SKILLS, JOB TASKS, AND PRODUCTIVITY IN TEACHING: EVIDENCE FROM A RANDOMIZED TRIAL OF INSTRUCTION PRACTICES

Eric S. Taylor[†]

Harvard Graduate School of Education

February 2016

I study how teachers' assigned job tasks—the basic practices they are asked to use in the classroom—affect the returns to math skills in teacher productivity. The results demonstrate the importance of distinguishing between workers' skills and worker's job tasks. I examine a randomized trial of different approaches to teaching math, each approach codified in a set of day-to-day tasks. Teachers were tested to measure their math skills. Teacher productivity—measured with student test scores—is increasing in math skills when teachers use conventional "direct-instruction" practices: explaining and modeling math rules and procedures. The relationship is weaker, perhaps negative, for newer "student-led" instruction tasks.

JEL No. I2, J24, M5

[†] Gutman Library 469, 6 Appian Way, Cambridge, MA 02138, eric_taylor@gse.harvard.edu. I thank Eric Bettinger, Brian Jacob, Susanna Loeb and seminar participants at Stanford University for helpful discussions and comments. I also thank the Institute for Education Sciences for providing access to data, and the original research teams who carried out the experiment and collected the data. Financial support was provided by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B090016 to Stanford University; and by the National Academy of Education/Spencer Dissertation Fellowship Program. Any mistakes are my responsibility.

Economists and managers have long studied whether and how differences in workers' skills (human capital) generate differences in workers' productivity. In the standard theoretical model, with foundations in the work of Becker (1964) and Tinbergen (1974), units of labor produce output and workers' varying skills make each unit of labor more (less) productive. Recently, Acemoglu and Autor (2011, 2012) have emphasized the distinction between workers' skills and the job tasks to which those skills are applied. Competed tasks produce output, not skills, and thus identically-skilled workers assigned different tasks can have different output.[1]

In this paper I demonstrate the value of this skills-tasks distinction using micro-data on elementary school math teachers. Specifically, I analyze a field-experiment in which teachers, with observed measures of math skills, were randomly assigned to follow one of four different instructional methods for teaching early-elementary math. Each instructional method is a set of specific tasks which teachers are asked to carry out in their classrooms.

Understanding how skills and job tasks translate into productivity is especially relevant and timely in public schools. A consistent empirical literature documents substantial between-teacher variation in job performance, especially teacher productivity as measured by teachers' contributions to student test score gains. In recent years, these differences in teacher performance have become central to political and managerial efforts to improve public schools. Yet relatively little is known about what causes this variation. In particular, several plausible measures of relevant skills do not consistently predict differences in teacher productivity; this lack of evidence is, however, not for lack of research effort. Evidence on how job tasks affect teacher

[1] While formalized in Acemoglu and Autor (2011, 2012), this model incorporates prior work by Autor, Levy, and Murnane (2003), Acemoglu and Zilibotti (2001), and Costinot and Vogel (2010).

productivity is, in comparison, even scarcer.[2] Moreover, interactions between skills and tasks have not, as far as I am aware, been explicitly studied with empirical data.[3] This lack of information on the causes of productivity differences constrains teacher policy and management.

I study teachers' math skills as measured by the Mathematical Knowledge for Teaching (MKT) test, and variation in teachers' tasks across four different approaches to teaching early-elementary math codified in commercially published materials. The MKT, administered pre-experiment, is designed to test a teacher's knowledge of math concepts and procedures per se, as well as her knowledge of how young students (mis)understand the math they are learning (Hill, Schilling, and Ball 2004, Hill, Rowan, and Ball 2005). The four approaches, randomly assigned to schools, can be primarily characterized as "direct-instruction" or "student-led" methods. In direct-instruction classrooms, the more conventional of the two approaches, teachers explicitly describe and model math concepts and procedures, and students practice skills frequently. In student-led classrooms the students are expected to reason-through and articulate math concepts with each other, while teachers "facilitate conversations" and "help students express their thoughts" with a "focus on [students'] understanding, rather than on students answering problems correctly" (Agodini et al. 2010, pp. xxi, 6-7).

I show, first, that students' math test scores are positively correlated with their teacher's math skills, as measured by MKT score; but this correlation goes to zero after accounting for the non-random sorting of students to teachers. Second, however, I show that this apparent zero correlation masks meaningful heterogeneity caused by the different tasks (instructional methods)

[2] For a review of the literature on teacher performance generally, including the evidence on skills and job tasks, see Jackson, Rockoff, and Staiger (2014). Rockoff et al. (2011) provide a thorough review of existing evidence on the role of teacher skills. I discuss the literature in Section 1 of this paper.

