scores under a simplifying assumption  $\gamma = 0$ . This is a simplification of the teacher's problem that abstracts away from the decision of how much leisure to consume. However, it is helpful for illustrating the economic forces at play. For a rigorous and complete treatment of the teacher problem, see the analytic solutions presented in Section E2. As described above,  $T$  is the stock of time available to each teacher;  $d$  and  $n$  are the units of time spent on lesson planning and other tasks, respectively;  $p_d$  and  $p_n$  denote teacher (in)effectiveness in lesson planning and other tasks;  $F$  is the fixed time cost of adopting off-the-shelf lessons;  $\underline{d}$  is lesson quality guaranteed by off-the-shelf lessons; and  $U$ 's are the indifference curves fixed at a certain level of average students' test score and a certain level of leisure. Specifically,  $U^*$  is the level of utility achieved by a teacher without the lessons;  $\tilde{U}$  is the hypothetical optimal level of utility with the lessons without the  $d \geq \underline{d}$  restriction; and finally,  $U^K$  is the maximal utility level achieved at the kink of the budget line where  $d = \underline{d}$ . The original budgets constraint without lessons is depicted by the dashed line connecting  $T/p_d$  and  $T/p_n$ . The feasible portion of the budget constraint with lesson use is depicted by the solid line that goes from  $T - F/p_d + \underline{d}$  and hits  $\underline{d}$ . The dashed continuation of this line is the infeasible portion of the budget constraint with lesson use without the  $d > \underline{d}$  restriction.

## Appendix F. Spillovers.

Table F1. Spillovers.

	2014 Standardized Math Score	2014 Standardized Math Score	2014 Standardized Math Score	2014 Standardized Math Score
	(1)	(2)	(3)	(4)
License Only	0.060* [0.033]	0.051 [0.033]	0.066* [0.034]	0.062* [0.032]
Full Treatment	0.100*** [0.035]	0.086** [0.037]	0.097** [0.038]	0.078** [0.039]
% License Only in School	0.111 [0.086]		0.141 [0.098]	
% Fully Treated in School	0.112 [0.092]		0.184* [0.103]	
District FE x Requested	Y	Y	Y	Y
District FE x Teacher-Level Lagged Test Scores	Y	Y	Y	Y
District FE x Individual Lagged Test Scores	N	N	Y	Y
School FE	N	Y	N	Y
All controls	Y	Y	Y	Y
Joint p-value for peer effects	0.384	N/A	0.536	N/A
Observations	363	363	27,613	27,613
Unit of Observation	Teacher	Teacher	Student	Student

*Notes:* \*\*\* - significance at less than 1%; \*\* - significance at 5%; \* - significance at 10%. Standard errors clustered at the teacher level are reported in square brackets. All specifications include controls for the requested indicator, average teacher-level 2013 math and reading test scores, and teacher-level shares of students with missing 2013 math and reading test scores – all interacted with district fixed effects. 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. Columns (2) and (4) include school-level fixed effects. Specifications in Columns (3) and (4) control for individual-level 2013 math and reading test scores. So that we can include all students with math scores in 2014 in regression models, students with missing 2013 standardized math and reading scores were given an imputed score of zero. To account for this in regression models, we also include indicators denoting these individuals in all specifications. Results are robust to restricting the sample to students with complete data. In the absence of exam type data for Hanover, test scores for that district were standardized by grade.

## Appendix G. Effect Heterogeneity by Teacher Quality.

One of the methodological innovations of paper is to present a way to test for treatment heterogeneity by teacher quality with a single year of value-added data. To motivate our strategy, we start out with the standard teacher value-added model as presented in [Jackson, Rockoff and Staiger \(2014\)](#).<sup>38</sup> We show that marginal effects in this standard value-added model, when aggregated up to the teacher level, yield a very intuitive interpretation in a conditional quantile regression model. Specifically, we show that when average student test scores (aggregated at the teacher level) is the dependent variable, the estimated coefficient of a randomized treatment using conditional quantile regression at quantile  $\tau$  is the estimated effect of that treatment on teachers at the  $\tau$ th percentile of the teacher quality distribution. We then present a Monte Carlo simulation showing that our method is valid.

### *The Standard Value Added Model*

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

$$Y_{it} = \mathbf{X}_{it}\delta + \mu_t + \theta_c + \varepsilon_{it}$$

Here  $Y_{it}$  is student  $i$ 's test score, where student  $i$  is being taught by teacher  $t$ ,  $\mathbf{X}_{it}$  is a matrix of observable student covariates,  $\varepsilon_{it}$  is the idiosyncratic student-level error,  $\theta_c$  is the idiosyncratic classroom-level error, and, finally,  $\mu_t$  is the teacher  $t$ 's fixed effect or value added. That is, a teacher's value added is the average increase (relative to baseline) in student test scores caused by the teacher.

Having laid out the standard value-added model, let us aggregate this model to the teacher level by taking averages. This results in the equation below:

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

Now we propose that the randomized treatment ( $T_t$ ) has a causal effect on each teachers value-added. Specifically, we propose that:

$$\mu_t = \beta T_t + v_t$$

, where  $\beta$  is the influence of Mathalicious lessons on teacher  $t$ 's value added, while  $v_t$  is the teacher fixed effect (or value added) before introducing the treatment. The full aggregated model is now:

$$\bar{Y}_t = \beta T_t + \bar{\mathbf{X}}_t\delta + v_t + \theta_c + \bar{\varepsilon}_t \quad (12)$$

As shown in (12), in a regression of average student test scores (aggregated to the teacher level) on treatment status and covariates, the unobserved error term is  $v_t + \theta_c + \bar{\varepsilon}_t$ . Accordingly, the residual from this regression is the teacher value added (without the treatment) plus noise (random classroom-level errors and aggregate student-level sampling variability). Following [Chetty et al. \(2011\)](#), we consider this teacher-level residual a noisy measure teacher quality (without the treatment). Accordingly, we refer to this teacher-level residual as teacher value added.

