American Economic Journal: Economic Policy 2018, 10(3): 226–254
https://doi.org/10.1257/pol.20170211

Can Online Off-the-Shelf Lessons Improve Student Outcomes? Evidence from a Field Experiment†

By Kirabo Jackson and Alexey Makarin*

Many websites now warehouse instructional materials designed to be taught by teachers in a traditional classroom. What are the potential benefits of the new resources? We analyze an experiment in which we randomly give middle school math teachers access to existing high-quality, off-the-shelf lessons, and in some cases, support to promote their use. Teachers receiving access alone increased students' math achievement by a marginally significant 0.06 of a standard deviation. Teachers who received access and support increased students' math achievement by 0.09 of a standard deviation. Weaker teachers experience larger gains, suggesting that these lessons substitute for teacher skill or efforts. The online materials are 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. (*JEL* C93, I21, J24, J45)

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 forthcoming). 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 such as designing and delivering lessons, managing the classroom, etc. While much research on teacher effectiveness has focused on how teachers deliver lessons more effectively (Pianta 2011, Taylor and Tyler 2012, Araujo et al. 2016), it stayed largely silent on the potentially important task of improving the lessons that teachers deliver.

Many lesson plans and instructional materials designed to be taught by teachers in a traditional classroom are now available online. One early site called Teachers Pay

* Jackson: School of Education and Social Policy, Northwestern University, 2120 Campus Drive, Room 204, Evanston, IL 60208 (email: kirabo-jackson@northwstern.edu); Makarin: Department of Economics, Northwestern University, 2211 Campus Drive, Room 3464, Evanston, IL 60208 (email: 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 (Grant B D 11107.R01). We're extremely grateful to Ginny Stuckey and Kate Novak at Mathalicious, Sarah Emmons of the University of Chicago Education Lab, and Tracy DellAngela 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.

† Go to https://doi.org/10.1257/pol.20170211 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

Teachers was launched in 2006 and allowed teachers to sell their lesson plans and instructional materials to other teachers. As of 2016, this site had an active membership of roughly four million users (more than all primary and secondary teachers in the United States). Other major players in this product space, such as LearnZillion, Pinterest, and Amazon Inspire, provide mostly free and openly licensed instructional materials (Madda 2016). There is considerable demand among teachers for these online resources. Opfer, Kaufman, and Thompson (2017) 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. These new resources have clearly altered how teachers plan and create their lessons.[1]

We present the first randomized experiment evaluating the benefits for students of providing teachers with access to, and support for, the use of online materials. In theory, lesson sharing websites create a public good 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 teachers do not possess themselves. Through these websites, one high-quality lesson can potentially improve the outcomes of millions of students. However, in practice, these lessons may have limited reach for two distinct reasons. First, many teachers may not use high quality lessons because of search costs associated with identifying high quality lessons or direct user fee costs for accessing content. Alternatively, these lessons may not be broadly applicable accross diverse classroom settings or may be costly to integrate into existing lesson plans. This paper tests between these hypotheses by removing these costs, identifying high-quality lessons and providing them to teachers free of charge to gauge the resulting benefits.

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 problems. Students must then memorize the definitions and practice the procedures (Stigler et al. 1999). 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 2006), the off-the-shelf lessons used in this study instead promote deep understanding, improve student engagement, and encourage retention of knowledge.[2] Each lesson is designed to be taught over two to five class periods, laying the foundation for between three and eight weeks of course material.

Under our experiment, teachers were randomly assigned to one of three treatment conditions. In the license-only condition, we informed teachers that these lessons were high quality and that they had free access to them. To promote lesson adoption, some teachers were randomly assigned to the full treatment condition in which teachers received email reminders to use the lessons and were invited to an online

[1] 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?"

[2] While there is 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 little experimental evidence on how promoting best practices among existing teachers impacts achievement tests.

social media group focused on lesson implementation (in addition to the license-only offerings). 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 and the full treatment groups experienced a 0.06σ and 0.09σ test score increase relative to those in the control condition, respectively. The full treatment effect is statistically significant at the 1 percent level, and has a similarly sized effect as that of moving from an average teacher to one at the eightieth percentile of quality, or reducing class size by 15 percent.[3] 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 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).

