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

By Kirabo Jackson and Alexey Makarin

Online Appendix

For Online Publication.

Appendix A. Treatment Allocation.

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

Treatment By District

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

Appendix B. Auxiliary Results Regarding Requested Teachers.

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

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

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

Table B2. Main Result by Requested Status.

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

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

Appendix C. Construction of Factors for The Student Survey.

Table C1. Factor Loadings in the Construction of Student Survey Factors.

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

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

Table C2. Pairwise Correlations Between Student Survey Factors.

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
	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
Factor 1	1					
Factor 2	0.503	1				
Factor 3	0.238	0.343	1			
Factor 4	0.265	0.379	0.371	1		
Factor 5	0.539	0.520	0.257	0.304	1	
Factor 6	0.379	0.412	0.193	0.284	0.595	1

Appendix D. Teacher Survey.

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

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

Table D1 summarizes our regression results. Unfortunately, there are large difference in survey response rates across the treatment arms for teachers. The fully treated teachers were 12 percentage points more likely to response to the surveys than control teachers. As such, one should interpret the teacher survey results with caution. Having presented the limitation of the teacher surveys, the data provide little evidence that either the full treatment or the license only treatment has any effect on teacher satisfaction, teacher classroom practices, or their perception of the classroom dynamics among students. The only practice for which the effect is on the borderline of being statistically significant is treatment teachers assigning more homework. Taken at face value, these patterns suggest that teacher in the full treatment condition simply substituted the off-the-shelf lessons for their own lessons and may have assigned more homework as a results. However, treated teachers did not appear to make many any other changes to their classroom practices or teaching style. This implies that the positive observed effects simply reflect off-the-shelf substituting for low teacher skills rather than any learning of change in teacher teaching style.

Table D1. Teacher Post-Treatment Survey Analysis.

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

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

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

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

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

Appendix E. Stylized Model of Teacher Multitasking.

E1. Set-up

Let us consider the general optimization problem for a teacher. In our model, a teacher cares about her students' test scores (y_i, where i is a student from a class of size s) and leisure (l). Student i's test score depends on how much time the teacher spends planning lessons (d) and other complementary teaching tasks (n). A teacher's (in)ability to plan lessons is modeled as a 'price' p_d that amplifies the time needed to achieve d units of lesson quality. Similarly, 'price' p_n denotes the teacher's ability to achieve n units of other teaching tasks. Note that the higher teacher abilities are, the lower are her corresponding p's. Each teacher chooses the allocation of her total time (T) toward leisure (l), lesson planning (d) and other teaching tasks (n) in order to maximize her utility. Formally, we write:

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

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

$$
\begin{gathered}
U\left(\left\{y_i(n,d)\right\}_{i=1}^{s}, l\right) \rightarrow \max_{\{n,d,l\}} \\
\text{s.t. } p_n n + p_d d + l \leq T + p_d\underline{d} - F \\
n \geq 0\,;\, d \geq \underline{d}\,;\, l \geq 0
\end{gathered} \tag{2}
$$

E2. Special Case with Functional Form Assumptions

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

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

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

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

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

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

E2.1. Baseline Results

No Off the shelf Lessons

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

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

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

With Off the shelf Lessons

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

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

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

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

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

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

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

However, it could be that $\tilde{d} < \underline{d}$ such that the adopting teacher would locate at the kink of the budget line, as in case B in Figure 1. To solve for maximum teacher utility at the kink, one sets d^K=$\underline{d}$ and maximizes teacher utility with respect to l and n. In this case, the values of the main variables are:

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

E2.2. Auxiliary Lemmas.

Before we move forward to proving the main results of the model, we state the following two lemmas:[2]

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

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

Lemma 2. A teacher adopts off-the-shelf lessons in one of two cases: (i) whenever $\tilde{U} \geq U^* \geq U^K$ and $\tilde{d} \geq \underline{d}$ (case A in Figure 1), (ii) whenever $\tilde{U} \geq U^K \geq U^*$ and $\tilde{U}$ is unattainable because $\tilde{d} < \underline{d}$ (case B in Figure 1).

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

E2.3. Predictions.

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

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

Consider the second case (ii) from Lemma 2. In this case, since $\tilde{U}$ is unattainable, it must be that $\underline{d} > \tilde{d}$. However, we still have that $\tilde{U} \geq U^* \iff F \leq p_d \underline{d}$. Therefore, we get that $\underline{d} \geq \tilde{d} = (\beta/(\alpha+\beta+\gamma))(T - F + p_d \underline{d})/p_d \geq (\beta/(\alpha+\beta+\gamma))(T/p_d) = d^*$. Hence, adoption of the off-the-shelf lessons does not lead to a decrease in lesson quality, d. ■ [3]

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

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

Consider the second case (ii) from Lemma 2. In this case, we get that whether $n^K \geq n^*$ or not will

[2]Note that these two lemmas do not require the Cobb-Douglas functional form assumptions and would hold under any other monotonic utility function.