---

<sup>38</sup>We suppress the time subscript, as there is no time dimension in our application.

### *Applying The Conditional Quantile Function*

The notation above implicitly assumes that the marginal effects of the treatment and the covariates were the same for all teachers. We now explicitly allow for the possibility that the treatment effect varies by teacher value added (as defined above). Using the nomenclature from [Koenker and Hallock \(2001\)](#), the conditional quantile ( $\tau$ ) is the  $\tau$ th quantile of the conditional distribution of the response variable. More simply, ( $\tau$ ) is the  $\tau$ th quantile of the distribution of the residual. As discussed above, the residual from (12) is precisely our measure of teacher value added. The conditional quantile estimator as introduced by [Koenker and Bassett \(1978\)](#) estimates the marginal effect of a treatment at different points of the conditional distribution of an outcome. We show below that this model, applied to the aggregate value-added model as in (12), yields consistent estimates of the marginal effect of the treatment for teacher at different points in the value-added distribution.

Allowing for the possibility that  $\beta$  and  $\delta$  may vary with the quantile  $\tau$ , let us apply the conditional quantile function to (12). This yields (13) below:

$$Q_\tau[\bar{y}_t|T, \bar{\mathbf{X}}] = \beta(\tau)T_t + \bar{\mathbf{X}}_t\delta(\tau) + Q_\tau[v_t(\tau) + \theta_c(\tau) + \bar{\epsilon}_t(\tau)|T, \bar{\mathbf{X}}] \quad (13)$$

Because the treatment was randomized across teachers,  $T_t$  is independent of all other random variables in the model, i.e.  $T_t \perp \{\bar{\mathbf{X}}_t, v_t, \theta_c, \bar{\epsilon}_t\}$ . In our setting, the conditional quantile regression formalized in [Koenker and Bassett \(1978\)](#) for conditional quantile  $\tau$ , solves for the  $\hat{\beta}(\tau)$  and the  $\hat{\delta}(\tau)$  that minimize

$$[\hat{\beta}(\tau), \hat{\delta}(\tau)] = \min_{b,d} \sum_{t=1}^T \rho_\tau[\bar{y}_t - bT_t - \bar{\mathbf{X}}_t d] \quad (14)$$

, where  $\rho_\tau = u_t[\tau - \mathbb{1}(u_t < 0)]$  is a re-weighting function of the residuals  $u_t = v_t + \theta_c + \bar{\epsilon}_t$ . As demonstrated in [Buchinsky \(1998\)](#), if  $Q_\tau[v_t(\tau) + \theta_c(\tau) + \bar{\epsilon}_t(\tau)|T, \bar{\mathbf{X}}] = 0$  for each quantile  $\tau$ , the conditional quantile regression coefficient  $\hat{\beta}(\tau)$  is a consistent estimate of  $\beta(\tau)$  in (13).<sup>39</sup>

To conclude, the conditional quantile regression model applied to teacher-level aggregate data provides marginal effect estimates for our randomized treatments at particular quantiles of the distribution of the residual, which in our case can be interpreted as teacher value-added (plus noise).

### *Mote Carlo Simulation*

Because the presentation above is somewhat theoretical, to provide concrete evidence that our procedure works, we assigned random treatments to the teachers in our data,<sup>40</sup> created simulated causal effects that varied based on each teacher’s residual,<sup>41</sup> and then estimated the conditional quantile model at each quantile. We ran this simulation with 1,000 random draws and plot the distribution of estimated causal effects for each quartile of the residual distribution between the

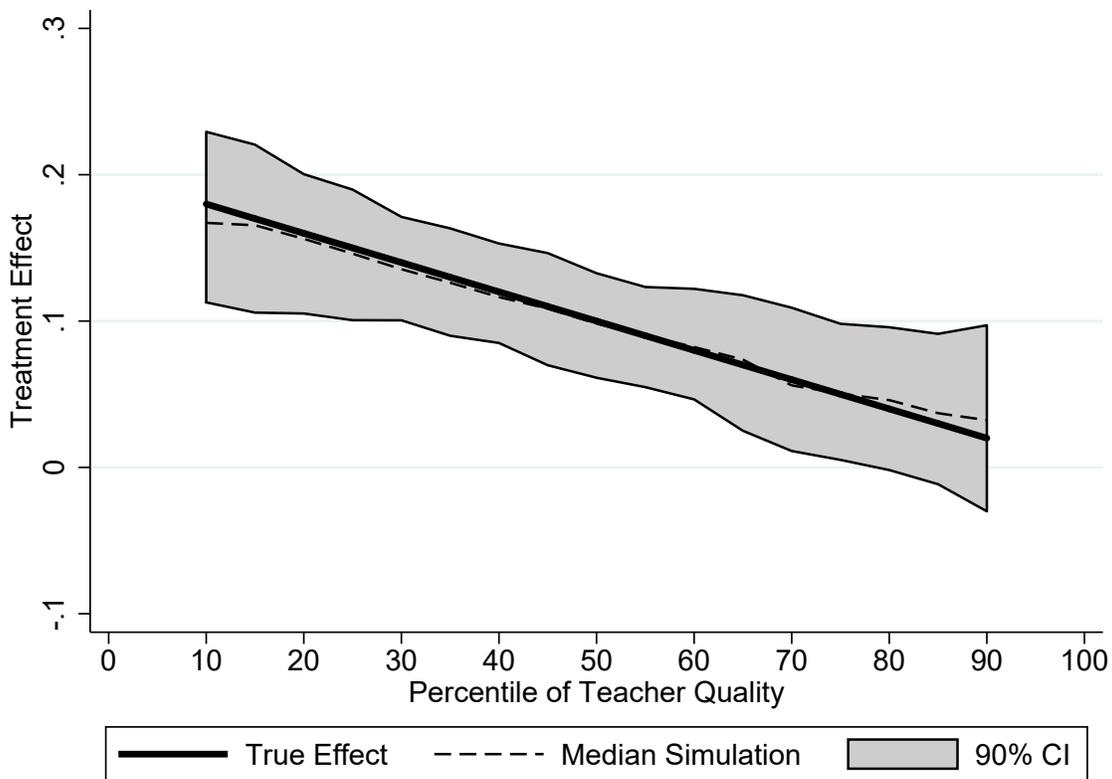
<sup>39</sup>This is a standard assumption in the quantile regression literature. For a reference, see e.g. [Buchinsky \(1998\)](#)