After estimating the base treatment effects, we explore the heterogeneity in, and mechanisms behind, the effects. We use conditional quantile regression models to estimate differential effects. Even though information technology is complementary to worker skill in many settings (e.g., Katz and Autor 1999; Akerman, Gaarder, and Mogstad 2015), the benefits of online lesson use are the largest for the least effective teachers (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. Recall that each lesson occupies between 2 and 5 days of class time, so that lesson use in the full treatment takes up about 7 days (roughly 5 percent of the school year prior to testing season[4]) and provides set up for about one-third of a years worth of material. Consistent with improved lesson quality and the aims of the intervention, treated students more frequently report that teachers emphasize deep learning, and that they feel that math has real life applications. The marginally significant test score gains in the license-only condition suggest that the improved outcomes in the full treatment condition are not driven solely by the additional supports but also by the increased lesson use. To provide more direct evidence of this, we show that the treatment arms

[3] 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).

[4] The school year is 180 days. However, state testing in Virginia begins in April. Accordingly, the number of instructional days reflected in the state test is about 140.

with the most lesson use also had the largest test score improvements, and, on average, the test score effects increase with lesson use.

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 2015).

Even though technology can change education in myriad ways, existing studies have focused on the effects of computer use among students (both inside and outside the classroom), and been silent on the potential role of *teacher's* use of technology outside the classroom to enhance their instruction inside the classroom.[5] The most closely related papers in this literature examine settings in which students interact with computers during class time so that software effectively replaces teacher instructional time (Barrow, Markman, and Rouse 2009, Taylor 2015). In contrast, we examine whether the dissemination of knowledge to *teachers* via technology can help teachers enhance their instructional time.[6]

Our findings also contribute to other related literatures. First, our intervention is a form of division of labor: teaching experts create instructional content, while classroom teachers focus on other tasks. Thus, 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.[7] Also, our work differs from studies on curriculum because we examine an approach that supplements existing curricula with online materials and lesson plans rather than changing the underlying curriculum per se. Finally, this work 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 I describes the intervention and outlines the experiment. Section II describes the data. Section III provides a

[5] For a recent review, see Bulman and Fairlie (2016). Angrist and Lavy (2002) provide first estimates of the effect of equipping classrooms with computers, while Beuermann et al. (2015) analyze the effect of giving laptops to students for home use. Several rigorous studies estimate the effect of specific educational software packages designed for in-class use (Rouse and Krueger 2004; Banerjee et al. 2007; Barrow, Markman, and Rouse 2009; Taylor 2015).

[6] In related work, Comi et al. (2017) find that effectiveness of technology at school depends on teachers' ability to incorporate it into their teaching practices.

[7] In one study with random assignment and a large sample of teachers, it has been shown that three particular math curricula are superior to the fourth one (Agodini et al. 2013).

stylized model, which is used to derive testable predictions. Section IV presents the empirical strategy and Section V describes the main results we obtained. Section VI explores the mechanisms, and Section VII concludes.

I. The Intervention

A. *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.[8] Lessons range from the simple to the more complex.

The lesson titled *Xbox Xponential* (see online 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 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.[9]

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

[8] 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.

[9] See the lesson titled *New-Tritional Info* in online Appendix Q for a less complex lesson.

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 five to seven 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 teachers 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 seven 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.

B. *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.[10] Importantly, participation in all components of the treatment was entirely voluntary.

TREATMENT CONDITION 2:

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

TREATMENT CONDITION 3:

Control Condition (*business-as-usual*). Teachers who were randomly assigned to the control condition continued business-as-usual. 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 (2017), over 90 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.[11] The Virginia

[10] 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).

[11] http://bit.ly/2u97XaZ.

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 (2017) 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 2017; Purcell et al. 2013). If there are nontrivial 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 VIB. 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.[12]

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 preselect certain regular classroom teachers that they wished to receive access to the Mathalicious licenses (i.e., receive either Treatment Condition 1 or Treatment Condition 2). We refer to these teachers as "requested" teachers. All requested teachers were identified and removed from the control condition. All of the remaining unrequested licenses in each district were allocated randomly to the remaining teachers.[13] 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 preselected 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.[14] Accordingly, all models condition on district and requested status and their interaction.[15] Moreover, our main results are robust to excluding the requested teachers.[16] Randomization ensured that conditional on

[12] We show in online Appendix F that the impacts of any spillovers on our estimates, if they exist, are negligible.

[13] 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.

[14] Table A1 of online Appendix A summarizes teacher participation by district, requested status, and treatment condition.

[15] This setup 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 setup, a regression with strata fixed effects and robust standard errors is also a valid specification.

[16] We present these results later in Section V. In online 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.

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.