[3]It is worth noting that this result holds regardless of the Cobb-Douglas functional form assumptions as long as n, d, and l are all normal goods. For intuition, observe that d increases in both cases A and B in Figure 1 and that similar figures can be drawn for any monotonic and quasi-concave utility function.

depend on the parameters of the model. Specifically, $n^K \geq n^*$ whenever $F/T \leq \beta/(\alpha+\beta+\gamma)$, while $n^K < n^*$ whenever $F/T > \beta/(\alpha+\beta+\gamma)$. ■ [4]

Proposition 3. *The gains in average test scores from using the off-the-shelf lessons are non-negative.*
Proof. Consider the first case (i) from Lemma 2. From Proposition 1 and Proposition 2, we know that in this case both n and d weakly go up with lesson adoption. Hence, average test scores, $bn^\alpha d^\beta$, weakly go up. That is, if a teacher adopts and does not locate at the kink, then test scores will weakly increase.

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

$$b(n^K)^\alpha (d^K)^\beta = b\underline{d}^\beta \left(\frac{\alpha}{\alpha+\gamma}\frac{T-F}{p_n}\right)^\alpha \geq b\left(\frac{\alpha^\alpha\beta^\beta}{(\alpha+\beta+\gamma)^{\alpha+\beta}}\right)\frac{T^{\alpha+\beta}}{p_n^\alpha p_d^\beta} = b(n^*)^\alpha (d^*)^\beta \tag{3}$$

$$(p_d\underline{d})^\beta \geq \frac{\beta^\beta(\alpha+\gamma)^\alpha}{(\alpha+\beta+\gamma)^{\alpha+\beta}}\frac{T^{\alpha+\beta}}{(T-F)^\alpha} = C \tag{4}$$

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

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

$$U^K = b\frac{\alpha^\alpha\gamma^\gamma}{(\alpha+\gamma)^{\alpha+\gamma}}\left(\frac{\underline{d}^\beta(T-F)^{\alpha+\gamma}}{p_n^\alpha}\right) \geq b\frac{\alpha^\alpha\beta^\beta\gamma^\gamma}{(\alpha+\beta+\gamma)^{\alpha+\beta+\gamma}}\left(\frac{T^{\alpha+\beta+\gamma}}{p_n^\alpha p_d^\beta}\right) = U^* \tag{5}$$

$$(p_d\underline{d})^\beta \geq \frac{\beta^\beta(\alpha+\gamma)^{\alpha+\gamma}}{(\alpha+\beta+\gamma)^{\alpha+\beta+\gamma}}\frac{T^{\alpha+\beta+\gamma}}{(T-F)^{\alpha+\gamma}} \tag{6}$$

$$(p_d\underline{d})^\beta \geq C\left[\frac{\alpha+\gamma}{\alpha+\beta+\gamma}\frac{T}{T-F}\right]^\gamma \tag{7}$$

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

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

However, if $\frac{\alpha+\gamma}{\alpha+\beta+\gamma}\frac{T}{T-F} < 1$, condition (7) does not tell us anything about condition (4) as for any $D < 1$ it could be that $C > (p_d\underline{d})^\beta \geq CD$. To make progress on this case, note that re-writing the first inequality leads to $F/T < \beta/(\alpha+\beta+\gamma)$. However, from Proposition 2, we know that in this case both $n^K > n^*$ and $d^K > d^*$, meaning that test scores must go up as $b(n^K)^\alpha(d^K)^\beta > b(n^*)^\alpha(d^*)^\beta$. Hence, adoption of off-the-shelf lessons leads to a non-negative effect on test scores. ■ [5]

[4]This result also does not require the Cobb-Douglas functional form assumptions. For intuition, observe that n increases in case A in Figure 1 but decreases in case B in Figure 1 and that similar figures can be drawn for any monotonic and quasi-concave utility function.

[5]Note that the first part of the proposition does not require the Cobb-Douglas functional form assumptions. In fact, it holds for any utility function for which n, d, and l are all normal. For such functions, as long as the teachers who

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

Proposition 4. *The relationship between the test score benefits of lesson use and teacher quality is ambiguous in sign as it depends on the definition of teacher quality. To see this, assume $\alpha+\beta=1$ and an interior (non-kink and non-corner) solution.*
Case 1*: In this scenario, better teachers are those that can produce better test scores with less time than weaker teachers. If teacher quality is defined as a set of (p_n, p_d), an increment in test scores $[b\tilde{n}^{\alpha}\tilde{d}^{\beta} - b(n^*)^{\alpha}(d^*)^{\beta}]$ goes down with a decrease in prices.*
Case 2*: Teacher quality is defined as ability to adopt new ways of teaching. As a simplification, consider that low teaching quality means that a teacher has a higher F. In this case, the difference $[b\tilde{n}^{\alpha}\tilde{d}^{\beta} - b(n^*)^{\alpha}(d^*)^{\beta}]$ goes up with a decrease in F.*
Proof. Conditional on $\alpha+\beta=1$ and an interior (non-kink and non-corner) solution, an increment in utility from adopting the lessons takes the following form:

$$[b\tilde{n}^{\alpha}\tilde{d}^{\beta} - b(n^*)^{\alpha}(d^*)^{\beta}] = b\frac{\alpha^{\alpha}\beta^{\beta}\gamma^{\gamma}}{(\alpha+\beta+\gamma)^{\alpha+\beta+\gamma}}\left(\frac{p_d\underline{d}-F}{p_n^{\alpha}p_d^{\beta}}\right) \tag{8}$$