<sup>40</sup>We retained the same distribution of treatments in the data. To do this we randomly “reassigned” teachers to the actual treatments in the data to create “simulated” treatment assignments. We then created simulated treatment effects based on these assignments.

<sup>41</sup>To obtain a measure of each teacher’s residual we estimate the main test score regression model in (1) without treatment indicators and stored each teacher’s residual. We then created a simulated treatment effect, or benefit, for all teachers that received a simulated treatment assignment. This benefit is a linear function of the teachers residual. Specifically, benefit= 0.002\*(100-percentile), where percentiles the percentile of the teachers residual and goes from 0 to 100.

10th and the 90th in increments of 5. If our procedure is valid, the distribution of estimated effects at each quartile should be centered on the real effect (as defined by the simulation). Figure G1 plots the real simulated effects at each quartile and also the 5th and 95th percentiles of the distribution of estimated effects by quartile. As one can see, the distribution of the estimates using our procedure is largely centered on the real effect. To provide a more formal test of this, the average deviation from the real effect across all 17 quartile estimates and 1,000 replications is  $-0.00016$ , and the test that this is equal to zero cannot be rejected at the 10 percent significance level. In sum, the simulation data indicate that our approach (with a randomly assigned treatment) yields consistent causal estimates of the treatment at each percentile of the teacher quality distribution.

Figure G1. Simulation Results.



*Notes:* The solid black line represents the simulated treatment effect that was artificially created to equal 0.18 at the 10th quantile of the teacher quality distribution and decrease by 0.01 each with each extra 5th quantile. Teacher quality is estimated as residuals from model (1). The dash black line displays the median treatment effect being evaluated at different quantiles of teacher quality using conditional quantile regression formally described in Appendix G. The shaded area depicts the empirical 90% confidence interval for each quartile calculated as the area between the 50th and 950th largest estimate obtained after 1,000 simulations.

## 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	1.669 [2.087]	4.291** [2.072]	0.017 [0.034]	0.055* [0.032]	2.096 [5.874]	0.015 [0.022]
Full Treatment	8.401*** [2.431]	7.905*** [2.234]	0.093** [0.039]	0.093*** [0.035]	1.637 [3.826]	0.003 [0.024]
District FE x Requested	Y	Y	Y	Y	Y	Y
District FE x Lagged Test Scores	Y	Y	Y	Y	Y	Y
All controls	N	Y	N	Y	Y	Y
Observations	363	363	363	363	363	363
Unit of Observation	Teacher	Teacher	Teacher	Teacher	Teacher	Teacher

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

## Appendix I. Heterogeneous Effects by Teacher Experience.

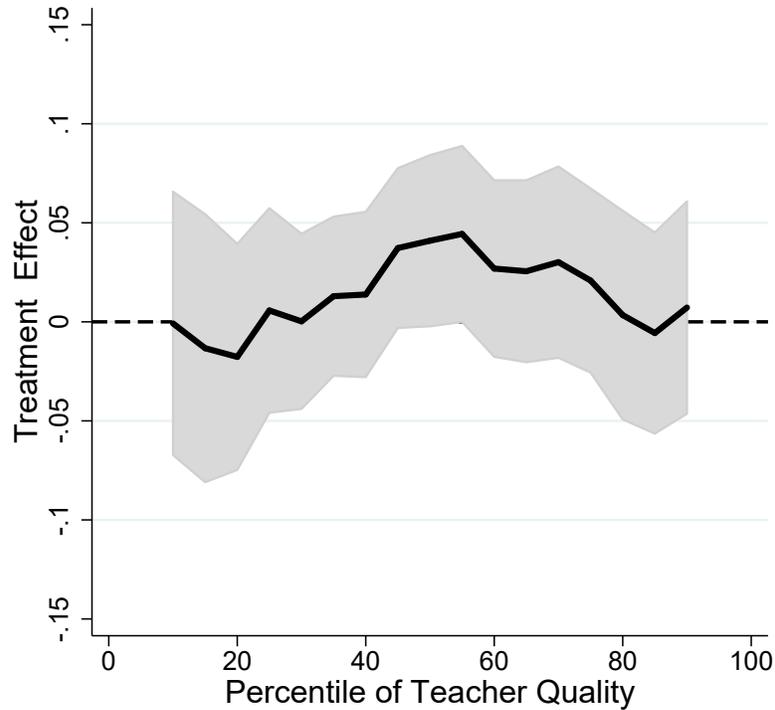
Table I1. Heterogeneous Effects by Teacher Experience.