II. 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 2013–2014 academic year (the year of the intervention). The teacher records included total years of teaching experience, gender, race, highest degree received, age, and years of teaching experience in the district. The administrative student records included grade level, gender, and race. Students were linked to their classroom teachers. These pretreatment 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 2012–2013 and 2013–2014. 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.[17] Reassuringly, like for all other incoming characteristics, this shows that incoming test scores are balanced across the three treatment conditions. Note that test scores in 2013 are similar between students in the control and full treatment groups (a difference of 0.04σ), but in 2014 are 0.163σ higher in the full treatment condition relative to the control condition.[18] The relative improvement in math scores over time is $0.163 - 0.04 = 0.123\sigma$ between the full treatment and the control group. By comparison, the relative improvement in English scores over time (where there should be no effect) between the full treatment and the control group is 0.003σ. These simple comparisons telegraph the more precise multiple regression estimates we present in Section V.

We use data from other sources to measure lesson use and to uncover underlying mechanisms. Each teacher was invited to answer two surveys: 22 percent and 61

[17] 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.

[18] 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.

Table 1—Summary Statistics

Variable	*N*	Mean	SD	Mean (control)	Mean (license only)	Mean (full treatment)	*p*-value for balance hypothesis (w/ district fixed effects and requested)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Teachers' characteristics							
Has MA degree	363	0.424	0.495	0.386	0.433	0.462	0.767
Has PhD degree	363	0.008	0.091	0.007	0.010	0.008	0.863
Teacher is female	363	0.802	0.399	0.793	0.769	0.840	0.852
Years teaching[a]	363	11.730	8.628	12.150	11.130	11.750	0.425
Teacher is white	363	0.884	0.320	0.879	0.865	0.908	0.622
Teacher is black	363	0.096	0.296	0.114	0.096	0.076	0.745
Grade 6	363	0.311	0.464	0.300	0.240	0.387	0.503
Grade 7	363	0.366	0.482	0.343	0.413	0.353	0.169
Grade 8	363	0.342	0.475	0.321	0.356	0.353	0.746
Participation across webinars	363	0.014	0.117	0.000	0.000	0.042	0.005
Total no. Mathalicious lessons the teacher taught	236[b]	0.818	2.123	0.275	0.750	1.519	0.053
Total no. Mathalicious lessons the teacher taught or read	236[b]	1.030	2.884	0.275	0.853	2.078	0.034
Total no. Mathalicious lessons the teacher downloaded	363	1.132	3.221	0.064	1.173	2.353	0.004
Total no. Mathalicious lessons the teacher looked at (downloaded, read, or taught)	256[c]	2.184	4.458	0.337	2.107	4.157	0.001
Students' chars. (*student level*)							
Student is male	27,613	0.516	0.074	0.515	0.519	0.513	0.798
Student is black	27,613	0.284	0.249	0.293	0.300	0.259	0.652
Student is white	27,613	0.541	0.261	0.534	0.535	0.553	0.588
Student is Asian	27,613	0.054	0.063	0.055	0.046	0.059	0.044
Student is Hispanic	27,613	0.083	0.078	0.081	0.078	0.089	0.395
Student is of other race	27,613	0.036	0.025	0.034	0.036	0.037	0.209
Math SOL scores, standardized by exam type, 2013	24,112[d]	0.0521	0.979	0.037	0.043	0.076	0.644
Math SOL scores, standardized by exam type, 2014	27,613	−0.002	1.001	−0.071	−0.021	0.092	0.887
Reading SOL scores, standardized by grade, 2013	24,878[d]	0.015	0.997	−0.010	−0.025	0.077	0.690
Reading SOL scores, standardized by grade, 2014	24,409[e]	0.008	0.997	−0.021	−0.027	0.068	0.969

[a]Using years in district for Henrico.

[b]The number of lessons taught and read were reported by teachers in the mid-year and end-of-year surveys. One hundred twenty-seven teachers did not take part in either of the surveys, hence the missing values.

[c]See (b) for an explanation of attrition. 20/127 teachers with missing values in (b) had nonzero values for the number of lessons downloaded.

[d]A small share of students have no recorded 2013 test scores. This is likely due to transfers into the district.

[e]Eighteen 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.

percent of teachers completed a midyear and an end-of-year survey, respectively.[19] 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

[19]Twenty percent of teachers completed both surveys and 61 percent of teachers completed either of them.

in Section VI. We supplement these data with the more objective measure of lessons downloaded. Specifically, for each lesson, we record whether it was downloaded for each teachers account using tracker data from the Mathalicious website. Based on both these data sources, our three measures of Mathalicious lesson use are (i) the number of lessons looked at, (ii) the number of lessons taught, and (iii) 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.[20] Survey questions were designed by the research team and Mathalicious to measure changes in factors hypothesized to be affected by the intervention (see for survey items). The student surveys were administered in the 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 VI.