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

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

Moreover, one can prove that, when both p_d and p_n are increased *simultaneously by the same percentage*, the difference strictly increases. Specifically, after taking the exact differential of (8), we show that simultaneous increases of p_n and p_d by the same percentage (i.e. such that $dp_n/p_n = dp_d/p_d = \varepsilon$) lead to an increase of the total difference:

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

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

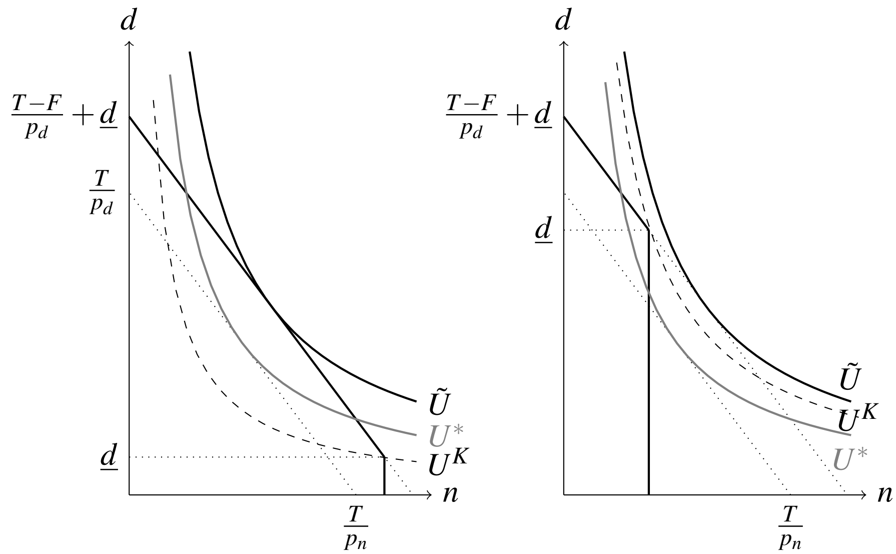
adopt the lessons put at least some extra effort into increasing lesson quality above the minimum, $\tilde{d} \geq \underline{d}$, test scores will not decrease.

Case 2: An alternative definition of teacher quality is the ability to adopt new technology. In principle, a more competent teacher should be able to adopt off-the-shelfs lessons at a lower cost, F. Taking a derivative of (8) with respect to F, one gets:

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

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

Figure 1: Illustration of the Model.



A. Teacher locates above the kink. B. Teacher locates at the kink.

Notes: This is an illustration of the stylized model presented in Section E2. As described there, T is the stock of time available to each teacher; d and n are the units of time spent on lesson planning and other tasks, respectively; p_d and p_n denote teacher (in)effectiveness in lesson planning and other tasks; F is the fixed time cost of adopting off-the-shelf lessons; $\underline{d}$ is lesson quality guaranteed by off-the-shelf lessons; U's are the indifference curves fixed at a certain level of average students' test score and a certain level of leisure.

Appendix F. Spillovers.

Table F1. Spillovers.

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

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

Appendix G. Effect Heterogeneity by Teacher Quality.

One of the methodological innovations of paper is to present a way to test for treatment heterogeneity by teacher quality with a single year of value-added data. To motivate out strategy, we start out with the standard teacher value-added model as presented in Jackson et al. (2014).[6] We show that marginal effects in this standard value-added model, when aggregated up to the teacher level, yield a very intuitive interpretation in a conditional quantile regression model. Specifically, we show that when average student test scores (aggregated at the teacher level) is the dependent variable, the estimated coefficient of a randomized treatment using conditional quantile regression at quantile τ is the estimated effect of that treatment on teachers at the τth percentile of the teacher quality distribution. We then present a Monte Carlo simulation showing that our method is valid.

The Standard Value Added Model

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

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

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

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

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

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

$$\mu_t = \beta T_t + \nu_t$$

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

$$\bar{Y}_t = \beta T_t + \bar{\mathbf{X}}_t\delta + \nu_t + \theta_c + \bar{\varepsilon}_t \tag{9}$$

As shown in (9), in a regression of average student test scores (aggregated to the teacher level) on treatment status and covariates, the unobserved error term is $\nu_t + \theta_c + \bar{\varepsilon}_t$. Accordingly, the residual from this regression is the teacher value added (without the treatment) plus noise (random classroom-level errors and aggregate student-level sampling variability). Following Chetty et al. (2011), we consider this teacher-level residual a noisy measure teacher quality (without the treatment). Accordingly, we refer to this teacher-level residual as teacher value added.