	2014 Standardized Math Score	2014 Standardized Math Score	2014 Standardized Math Score	2014 Standardized Math Score
	(1)	(2)	(3)	(4)
License Only	0.045 [0.052]	0.047 [0.033]	0.055 [0.054]	0.049 [0.034]
Full Treatment	0.024 [0.052]	0.090** [0.035]	0.025 [0.057]	0.081** [0.038]
License Only x Years of Experience	0.001 [0.004]		0.000 [0.004]	
Full Treatment x Years of Experience	0.006 [0.004]		0.005 [0.004]	
License Only x First/Second Year Teachers		0.252** [0.125]		0.276** [0.135]
Full Treatment x First/Second Year Teachers		0.079 [0.104]		0.056 [0.097]
District FE x Requested	Y	Y	Y	Y
District FE x Teacher-Level Lagged Test Scores	Y	Y	Y	Y
District FE x Individual Lagged Test Scores	N	N	Y	Y
All controls	Y	Y	Y	Y
Joint p-value for Treatment x Experience Var	0.275	0.126	0.404	0.124
Observations	363	363	27,613	27,613
Unit of Observation	Teacher	Teacher	Student	Student

*Notes:* \*\*\* - significance at less than 1%; \*\* - significance at 5%; \* - significance at 10%. Standard errors clustered at the teacher level are reported in square brackets. All specifications include controls for the requested indicator, average teacher-level 2013 math and reading test scores, and teacher-level shares of students with missing 2013 math and reading test scores – all interacted with district fixed effects. 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. Specifications in Columns (3) and (4) control for individual-level 2013 math and reading test scores. So that we can include all students with math scores in 2014 in regression models, students with missing 2013 standardized math and reading scores were given an imputed score of zero. To account for this in regression models, we also include indicators denoting these individuals in all specifications. Results are robust to restricting the sample to students with complete data. In the absence of exam type data for Hanover, test scores for that district were standardized by grade.

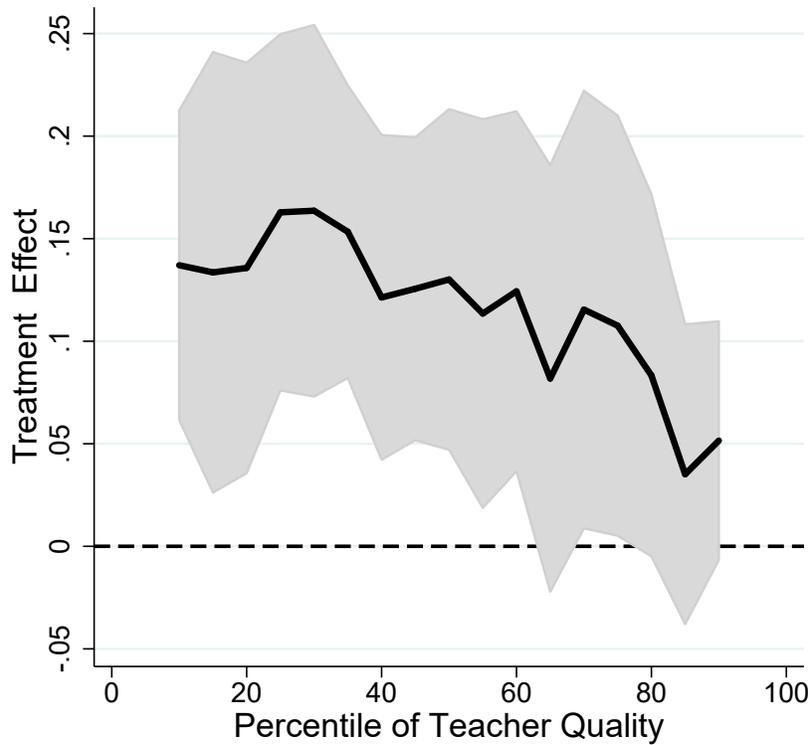
## Appendix J. Quantile Regression: Robustness Checks.

Figure J1. 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](#).

Figure J2. Marginal Effect of the Full Treatment by Classroom Quality.  
Excluding Requested Teachers.



*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 Math test scores serve as the main outcome. All specifications exclude teachers with a requested status. The shaded area depicts the 90% confidence interval for each regression estimate. For a formal discussion of the method, see [Appendix G](#).

## Appendix K. Survey Response and Lesson Downloads.

Table K1. Survey Response and Lessons Downloads.

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

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

## Appendix L. Auxiliary Results on Lesson Use.

Table L1. Effects on Lesson Use Calculated Based on Complete Data.

Panel A: Subsample of Teachers Who Answered Both Mid-Year and End-of-Year Surveys (~20%).

	Lessons Looked	Lessons Taught	Lessons Downloaded	Webinars Viewed
	(1)	(2)	(3)	(4)
License Only	1.404 [5.018]	0.092 [1.650]	1.969 [4.178]	0.168 [0.175]
Full Treatment	5.103 [5.021]	2.284 [1.912]	3.699 [4.225]	0.499** [0.231]
District FE x Requested	Y	Y	Y	Y
All controls	Y	Y	Y	Y
Observations	69	69	69	69

Panel B: Subsample of Teachers Who Answered either Mid-Year or End-of-Year Survey (~60%).