III. 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 (i) guarantee a minimum level of lesson quality, and (ii) free up teacher time that would have been spent planning lessons, but (iii) 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.

[20] 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 online Appendix I.

To shed further light on this we model the teacher's problem (see online 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 nonobvious and testable results.

RESULT 1 (*The gains in average test scores from using the off-the-shelf lessons are nonnegative*): This is because (i) teachers adopt lessons if and only if the time savings are large enough to allow test scores to weakly increase, and (ii) 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 nonnegative*): 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.[21]

IV. 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:

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

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

[21] The online lessons produce a kink in the teachers 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.

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, α_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 π varies by district d. To improve precision,[22] 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 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 equation using OLS:

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

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.

V. Main Results

A. *Effects on Student Achievement in Mathematics*

The first result from the theoretical section is that the intervention effect on math scores should be nonnegative. 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 percent 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

[22] 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).

TABLE 2—EFFECTS ON STUDENT TEST SCORES

	Mathematics						Falsification: English	
	2014 raw score	2014 raw score	2014 standardized score	2014 standardized score	2014 standardized score	2014 standardized score	2014 raw score	2014 standardized score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. Baseline results								
License only	2.653	3.583	0.050	0.061	0.060	0.055	1.105	0.025
	[2.136]	[1.926]	[0.040]	[0.034]	[0.033]	[0.032]	[1.041]	[0.019]
Full treatment	7.899	7.057	0.105	0.094	0.086	0.093	0.460	0.008
	[2.662]	[2.308]	[0.046]	[0.038]	[0.038]	[0.035]	[1.223]	[0.022]
District FE × requested	Y	Y	Y	Y	Y	Y	Y	Y
District FE × teacher-level lagged test scores	Y	Y	Y	Y	Y	Y	Y	Y
District FE × 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
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Panel B. Baseline results without requested teachers								
License only	2.125	2.684	0.039	0.042	0.043	0.048	−0.688	−0.012
	[2.111]	[2.023]	[0.038]	[0.036]	[0.036]	[0.035]	[1.050]	[0.019]
Full treatment	9.382	8.714	0.124	0.108	0.101	0.117	1.880	0.030
	[2.904]	[2.692]	[0.046]	[0.045]	[0.044]	[0.043]	[1.450]	[0.026]
District FE × requested	Y	Y	Y	Y	Y	Y	Y	Y
District FE × teacher-level lagged test scores	Y	Y	Y	Y	Y	Y	Y	Y
District FE × 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: 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.

by 10.5 percent 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 percent 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 percent 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 percent 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 percent of a standard deviation relative to those in the control condition (p-value < 0.01). In sum, there is suggestive evidence of a positive effect of the license-only treatment (relative to the control condition) on student math scores of between 4 and 6 percent of a standard deviation, and a robust positive statistically significant effect of the full treatment of about 9 percent of a standard deviation.

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 online 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.

B. *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 pretreatment 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 online 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 pth percentile from the conditional quantile regression of equation (1) is the marginal effect of the treatment for teachers at the pth percentile of effectiveness. To verify this claim computationally, we implement a Monte Carlo simulation (see online Appendix G) and are able to consistently uncover treatment effects at different percentiles of the teacher quality distribution.[23]

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 tenth through ninetieth percentiles in intervals of five 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 nonlinear relationship, we fit a piece-wise linear function with a structural break at the sixtieth percentile. At and below the sixtieth quantile, the slope is 0.0003 and not statistically significant, while above the sixtieth quantile the slope is -0.00314 (p-value $= 0.001$).[24] In sum, for the bottom 60 percent of teachers, the

[23] Online 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.