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

Applying The Conditional Quantile Function

The notation above implicitly assumes that the marginal effects of the treatment and the covariates were the same for all teachers. We now explicitly allow for the possibility that the treatment effect varies by teacher value added (as defined above). Using the nomenclature from Koenker and Hallock (2001), the conditional quantile (τ) is the τth quantile of the conditional distribution of the response variable. More simply, (τ) is the τth quantile of the distribution of the residual. As discussed above, the residual from (9) is precisely our measure of teacher value added. The conditional quantile estimator as introduced by Koenker and Bassett (1978) estimates the marginal effect of a treatment at different points of the conditional distribution of an outcome. We show below that this model, applied to the aggregate value-added model as in (9), yields consistent estimates of the marginal effect of the treatment for teacher at different points in the value-added distribution.

Allowing for the possibility that β and δ may vary with the quantile τ, let us apply the conditional quantile function to (9). This yields (11) below:

$$Q_\tau[\bar{y}_t|T,\bar{\mathbf{X}}] = \beta(\tau)T_t + \bar{\mathbf{X}}_t\delta(\tau) + Q_\tau[\nu_t(\tau) + \theta_c(\tau) + \bar{\varepsilon}_t(\tau)|T,\bar{\mathbf{X}}] \tag{10}$$

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

$$[\hat{\beta}(\tau), \hat{\delta}(\tau)] = \min_{b,d} \sum_{t=1}^{T} \rho_\tau[\bar{y}_t - bT_t - \bar{\mathbf{X}}_t d] \tag{11}$$

, where $\rho_\tau = u_t[\tau - \mathbb{1}(u_t < 0)]$ is a re-weighting function of the residuals $u_t = \nu_t + \theta_c + \bar{\varepsilon}_t$. As demonstrated in Buchinsky (1998), if $Q_\tau[\nu_t(\tau) + \theta_c(\tau) + \bar{\varepsilon}_t(\tau)|T,\bar{\mathbf{X}}] = 0$ for each quantile τ, the conditional quantile regression coefficient $\hat{\beta}(\tau)$ is a consistent estimate of $\beta(\tau)$ in (11).[7]

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

Mote Carlo Simulation

Because the presentation above is somewhat theoretical, to provide concrete evidence that our procedure works, we assigned random treatments to the teachers in our data,[8] created simulated causal effects that varied based on each teacher's residual,[9] and then estimated the conditional

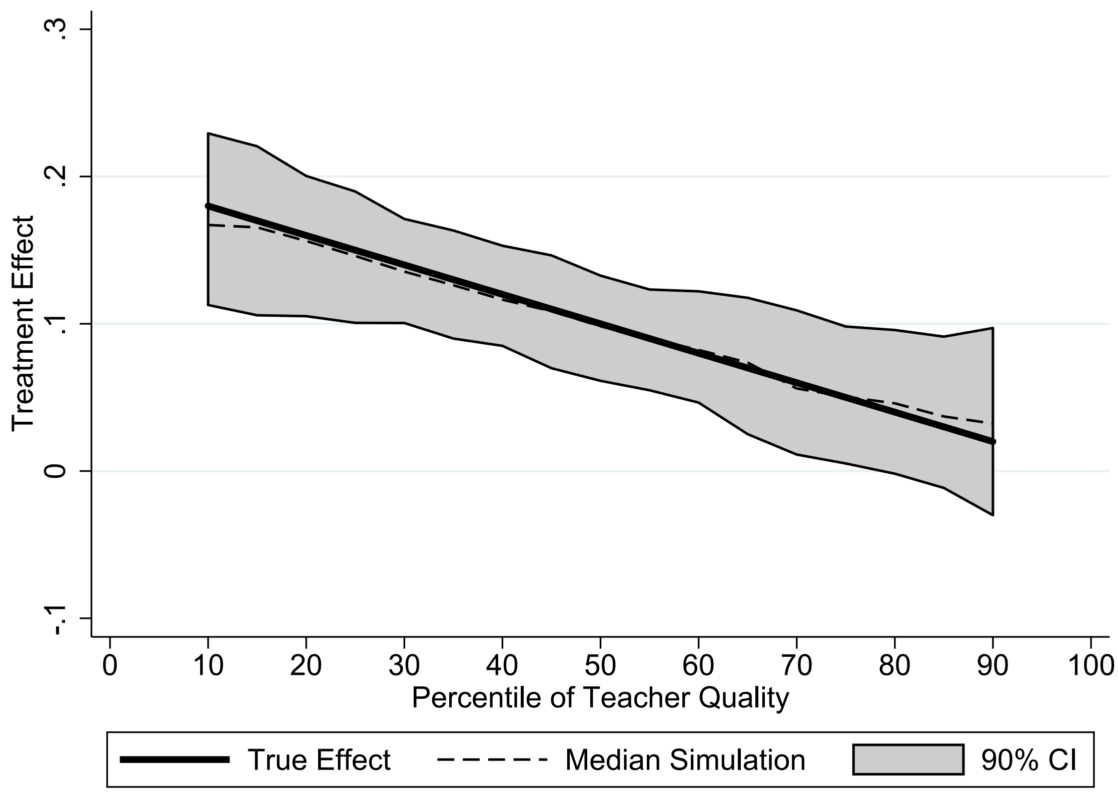
[7] This is a standard assumption in the quantile regression literature. For a reference, see e.g. Buchinsky (1998)

[8] We retained the same distribution of treatments in the data. To do this we randomly "reassigned" teachers to the actual treatments in the data to create "simulated" treatment assignments. We then created simulated treatment effects based on these assignments.