	Lessons Looked	Lessons Taught	Lessons Downloaded	Webinars Viewed
	(5)	(6)	(7)	(8)
License Only	1.396** [0.700]	0.466 [0.407]	1.034** [0.490]	-0.027 [0.018]
Full Treatment	2.618*** [0.720]	0.983** [0.390]	2.134*** [0.588]	0.097** [0.041]
District FE x Requested	Y	Y	Y	Y
All controls	Y	Y	Y	Y
Observations	236	236	236	236

Notes: \*\*\* - significance at less than 1%; \*\* - significance at 5%; \* - significance at 10%. Robust standard errors are reported in square brackets. All specifications include controls for the requested indicator, average teacher-level 2013 math and reading test scores, and teacher-level shares of students with missing 2013 math and reading test scores – all interacted with district fixed effects. Additional controls include teachers' education level, years of experience, sex, race, grade fixed effects, as well as the percentage of male, black, white, Asian, and Hispanic students in their class. The data on lessons downloaded and webinars watched are available for all 363 teachers. The number of lessons taught or read was missing for some teachers because of survey non-response: 69 teachers completed both mid-year and end-of-year surveys, 236 teachers completed either of the two. Panel A restricts the sample to 69 teachers who completed both surveys. Panel B restricts the sample to 236 teachers who completed either survey.

Table L2. Effects on Lesson Use by Requested Status.

Panel A: Multiple Imputation Estimates by Requested Status. Missing Outcome Data Imputed Using Multiple Imputation.				
	Lessons Looked	Lessons Taught	Lessons Downloaded	Webinars Viewed
	(1)	(2)	(3)	(4)
License Only	1.914*** [0.519]	0.653*** [0.185]		
Full Treatment	3.229*** [0.621]	2.100*** [0.461]	N/A	N/A
License Only x Requested	-0.786 [0.936]	-0.073 [0.334]		
Full Treatment x Requested	2.348* [1.246]	-0.443 [0.594]		
District FE x Requested	Y	Y		
All controls	Y	Y		
Observations	363	363		

Panel B: Full Sample Estimates by Requested Status. Missing Data for Lessons Looked and Taught Replaced with Zero (Lower Bound).				
	Lessons Looked	Lessons Taught	Lessons Downloaded	Webinars Viewed
	(5)	(6)	(7)	(8)
License Only	1.545*** [0.553]	0.383 [0.290]	1.249*** [0.421]	-0.007 [0.007]
Full Treatment	1.125** [0.506]	0.551* [0.333]	0.740* [0.417]	0.006 [0.011]
License Only x Requested	-1.222 [0.907]	-0.453 [0.440]	-0.834 [0.806]	-0.008 [0.014]
Full Treatment x Requested	1.987* [1.036]	-0.109 [0.468]	2.227** [0.953]	0.089* [0.049]
District FE x Requested	Y	Y	Y	Y
All controls	Y	Y	Y	Y
Joint p-value for Treatment x Requested	0.026	0.588	0.014	0.179
Observations	363	363	363	363

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

## Appendix M. Auxiliary Results on Student Surveys.

Table M1. Students' Post-Treatment Survey Analysis Without Controls (Chesterfield and Hanover only).

	Share of Completed Surveys	Standardized Factors					
		Math has Real Life Application	Increased Interest in Math Class	Increased Effort in Math Class	Increased Motivation for Studying in General	Math Teacher Promotes Deeper Understanding	Math Teacher Gives Individual Attention
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
License Only	-0.052 [0.083]	-0.017 [0.072]	-0.030 [0.075]	0.010 [0.046]	-0.021 [0.053]	0.052 [0.076]	0.085 [0.078]
Full Treatment	-0.036 [0.095]	0.158** [0.076]	0.058 [0.074]	0.030 [0.045]	0.036 [0.050]	0.204** [0.081]	0.187*** [0.072]
End-of-Year Indicator	Y	Y	Y	Y	Y	Y	Y
District FE x Requested	N	N	N	N	N	N	N
All controls	N	N	N	N	N	N	N
Observations	27,450	18,013	17,855	18,010	17,822	17,899	18,503

Notes: \*\*\* - significance at less than 1%; \*\* - significance at 5%, \* - significance at 10%. Standard errors clustered at the teacher level are reported in square brackets. For details on the estimating strategy, see (3). Each outcome, except for the share of completed surveys, is a result of factor analysis and encompasses variation from several individual questions. For details on how the factors were formed, see Appendix C. The specification do not contain any covariates other than the treatment and end-of-year indicators. The fact that the survey was anonymous prevented us from including any student-level covariates. The regressions presented in Column (1) are estimated at the teacher level. The share of completed surveys for each teacher was calculated by comparing the number of completed student surveys with the number of students with complete data on math test scores.