[24] 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 standard errors 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

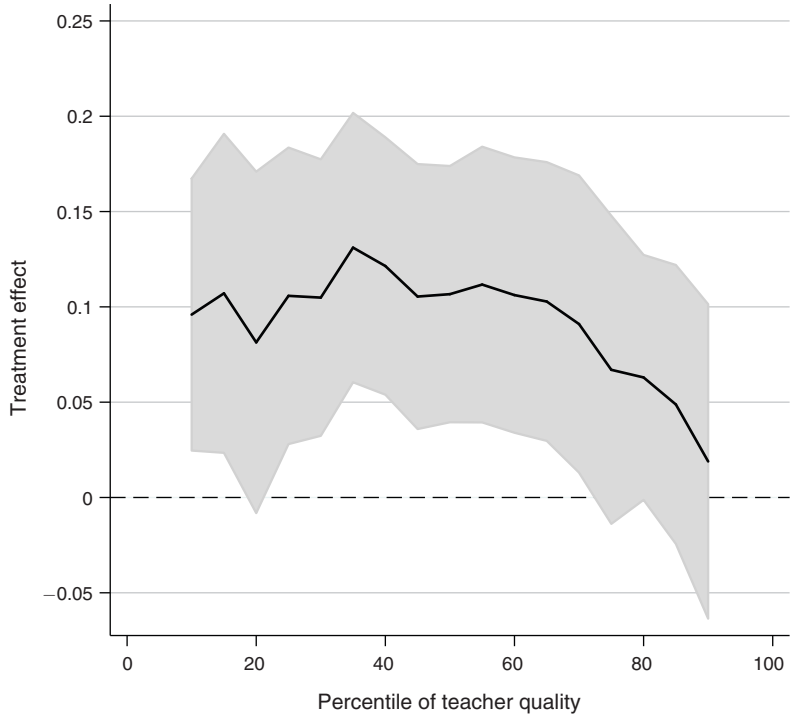

Figure 1. Marginal Effect of the Full Treatment by Teacher Quality. Mathematics Test Scores

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 percent confidence interval for each conditional quantile regression estimate. For a formal discussion of the method, see online 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.

marginal effect of the full treatment is roughly 0.11σ, and the full 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 ninety-ninth percentile of classroom quality, the semi-parametric point estimate is positive (albeit not statistically different from zero).[25] This is consistent

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 sixtieth percentile (-0.00314) lies below the fifth percentile of the distribution of placebo slopes.

[25] 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 online 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 0 at the 10 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.

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 online 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 0 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 (i) less effective and/or (ii) more likely to spend a lot of time planning lessons.

VI. Mechanisms

A. *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 midyear and end-of-year surveys and may suffer from bias due to survey nonresponse. 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 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.

Table 3—Effects on Lesson Use

	Lessons looked (1)	Lessons taught (2)	Lessons looked (3)	Lessons taught (4)
Panel A. Multiple imputation estimates. Missing outcome data imputed using multiple imputation				
License only	1.586	0.657	1.726	0.640
	[0.418]	[0.191]	[0.499]	[0.184]
Full treatment	4.404	1.925	3.058	2.017
	[0.605]	[0.282]	[0.594]	[0.467]
District FE × 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
	Lessons looked (5)	Lessons taught (6)	Lessons downloaded (7)	Webinars viewed (8)
Panel B. Full sample estimates. Missing data for lessons looked and taught replaced with zero (*lower bound*)				
License only	1.115	0.262	0.916	−0.013
	[0.422]	[0.221]	[0.328]	[0.009]
Full treatment	2.236	0.573	1.900	0.048
	[0.506]	[0.238]	[0.457]	[0.022]
District FE × requested	Y	Y	Y	Y
All controls	Y	Y	Y	Y
Observations	363	363	363	363

Notes: 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 nonresponse: 69 teachers completed both midyear 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.

We analyze the effect of the treatment on 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 20 percent of teachers who completed the surveys during both waves. We address this by using multiple imputation (Rubin 2004, Schafer 1997) to impute use for those individuals who did not complete the surveys.[26] 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 online Appendix K, where we show that conditional on treatment status, survey participation is unrelated to lessons downloaded, so that the imputation method is

[26] 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).

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 online 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 in 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.[27] 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.[28] 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.

To put these estimates in perspective, each Mathalicious lesson occupies about 3.5 days of class time (about 2.5 percent of the pretesting instructional days[29]) and provides intuition for topics that span between 3 and 8 weeks. As such, teachers in the full treatment report teaching lessons that take up about 7 classroom days (roughly 5 percent of the pretesting instructional days) and may impact about one-third of the school year. Accordingly, while the full treatment group never reached full fidelity with the Mathalicious model (which is between five and seven lessons

[27] As an additional check on our method, we compute lesson use based on the 60 percent of teachers that completed either the midyear 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 online 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.

[28] 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 midyear 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 that 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 online 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.

[29] While the academic calendar is 180 days. Testing begins in April so that there are roughly 140 pretesting instructional days. Accordingly, 3.5 classrooms days corresponds to about $3.5/140 = 2.5$ percent of the pretesting days.