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

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

Figure G1. Simulation Results.



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

Appendix H. Test Score Regressions - Teacher Level.

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

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

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

Appendix I. Heterogeneous Effects by Teacher Experience.

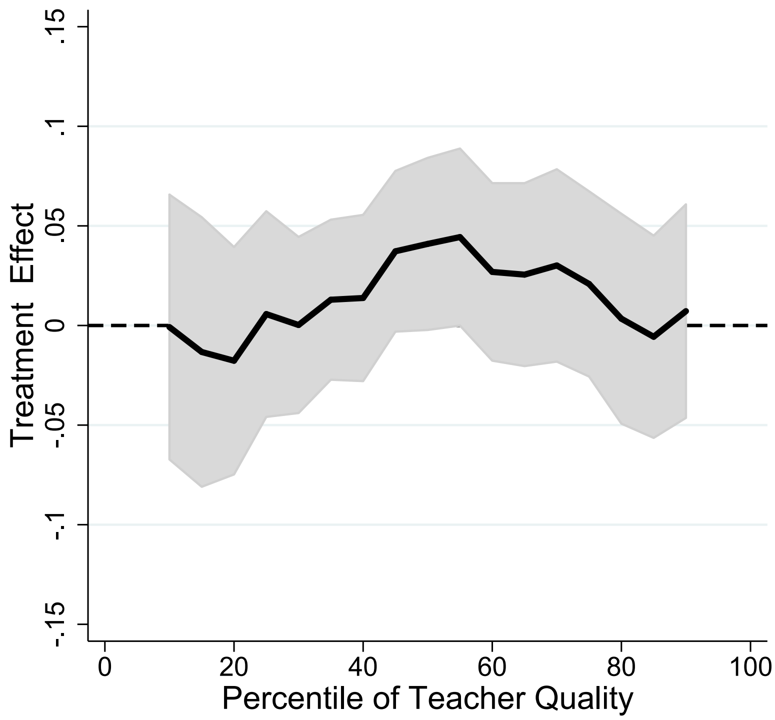
Table I1. Heterogeneous Effects by Teacher Experience.

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

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

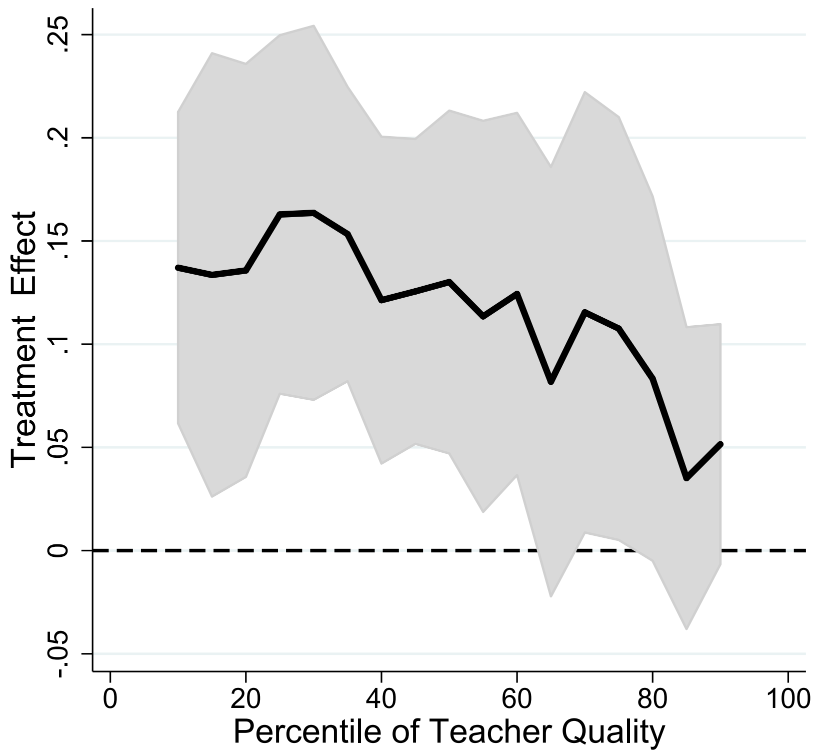
Appendix J. Quantile Regression: Robustness Checks.

Figure J1. Marginal Effect of the Full Treatment by Classroom Quality.
Falsification Test: English Test Scores.



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

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



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

Appendix K. Survey Response and Lesson Downloads.

Table K1. Survey Response and Lessons Downloads.

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

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

Appendix L. Auxiliary Results on Lesson Use.