## Appendix N. Instrumental Variables Estimation.

As an additional test of whether lesson use is indeed responsible for an increase in math scores, we estimate instrumental variables regressions of test scores against lesson use using indicators for the six treatments as instruments. Note that we impute lesson use for those with missing or incomplete use data. The results are presented in [Table N1](#). Looking at the student level regression (Column 2), the instrumental variable coefficient on lessons taught is  $0.033\sigma$  and is statistically significant at the 5 percent level. The effects are similar at the teacher level (Column 4). Note that in both these models the first stage F-statistic is above 10. In our placebo tests, the effects for English scores are very close to zero and are not statistically significant (Columns 8). To directly test for the possibility that the additional supports may have a positive effect irrespective of lesson use, we estimate the same instrumental variables regression while controlling for receiving the full treatment. In such models (Column 3 and 6), conditional on lesson use, the coefficient on the full treatment dummy is negative and not statistically significant, while the coefficient on lesson use is slightly larger (albeit no longer statistically significant due to larger standard errors). This is very similar to the results based on comparisons across the different treatments. Overall the patterns presented are inconsistent with the benefits being due to the extra supports, and provide compelling evidence that all of our effects are driven by the increased lesson use itself.

Table N1. Instrumental Variables (IV) Estimation with Lessons Taught as an Endogenous Variable.

	Mathematics								Falsification: English	
	2014 Standardized Score									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Lessons Taught	0.010 [0.006]	0.038** [0.018]	0.033** [0.015]	0.039 [0.033]	0.011* [0.006]	0.044** [0.018]	0.039** [0.016]	0.032 [0.031]	0.002 [0.010]	0.004 [0.008]
Full Treatment				-0.014 [0.076]				0.018 [0.071]		
District FE x Requested	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
District FE x Teacher-Level Lagged Test Scores	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
District FE x Individual Lagged Test Scores	Y	Y	Y	Y	N	N	N	N	Y	Y
All controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	27,613	27,613	27,613	27,613	363	363	363	363	25,038	25,038
Estimation method	OLS	2SLS	2SLS	2SLS	OLS	2SLS	2SLS	2SLS	2SLS	2SLS
First Stage F-stat	-	23.84	41.87	4.607	-	15.51	16.69	3.252	20.94	46.52
Unit of Observation	Student	Student	Student	Student	Teacher	Teacher	Teacher	Teacher	Student	Student
Instruments	-	Treatment	Treatment X District	Treatment X District	-	Treatment	Treatment X District	Treatment X District	Treatment	Treatment X District

Notes: \*\*\* - significance at less than 1%; \*\* - significance at 5%, \* - significance at 10%. Standard errors clustered at the teacher level are reported in square brackets. Columns (1) and (5) report the results of OLS estimation, while Columns (2)-(4) and (6)-(10) contain the results of 2SLS estimation where the number of Mathalicious lessons taught is instrumented by the treatment status. All specifications include controls for the requested indicator, average teacher-level 2013 math and reading test scores, and a teacher-level shares of students with missing 2013 math and reading test scores - all interacted with district fixed effects. Additional controls include teachers' education level, years of experience, sex, race, grade fixed effects, as well as the percentage of male, black, white, Asian, and Hispanic students in their class. In addition, the student-level specifications in Columns (1)-(4) and (9)-(10) control for individual-level math and reading test scores and all student level demographics. Standardized test scores refer to the raw test scores standardized by exam type. In the absence of exam type data for Hanover, test scores for that district were standardized by grade.

## Appendix O. Patterns of Lesson Use Over Time

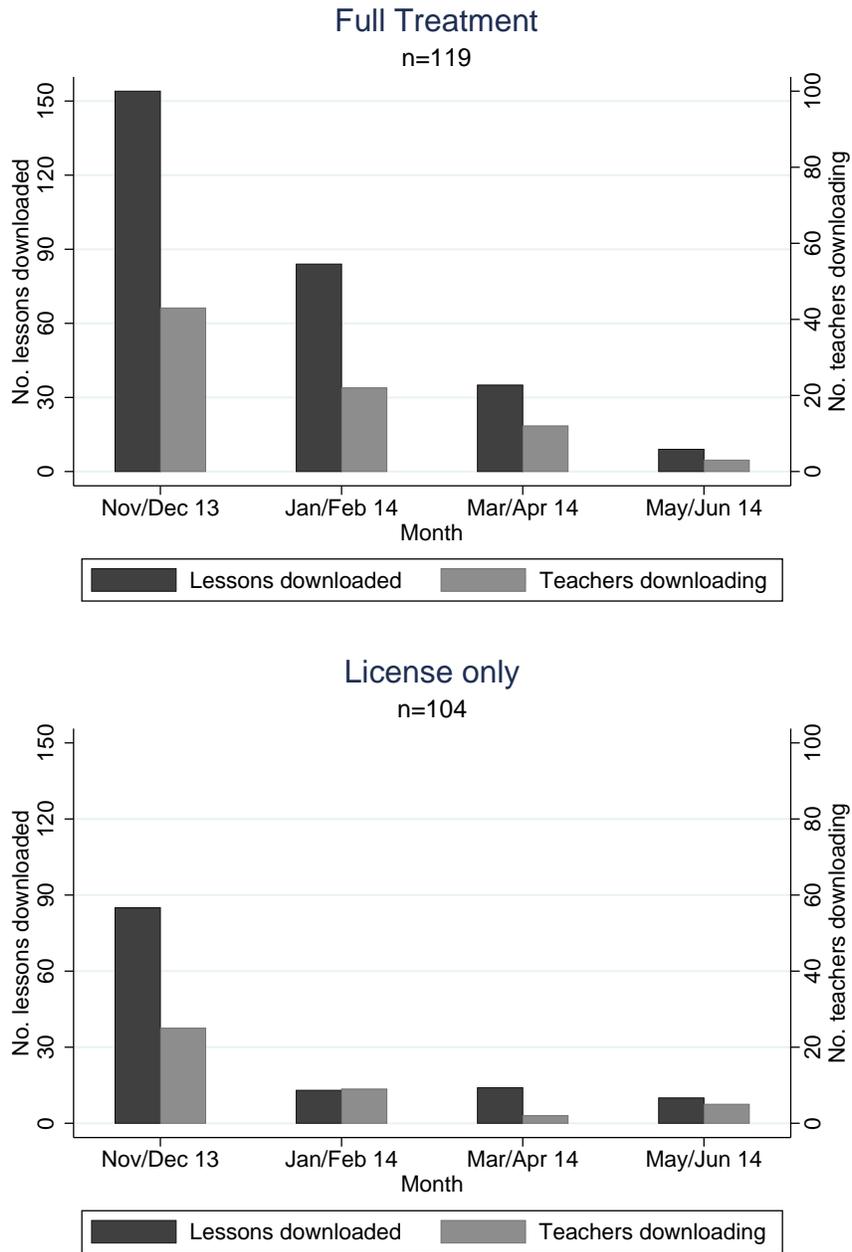
Given the sizable benefits to using the off-the-shelf lessons, one may wonder why lesson use was not even more widespread. To gain a sense of this, we present some graphical evidence of lesson use over time. [Figure O1](#) shows the number of lessons downloaded by license only and full treatment groups in different months. As expected, lesson use was much larger in the full treatment condition than that in the license only condition. However, [Figure O1](#) reveals a few other interesting patterns. There was a steady decline in the number of lessons downloaded over time within groups. While there were 97 downloads in the full treatment in November 2014, there were only 8 downloads in May 2015. Similarly, in the license only group, while there were 59 downloads in the November 2014, there were only 4 downloads in May 2015. To determine whether this decline is driven by the same number of teachers using Mathalicious less over time, or a decline in the number of teachers using Mathalicious over time, we also plot the number of teachers downloading lessons by treatment group over time. There is also a steady decline in the number of teachers downloading lessons so that the reduced use is driven by both reductions in downloads among teachers, and a reduction in the number of teachers downloading lessons over time.