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.

B. *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 six distinct factors.[30] Each factor is standardized to be mean zero, unit variance.

Teachers are only partially treated at the time of the midyear 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 midyear 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 midyear survey.[31] 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 observation that came from the end-of-year survey and zero otherwise:

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

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.[32]

[30] 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 factor. Factor loadings for each individual question are presented in online Appendix C.

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

[32] 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.

Table 4—Students' Early- and Posttreatment Survey Analysis (*Chesterfield and Hanover only*)

		Standardized factors					
	Share of completed surveys	Math has real life application	Increased interest in math class	Increased effort in math class	Increased motivation for studying in general	Math teacher promotes deeper understanding	Math teacher gives individual attention
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
License only	0.100	−0.012	−0.018	0.045	−0.021	0.001	0.033
	[0.082]	[0.060]	[0.062]	[0.035]	[0.035]	[0.065]	[0.063]
Full treatment	0.012	0.162	0.087	0.003	0.039	0.175	0.144
	[0.099]	[0.063]	[0.074]	[0.044]	[0.035]	[0.070]	[0.069]
End-of-year indicator	Y	Y	Y	Y	Y	Y	Y
District FE × 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: Standard errors clustered at the teacher level are reported in square brackets. For details on the estimating strategy, see equation (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 online 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.

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 nontrivial 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.[33]

The first factor measures whether students believe that math has real life applications. The results in column 2 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 0.162σ 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 nontrivial 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 motivation to study in general. None of these is directly targeted by the

[33] In fact, in the model with no controls (Table M1 in online 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 nonresponse.

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 (0.087σ). However, this effect is not statistically significant at traditional levels. The estimated coefficient on effort in math class is 0.045σ for the license-only condition 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.[34] The results support the premise of our model that teachers who used the lessons improved lesson quality. Students from the full treatment group are 0.175σ (p-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 0.144σ higher in the full treatment condition than in the control condition (p-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 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

[34] Jackson (2016) also uses more one-on-one time as a measure of teacher time.

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.[35]

C. *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 online 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 0.033σ. These patterns indicate that those treatments with larger effects on lesson use had larger test score gains suggesting that the reason the full treatments had a larger effect on test scores is, in part, because they had a larger effect on lesson use.

[35] 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 online Appendix D.

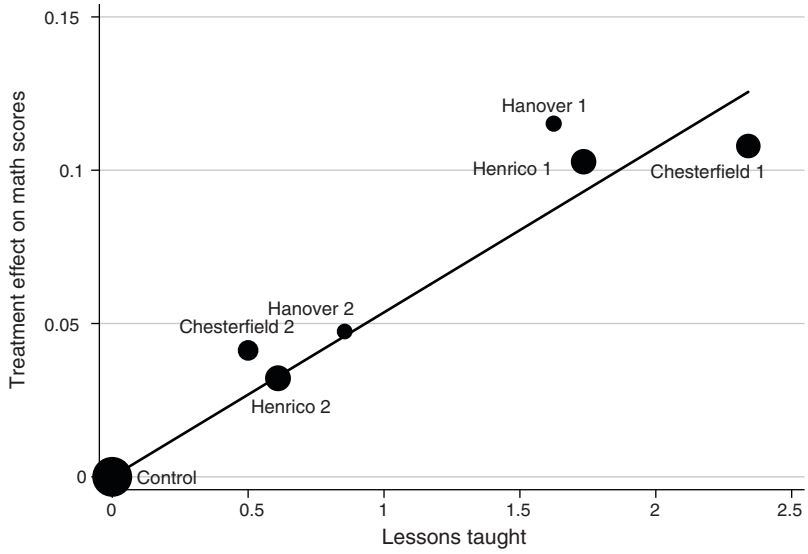

Figure 2. Estimated Effect on Math Test Scores by Estimated Effect on Lessons Taught

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 column 5 of Table 2. The *x*-axis displays coefficients for specifications similar to those estimated, 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 seven 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.

VII. 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 2017) 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 students 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 marginally statistically significant effects on lesson use and modest (but economically meaningful) effects on test scores (0.06σ). However, fully treated teachers (who also received online supports to promote lesson use) used the lessons more and improved their students test scores by about 0.09σ relative to teachers in the control condition. These positive effects are associated with 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 online Appendix O, we document that lesson use was moderate during the first couple of 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 2 percent of treated teachers mentioned that low quality 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 intervention is low cost. An upper bound estimate of the program cost (lessons and supports) is $431 per teacher.[36] Chetty, Friedman, and Rockoff (2014) estimate that a teacher who raises test scores by 0.14σ generates marginal gains of about $7,000 per student in present value future earnings. Using this estimate, the test score effect of about 0.09σ would generate roughly $4,500 in present value of future earnings per student.[37] 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

[36] 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.