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

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

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

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

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

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

Table L2. Effects on Lesson Use by Requested Status.

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

	Lessons Looked	Lessons Taught	Lessons Downloaded	Webinars Viewed
	(1)	(2)	(3)	(4)
License Only	1.914***	0.653***		
	[0.519]	[0.185]		
Full Treatment	3.229***	2.100***	N/A	N/A
	[0.621]	[0.461]		
License Only x Requested	-0.786	-0.073		
	[0.936]	[0.334]		
Full Treatment x Requested	2.348*	-0.443		
	[1.246]	[0.594]		
District FE x Requested	Y	Y		
All controls	Y	Y		
Observations	363	363		

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

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

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

Appendix M. Auxiliary Results on Student Surveys.

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

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

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

Appendix N. Instrumental Variables Estimation.

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

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

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

Notes: *** - significance at less than 1%; ** - significance at 5%, * - significance at 10%. Standard errors clustered at the teacher level are reported in square brackets. Columns (1) and (4) report the results of OLS estimation, while Columns (2)-(3) and (5)-(8) contain the results of 2SLS estimation where the number of Mathalicious lessons taught is instrumented by the treatment status. All specifications include controls for the requested indicator, average teacher-level 2013 math and reading test scores, and a teacher-level shares of students with missing 2013 math and reading test scores - all interacted with district fixed effects. Additional controls include teachers' education level, years of experience, sex, race, grade fixed effects, as well as the percentage of male, black, white, Asian, and Hispanic students in their class. In addition, the student-level specifications in Columns (1)-(3) and (7)-(8) control for individual-level math and reading test scores and all student level demographics. Standardized test scores refer to the raw test scores standardized by exam type. In the absence of exam type data for Hanover, test scores for that district were standardized by grade.

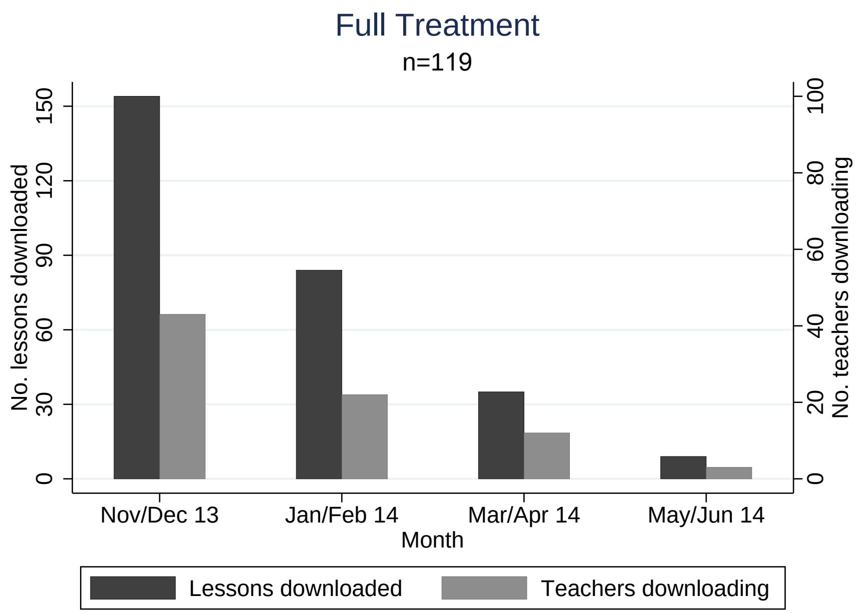
Appendix O. Patterns of Lesson Use Over Time

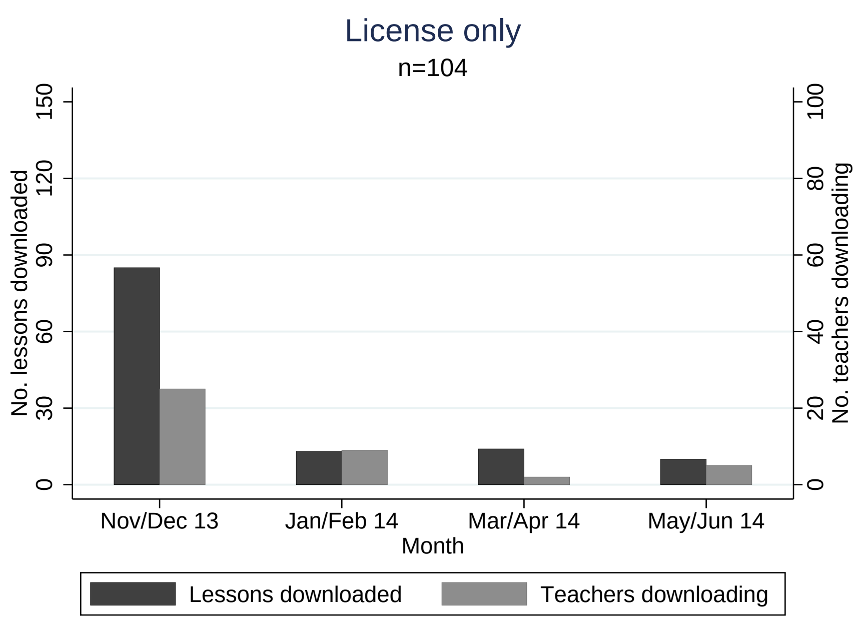
Given the sizable benefits to using the off-the-shelf lessons, one may wonder why lesson use was not even more widespread. To gain a sense of this, we present some graphical evidence of lesson use over time. Figure X1 shows the number of lessons downloaded by license only and full treatment groups in different months. As expected, lesson use was much larger in the full treatment condition than that in the license only condition. However, Figure X1 reveals a few other interesting patterns. There was a steady decline in the number of lessons downloaded over time within groups. While there were 97 downloads in the full treatment in November 2014, there were only 8 downloads in May 2015. Similarly, in the license only group, while there were 59 downloads in the November 2014, there were only 4 downloads in May 2015. To determine whether this decline is driven by the same number of teachers using Mathalicious less over time, or a decline in the number of teachers using Mathalicious over time, we also plot the number of teachers downloading lessons by treatment group over time. There is also a steady decline in the number of teachers downloading lessons so that the reduced use is driven by both reductions in downloads among teachers, and a reduction in the number of teachers downloading lessons over time.