Even though we have no dispositive evidence on why lesson use was not higher, or why lesson use dropped off over time, we speculate that it may have to do with behavioral biases and time management. The patterns of attrition from lesson downloads over time are remarkably similar to the patterns of attrition at online courses ([Koutropoulos et al., 2012](#)), gym attendance ([DellaVigna and Malmendier, 2006](#)), and fitness tracker use ([Ledger and McCaffrey, 2014](#)). Economists hypothesize that such behaviors may be due to individuals underestimating the odds that they will be impatient in the future and then procrastinate ([O'Donoghue and Rabin, 1999](#); [Duflo, Kremer and Robinson, 2011](#)). Similar patterns in [Figure O1](#) provide a reason to suspect that similar behaviors may be at play. In our context, these patterns may reflect teachers being optimistic about their willpower to use the lessons such that they started out strong, but when the time came, they procrastinated and did not make the time to implement them later on. However, it is also possible that as teachers use the lessons, they perceive that they are not helpful and decide to discontinue their use after downloading the first few lessons. Most of the empirical patterns support the former explanation. First, the rate of decay of lesson use is more rapid in the license only treatment than in the full treatment group. Specifically, without the additional supports to implement the lessons, the drop-off in lesson use was more rapid. In the full treatment group, downloads fell by about 45 percent between Nov/Dec and January/Feb, while it fell by over 80 percent during that same time period in the license only group. If the reason for the drop-off was low lesson quality, drop-off should have been similarly rapid for both groups. The second piece of evidence is that there is a sizable reduction in lessons downloaded in the full treatment condition after February when Mathalicious ceased sending out email reminders to teachers, while lesson use was stable in the license only condition. The third piece of evidence comes from surveys. We employed data from the end of year survey that asked treated teachers why they did not use off-the-shelf lessons more. Looking specifically at the question of whether the lessons were low quality, only 2 percent of teachers mentioned this was a major factor and almost 89% stated that it was not a factor at all. In sum, poor lesson quality does not explain the drop-off in lesson use, being reminded mattered, and the patterns of drop-off are very similar to other contexts in which behavioral biases played a key role – suggesting that procrastination is a plausible explanation.

The last piece of evidence to support the procrastination hypothesis also comes from the survey