[37] We assume that the full treatment effect is an average treatment effect (ATE) such that if all teachers were offered the full treatment the average effect would have been the same as our estimated full treatment effect. This assumption is supported by the similarity between the effects of the full treatment both with or without Requested teachers.

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 preexisting 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 may 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 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

Agodini, Roberto, Barbara Harris, Neil Seftor, Janine Remillard, and Melissa Thomas. 2013. *After Two Years, Three Elementary Math Curricula Outperform a Fourth*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences. Washington, DC, September.

Akerman, Anders, Ingvil Gaarder, and Magne Mogstad. 2015. "The Skill Complementarity of Broadband Internet." *Quarterly Journal of Economics* 130 (4): 1781–1824.

Angrist, Joshua, and Victor Lavy. 2002. "New Evidence on Classroom Computers and Pupil Learning." *Economic Journal* 112 (482): 735–65.

Araujo, M. Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady. 2016. "Teacher Quality and Learning Outcomes in Kindergarten." *Quarterly Journal of Economics* 131 (3): 1415–53.

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics* 122 (3): 1235–64.

Barrow, Lisa, Lisa Markman, and Cecilia Elena Rouse. 2009. "Technology's Edge: The Educational Benefits of Computer-Aided Instruction." *American Economic Journal: Economic Policy* 1 (1): 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." *Academy of Management Perspectives* 26 (1): 12–33.

Bonwell, Charles C., and James A. Eison. 1991. *Active Learning: Creating Excitement in the Classroom*. ASHE-ERIC Higher Education Report Series. Washington, DC: George Washington University.

Brown, John Seely, Allan Collins, and Paul Duguid. 1989. "Situated Cognition and the Culture of Learning." *Educational Researcher* 18 (1): 32–42.

Bugni, Federico A., Ivan A. Canay, and Azeem M. Shaikh. 2017. "Inference under Covariate-Adaptive Randomization with Multiple Treatments." Centre for Microdata Methods and Practice (CEMMAP) Working Paper 34/17.

Bulman, George, and Robert W. Fairlie. 2016. "Technology and Education: Computers, Software, and the Internet." National Bureau of Economic Research (NBER) Working Paper 22237.

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, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *Quarterly Journal of Economics* 126 (4): 1593–1660.

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): 2633–79.

Chingos, Matthew M., and Grover J. Whitehurst. 2012. *Choosing Blindly: Instructional Materials, Teacher Effectiveness, and the Common Core*. Brown Center on Education Policy at Brookings. Washington, DC, April.

Comi, Simona Lorena, Gianluca Argentin, Marco Gui, Federica Origo, and Laura Pagani. 2017. "Is it the way they use it? Teachers, ICT and student achievement." *Economics of Education Review* 56: 24–39.

Darling-Hammond, Linda, Ruth Chung Wei, Alethea Andree, Nikole Richardson, and Stelios Orphanos. 2009. *Professional Learning in the Learning Profession: A Status Report on Teacher Development in the United States and Abroad*. National Staff Development Council. Oxford, OH, February.

Deming, David. 2009. "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1 (3): 111–34.

Dostál, Jiří. 2015. *Inquiry-based instruction: Concept, essence, importance and contribution*. Moravia, Czech Republic: Palacký University, Olomouc.

Eichholtz, Piet, Nils Kok, and John M. Quigley. 2010. "Doing Well by Doing Good? Green Office Buildings." *American Economic Review* 100 (5): 2492–2509.

Fryer, Roland G., Jr. 2016. "The 'Pupil' Factory: Specialization and the Production of Human Capital in Schools." National Bureau of Economic Research (NBER) Working Paper 22205.

Hanushek, Eric A. 1974. "Efficient Estimators for Regressing Regression Coefficients." *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): 446–93.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization* 7 (Special Issue): 24–52.

Jackson, C. Kirabo. 2016. "The Effect of Single-Sex Education on Test Scores, School Completion, Arrests, and Teen Motherhood: Evidence from School Transitions." National Bureau of Economic Research (NBER) Working Paper 22222.

Jackson, C. Kirabo. Forthcoming. "What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test-Score Outcomes." *Journal of Political Economy*.