Even though we have no dispositive evidence on why lesson use was not higher, or why lesson use dropped off over time, we speculate that it may have to do with behavioral biases and time management. The patterns of attrition from lesson downloads over time are remarkably similar to the patterns of attrition at online courses (Koutropoulos et al., 2012), gym attendance (DellaVigna and Malmendier, 2006), and fitness tracker use (Ledger and McCaffrey, 2014). Economists hypothesize that such behaviors may be due to individuals underestimating the odds that they will be impatient in the future and then procrastinate (O'Donoghue and Rabin, 1999; Duflo et al., 2011). Similar patterns in Figure 4 provide a reason to suspect that similar behaviors may be at play. In our context, these patterns may reflect teachers being optimistic about their willpower to use the lessons such that they started out strong, but when the time came, they procrastinated and did not make the time to implement them later on. However, it is also possible that as teachers use the lessons, they perceive that they are not helpful and decide to discontinue their use after downloading the first few lessons. Most of the empirical patterns support the former explanation. First, the rate of decay of lesson use is more rapid in the license only treatment than in the full treatment group. Specifically, without the additional supports to implement the lessons, the drop-off in lesson use was more rapid. In the full treatment group, downloads fell by about 45 percent between Nov/Dec and January/Feb, while it fell by over 80 percent during that same time period in the license only group. If the reason for the drop-off was low lesson quality, drop-off should have been similarly rapid for both groups. The second piece of evidence is that the there is a sizable reduction in lessons downloaded in the full treatment condition after February when Mathalicious ceased sending out email reminders to teachers, while lesson use was stable in the license only condition. The third piece of evidence comes from surveys. We employed data from the end of year survey that asked treated teachers why they did not use off-the-shelf lessons more. Looking specifically at the question of whether the lessons were low quality, only 2 percent of teachers mentioned this was a major factor and almost 89% stated that is was not a factor at all. In sum, poor lesson quality does not explain the drop-off in lesson use, being reminded mattered, and the patterns of drop-off are very similar to other contexts in which behavioral biases played a key role – suggesting that procrastination is a plausible explanation.

The last piece of evidence to support the procrastination hypothesis also comes from the survey evidence shown in Figure X2. The main reason cited for not using more lessons was a lack of time. Taken at face value, one might argue that the pressures on teacher time increased over the course of the year such that lesson use declined over time. However, this cannot explain the large differences in the trajectory of lesson use over time across the treatment arms. The explanation that best fits the observed patterns and the survey evidence is that, without the reminders and extra supports (i.e. Edmodo groups), teachers were unable to hold themselves to make the time to implement the lessons. The patterns also suggest that providing ways to reduce procrastination during the school year (such as sending constant reminders or providing some commitment mechanism) may be fruitful ways to increase lesson use. Other simple approaches may reduce the incentive to procrastinate at the moment by providing designated lesson planning time, or granting lesson access the summer before the school year when the demands on teachers' time may be lower.