[3] In the closest work that I am aware of, Stein and Kaufman (2010) study the extent to which elementary math teachers successfully or faithfully follow the instructional methods they are asked to use. They do not find any correlation between implementation and teachers' knowledge, education, or job experience.

teachers are assigned. The correlation between teacher skills and productivity is positive and meaningful when teachers use "direct-instruction" methods. By contrast, the correlation is much weaker, perhaps even negative, when teachers use a "student-led" approach to teaching math. In short, whether and how a teacher's math skills contribute to her productivity depends on how she is asked to teach math. Student-led and direct-instruction methods generate quite different relationships between skills and productivity, as measured by teachers' contributions to testable student learning.

Importantly, productivity differences between student-led and direct-instruction methods are apparently driven by high-skilled teachers not their average- and low-skilled colleagues. Comparing only classrooms with teachers in the top-tercile of MKT scores, students taught with the strongly student-led methods score 0.13-0.16 standard deviations lower at the end of the year than their peers taught with more-conventional direct-instruction. By contrast, there is little difference across classrooms with teachers in the bottom- or middle-tercile of MKT rank.

These higher returns to math skills for teachers using direct-instruction are consistent with the key differences in tasks between direct-instruction and student-led methods. Direct-instruction methods, far more frequently than student-led methods, rely on the teacher to demonstrate and explain math concepts and procedures to her students. The ability to *correctly* demonstrate and explain—an ability measured by the MKT test—will have more value in direct-instruction classrooms.

Moreover, the teaching tasks or methods teachers adopt can substantially shrink (expand) the variation in teacher productivity. The standard deviation of productivity among teachers using direct-instruction is 0.12-0.19 student standard deviations. Among teachers using student-led methods the variation was at least one-third smaller: 0.08 student standard deviations. These

treatment effects on the variance of productivity are total effects not just changes in the returns to skills.

Causal interpretation of these estimates relies largely on the random assignment of schools and teachers to the four instructional-method treatment conditions. Two points on identification are notable. First, within any given treatment condition the slope of the relationship between teacher MKT and student test scores is *not* causally identified, but the differences in slope across treatment conditions are casually identified. The slopes may be biased by omitted variables, though I show that my estimates are robust to several common and less-common measures of teachers. Second, differences in the variance of teacher productivity are causal even though students were not randomly assigned to teachers. To identify the difference in variance requires only that, if there is any (residual) bias in estimating teachers' contributions to student test scores, the bias is independent of treatment assignment.

This paper is the first, of which I am aware, to demonstrate that the job tasks teachers' are assigned partly determine the returns to teacher skills in education production, and partly determine the variability in teacher productivity more generally. The results suggest decisions about how teachers' are asked to teach can be as important as decisions about who is hired to teach. More generally, these results show the value of conceptually separating workers' skills and job tasks when proposing empirical tests of the relationships between skills and productivity.

1. Existing evidence on teacher productivity, skills, and job tasks

A consistent empirical literature documents substantial between-teacher variation in job performance—variation revealed by differences in observable student outcomes (for a review see Jackson, Rockoff, and Staiger 2014). Indeed, economists have been studying teacher

productivity, as measured by contributions to student learning growth, for more than four decades (with original work by Hanushek 1971 and Murnane 1975). In a typical result, students assigned to a teacher at the 75th percentile of the job performance distribution will score between 0.07-0.15 standard deviations higher on achievement tests than their peers assigned to the average teacher. Newer evidence documents equally important between-teacher variation with non-test-score outcomes, including students' non-cognitive skills (Jackson 2013) and students' long-run economic and social success as adults (Chetty, Friedman, and Rockoff 2014b). Evidence on what causes these differences in teacher performance is much scarcer.

Differences in teachers' skills—each teacher's stock of current capabilities whether innate, or acquired by training or experience, or both—are an intuitive explanation for differences in productivity. Indeed, a large body of research has examined several types of relevant skills and several plausible measures, including: (i) general cognitive ability, often measured by prior academic success; (ii) specific knowledge of the subject the teacher is assigned to teach; (iii) teaching-specific skills, often measured by certification exams; and (iv) non-cognitive skills, interpersonal skills, and relevant personality traits. No consistent patterns emerge from reading this research; some skill measures explain performance differences in one setting but not in another (for reviews see Rockoff et al. 2011 and Hanushek 1997).

There are at least two hypotheses for the lack of consistent patterns. First, the returns to specific skills depend on the job tasks those skills are applied to; different research results may partly reflect differences in sample teachers' jobs. Second, many (most) easily-observable measures of teaching skills are empirically poor measures (e.g., noisy, little variation). This paper is partly motivated by the first hypothesis. The second hypothesis is a critical consideration in selecting a measure of skills to test the first hypothesis.