Jackson, Kirabo, and Alexey Makarin. 2018. "Can Online Off-the-Shelf Lessons Improve Student Outcomes? Evidence from a Field Experiment." *American Economic Journal: Economic Policy*. https://doi.org/10.1257/pol.20170211.

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." *Quarterly Journal of Economics* 131 (1): 157–218.

Jackson, C. Kirabo, Jonah E. Rockoff, and Douglas O. Staiger. 2014. "Teacher Effects and Teacher-Related Policies." *Annual Review of Economics* 6: 801–25.

Jackson, C. Kirabo, and Henry S. Schneider. 2015. "Checklists and Worker Behavior: A Field Experiment." *American Economic Journal: Applied Economics* 7 (4): 136–68.

Jacob, Brian A., and Jonah E. Rockoff. 2012. "Organizing Schools to Improve Student Achievement: Start Times, Grade Configurations, and Teacher Assignments." *Education Digest* 77 (8): 28–34.

Kane, Thomas J., Antoniya M. Owens, William H. Marinell, Daniel R. C. Thal, and Douglas O. Staiger. 2016. "Teaching Higher: Educators' Perspectives on Common Core Implementation." https://cepr.harvard.edu/files/cepr/files/teaching-higher-report.pdf.

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." National Bureau of Economic Research (NBER) Working Paper 14607.

Katz, Lawrence F., and David H. Autor. 1999. "Changes in the Wage Structure and Earnings Inequality." In *Handbook of Labor Economics*, Vol. 3A, edited by Orley Ashenfelter and David Card, 1463–1555. Amsterdam: North-Holland.

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): 1–22.

Koenker, Roger, and Gilbert Bassett, Jr. 1978. "Regression Quantiles." *Econometrica* 46 (1): 33–50.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114 (2): 497–532.

Lave, Jean, and Etienne Wenger. 1991. *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Lesh, Richard, and Helen M. Doerr. 2003. Foundations of a Models and Modeling Perspective on Mathematics Teaching, Learning, And Problem Solving." In *Beyond Constructivism: Models and Modeling Perspectives on Mathematics, Problem Solving, Learning, and Teaching,* edited by Richard Lesh and Helen M. Doerr, 3–33. New York: Routledge.

Lewis, Jeffrey B., and Drew A. Linzer. 2005. "Estimating Regression Models in Which the Dependent Variable Is Based on Estimates." *Political Analysis* 13 (4): 345–64.

Madda, Mary Jo. 2016. "Amazon Launches 'Inspire,' a Free Education Resource Search Platform for Educators." https://www.edsurge.com/news/2016-06-27-amazon-launches-inspire-a-free-education-resource-search-platform-for-educators (accessed July 3, 2018).

Mihaly, Kata, Daniel F. McCaffrey, Douglas O. Staiger, and J. R. Lockwood. 2013. "A Composite Estimator of Effective Teaching." https://pdfs.semanticscholar.org/68ae/e9b3f3ffdab84138bad40f3bc87db0281949.pdf.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2013. "Contract Teachers: Experimental Evidence from India." National Bureau of Economic Research (NBER) Working Paper 19440.

Opfer, V. Darleen, Julia H. Kaufman, and Lindsey E. Thompson. 2017. *Implementation of K–12 State Standards for Mathematics and English Language Arts and Literacy*. RAND. Santa Monica, CA, April.

Pianta, Robert C. 2011. *Teaching Children Well: New Evidence-Based Approaches to Teacher Professional Development and Training*. Center for American Progress. Washington, DC, November.

Purcell, Kristen, Alan Heaps, Judy Buchanan, and Linda Friedrich. 2013. *How Teachers Are Using Technology at Home and in Their Classrooms*. Pew Research Center. Washington, DC, February.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (2): 417–58.

Rothstein, Jesse. 2015. "Teacher Quality Policy When Supply Matters." *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): 323–38.

Rubin, Donald B. 2004. *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: John Wiley and Sons (Orig. pub. 1987).

Sawyer, R. Keith. 2006. *The Cambridge Handbook of the Learning Sciences*. Cambridge: Cambridge University Press.

Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data*. Monographs on Statistics and Probability. London: Chapman and Hall.

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*. National Center for Education Statistics. Washington, DC.

Taylor, Eric S. 2015. "New Technology and Teacher Productivity." Paper presented as CESifo Area Conference on the Economics of Education, Munich, September 11–12.

Taylor, Eric S., and John H. Tyler. 2012. "The Effect of Evaluation on Teacher Performance." *American Economic Review* 102 (7): 3628–51.