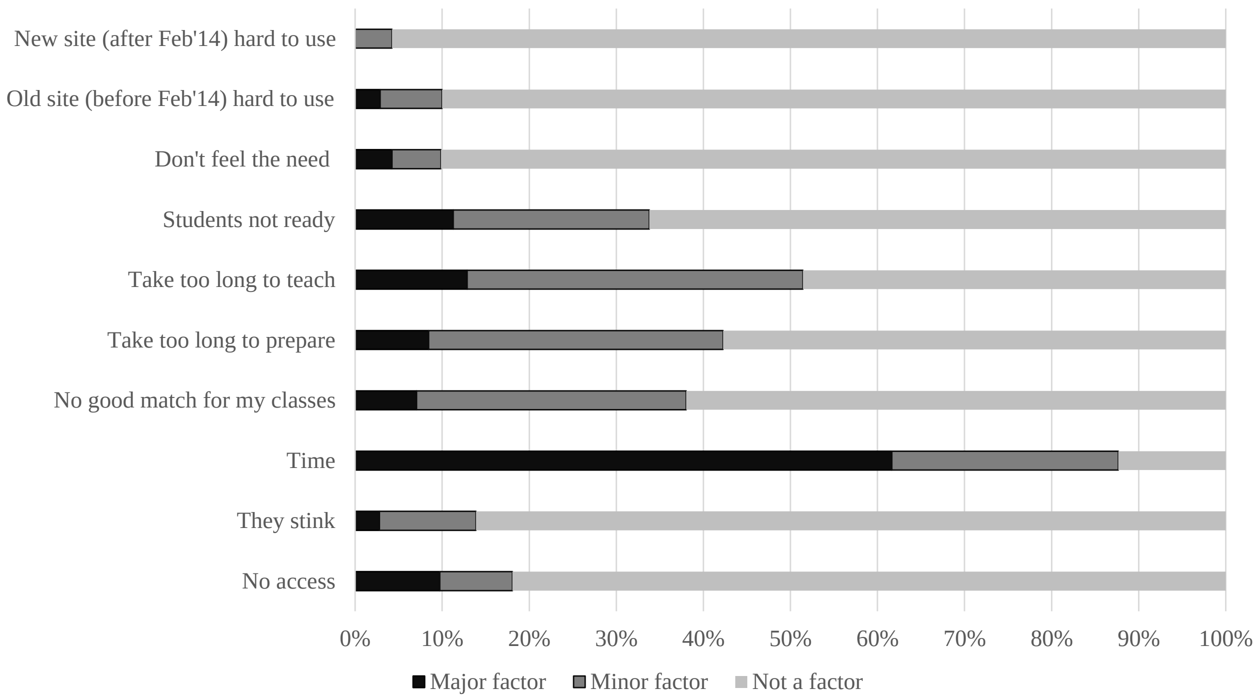
Figure O1. Downloads of Mathalicious Lessons Over Time





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

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



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

Appendix P. Sample Mathalicious Lesson #1.

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

XBOX XPONENTIAL

How have video game console speeds changed over time?

lesson guide

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

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

Primary Objectives

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

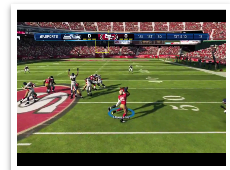
Content Standards (CCSS)	Mathematical Practices (CCMP)	Materials
Functions: IF.8b, BF.1a, LE.2, LE.5 Statistics: ID.6a	MP.4, MP.7	• Student handout • LCD projector • Computer speakers • Graphing calculators • Computers with Internet access

Before Beginning…

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

Preview & Guiding Questions

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



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

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

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

Act One

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

Act Two

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

Act One: Moore Fast

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

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

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

×2 *×2×2* *×2×2×2×2×2* *×2×2×2×2* *×2×2*

Explanation & Guiding Questions

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

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

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

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

Deeper Understanding

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

Appendix Q. Sample Mathalicious Lesson #2.

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

NEW-TRITIONAL INFO

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

lesson guide

mathalicious

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

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

Primary Objectives

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

Content Standards (CCSS)	Mathematical Practices (CCMP)	Materials
Grade 6 RP.3d, NS.3	MP.3, MP.6	• Student handout • LCD projector • Computer speakers

Before Beginning…

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

Preview & Guiding Questions

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

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

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

Act One

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

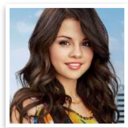
Act Two

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

Act One: Burn It

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

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

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

Explanation & Guiding Questions

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

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

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

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

Deeper Understanding

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

Appendix References

M. Buchinsky. Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of Human Resources*, pages 88–126, 1998.

R. Chetty, J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan. How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics*, 126(4):1593–1660, 2011.

S. DellaVigna and U. Malmendier. Paying not to go to the gym. *The American Economic Review*, pages 694–719, 2006.

E. Duflo, M. Kremer, and J. Robinson. Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya. *The American Economic Review*, pages 2350–2390, 2011.

C. K. Jackson, J. E. Rockoff, and D. O. Staiger. Teacher Effects and Teacher-Related Policies. *Annual Review of Economics*, 6(1):801–825, 2014.

R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, pages 33–50, 1978.

R. Koenker and K. Hallock. Quantile regression: An introduction. *Journal of Economic Perspectives*, 15(4):43–56, 2001.

A. Koutropoulos, M. S. Gallagher, S. C. Abajian, I. de Waard, R. J. Hogue, N. O. Keskin, and C. O. Rodriguez. Emotive vocabulary in MOOCs: Context & participant retention. *European Journal of Open, Distance and E-Learning*, 15 (1), 2012.

D. Ledger and D. McCaffrey. Inside wearables: How the science of human behavior change offers the secret to long-term engagement. *Endeavour Partners*, 2014.

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

D. B. Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.