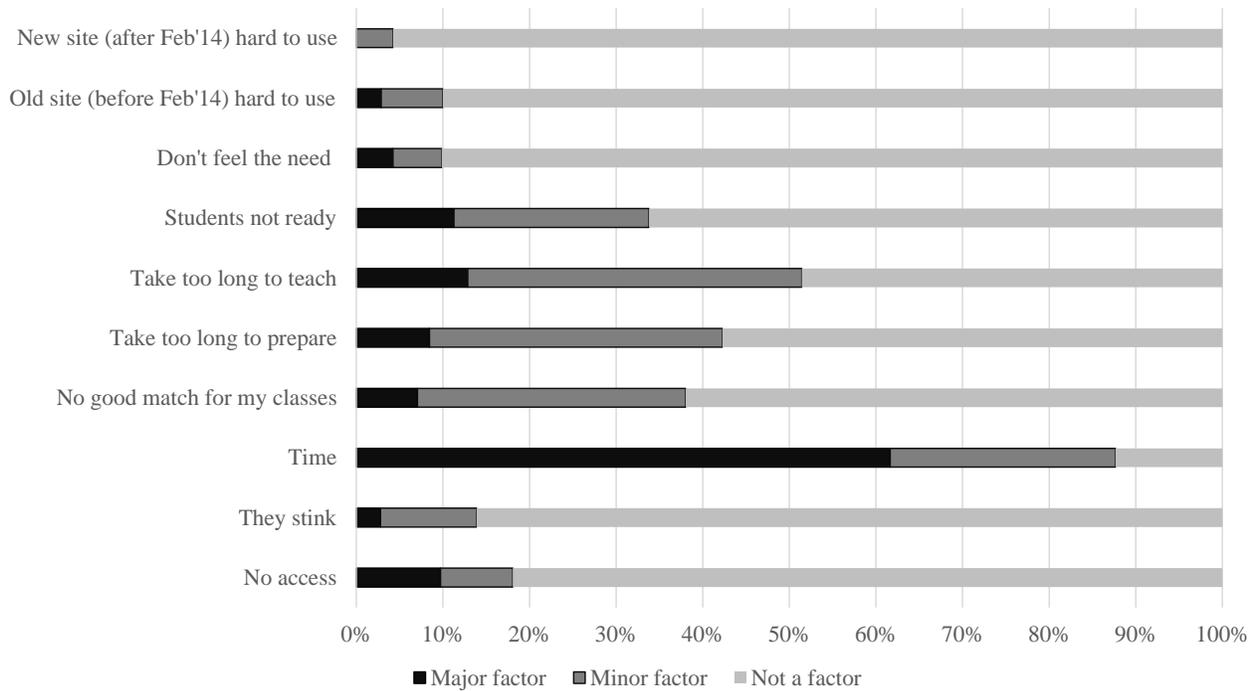
evidence shown in [Figure O2](#). The main reason cited for not using more lessons was a lack of time. Taken at face value, one might argue that the pressures on teacher time increased over the course of the year such that lesson use declined over time. However, this cannot explain the large differences in the trajectory of lesson use over time across the treatment arms. The explanation that best fits the observed patterns and the survey evidence is that, without the reminders and extra supports (i.e. Edmodo groups), teachers were unable to hold themselves to make the time to implement the lessons. The patterns also suggest that providing ways to reduce procrastination during the school year (such as sending constant reminders or providing some commitment mechanism) may be fruitful ways to increase lesson use. Other simple approaches may reduce the incentive to procrastinate at the moment by providing designated lesson planning time, or granting lesson access the summer before the school year when the demands on teachers' time may be lower.

Figure O1. 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 O2. Reasons for Lack of Mathalicious Lesson Use.  
License Only and Full Treatment Teachers Combined (n=71).



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

# Appendix P. Sample Mathalicious Lesson #1.

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

licensed under CC-BY-NC

## XBOX XPONENTIAL

How have video game console speeds changed over time?

lesson  
guide



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

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

### Primary Objectives

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

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

### Before Beginning...

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

### Preview & Guiding Questions

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



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

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

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

### Act One

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

### Act Two

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

## Act One: Moore Fast

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

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

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

### Explanation & Guiding Questions

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

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

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

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

### Deeper Understanding

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

## Appendix Q. Sample Mathalicious Lesson #2.

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

licensed under CC-BY-NC

### NEW-TRITIONAL INFO

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

lesson  
guide



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

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

#### Primary Objectives

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

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

#### Before Beginning...

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

**Preview & Guiding Questions**

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

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

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

**Act One**

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

**Act Two**

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

## Act One: Burn It

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

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



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

### Explanation & Guiding Questions

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

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

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

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

### Deeper Understanding

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

## Appendix References

- Buchinsky, Moshe.** 1998. "Recent advances in quantile regression models: a practical guideline for empirical research." *Journal of Human Resources*, 88–126.
- Chetty, Raj, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan.** 2011. "How does your kindergarten classroom affect your earnings? Evidence from Project STAR." *The Quarterly Journal of Economics*, 126(4): 1593–1660.
- Della Vigna, Stefano, and Ulrike Malmendier.** 2006. "Paying not to go to the gym." *The American Economic Review*, 694–719.
- Duflo, Esther, Michael Kremer, and Jonathan Robinson.** 2011. "Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya." *The American Economic Review*, 2350–2390.
- Jackson, C Kirabo, Jonah E Rockoff, and Douglas O Staiger.** 2014. "Teacher Effects and Teacher-Related Policies." *Annual Review of Economics*, 6(1): 801–825.
- Koenker, Roger, and Gilbert Bassett.** 1978. "Regression quantiles." *Econometrica*, 33–50.
- Koenker, Roger, and Kevin Hallock.** 2001. "Quantile regression: An introduction." *Journal of Economic Perspectives*, 15(4): 43–56.
- Koutropoulos, Apostolos, Michael Sean Gallagher, Sean C Abajian, Inge de Waard, Rebecca Joanne Hogue, Nilgun Ozdamar Keskin, and C Osvaldo Rodriguez.** 2012. "Emotive vocabulary in MOOCs: Context & participant retention." *European Journal of Open, Distance and E-Learning*, 15(1).
- Ledger, Dan, and Daniel McCaffrey.** 2014. "Inside wearables: How the science of human behavior change offers the secret to long-term engagement." *Endeavour Partners*.
- O'Donoghue, Ted, and Matthew Rabin.** 1999. "Doing it now or later." *American Economic Review*, 103–124.
- Rubin, Donald B.** 2004. *Multiple imputation for nonresponse in surveys*. Vol. 81, John Wiley & Sons.