Two notable results suggest the importance of hypothesis two and of selecting skill measures. First, Rockoff and coauthors (2011) and Dobbie (2011), each studying different data, show that composite indices composed of several skill measures do meaningfully predict teacher performance, but that the individual components are not predictors. In Dobbie's data the index may explain more than half of the variance in teacher test score effects (Jackson et al. 2014 p. 806). Second, several recent studies examine the highly teaching-specific skills measured in formal classroom observations; in these observations trained raters score teachers on a dozen or more specific instructional practices. These observed-skills measures also meaningfully predict teacher job performance (for example, Kane et al. 2011, Kane et al. 2013, and Jacob et al. 2015).[4]

In this paper I measure teachers' math skills using scores from the Knowledge of Mathematics for Teaching (MKT) test developed by Heather Hill and Deborah Ball (Hill, Schilling, and Ball 2004, Hill, Rowan, and Ball 2005). As I describe in greater detail in Section 2, the MKT is designed to measure math skills which are particularly relevant to teaching elementary-level math, not simply to test knowledge of mathematics per se. Empirically, teachers' MKT scores have been shown to predict their students' math test outcomes (see for examples, Hill et al. 2005, Rockoff et al. 2011, Hill, Umland, Litke, and Kapitula 2012).

Like differences in skills, differences in teachers' assigned job tasks likely contribute to differences in observed teacher productivity, but empirical studies on this topic have been comparatively rare. The most common, relevant evidence focuses on differences in job tasks

[4] A third notable result is also highly suggestive. Several studies now provide convincing evidence of returns to on-the-job experience in teaching (for example, Rockoff 2004, Rivkin, Hanushek, and Kain 2005, Papay and Kraft forthcoming). Of course, "years of experience" is not a direct measure of specific skills (different teachers learn different skills on-the-job), but it is likely correlated with a number of different skills. Thus years of experience is another kind of composite skill measure.

vary between grade levels or course subjects, or across schools; performance does change when a teacher's job changes on these dimensions (for a review see Jackson, Rockoff, and Staiger 2014). Differences in schools' use of instructional computer technology also affect teacher performance, and the effects appear to vary depending on teachers' skills (Taylor 2015).

2. Experimental setting, treatments, and data

Data for this paper were collected during a field-experiment in first and second grade math instruction. Schools, and thus their teachers, were randomly assigned to follow one of four different *instructional methods* for teaching math—the four treatment conditions—but the math *concepts* teachers were asked to cover during the school year did not vary. Alongside this experimental variation in teachers' job tasks, the study team tested teachers to measure math skills at baseline, and tested students pre- and post-experiment to measure learning growth. All data were collected during the first year of treatment (either 2006-07 or 2007-08); for most teachers and schools this was the first school year using the assigned instructional methods.[5]

In this paper I examine whether and how teachers' skills and assigned job tasks interact in the production of student learning. The critical features of the data, which I discuss in this section, are measures of teachers' skills and student learning, and exogenous variation in

[5] The original study was funded by the Institute for Education Sciences, U.S. Department of Education, and carried out by Mathematica Policy Research and SRI International. The discussion in this section focuses on topics most relevant to the current study. Additional topics and details can be found in the original experiment study report (Agodini, Harris, Thomas, Murphy, and Gallagher 2010), including extensive descriptions of the four treatment conditions' instructional methods and approaches.

Original descriptions of the experiment and results refer to the four treatment conditions as four different "curricula." I use the term "instructional methods" or "methods" since, for many readers, the word "curriculum" would imply treatment variation in the math concepts (or standards) teachers were asked to teach.

The experiment team collected (planned to collect) data during the second year of implementation in some schools. Those data are not yet available.

teachers' job tasks. While the data necessary for the current paper's research questions were collected by the original experiment team, the questions were not addressed in their analysis.[6]

Table 1 describes the study participants, including 110 schools and nearly 800 teachers and 9,000 students.[7] By design, the sample focuses on relatively high-poverty settings: three-quarters of schools were eligible for school-wide Title I funding, and about half of students are eligible for free or reduced price lunch. More than half of students were Latino or African-American, and one in seven was an English language learner. Teachers had, on average, 12 years of experience, nearly eight years in their current school. Just under half of teachers had a master's degree (in any field), and half reported having taken one or more advanced math classes in college.

2.1 Experimental treatments—variation in teachers' job tasks

Schools were randomly assigned, within blocks defined by district and observable characteristics, to one of four treatment conditions.[8] Each condition is a distinct method or approach for teaching early elementary math concepts, and each method is codified in a commercially published set of teacher instructions and classroom materials. The methods' commercial names are: *Investigations in Number, Data, and Space (Investigations)*; *Math Expressions (Expressions)*; *Saxon Math (Saxon)*; and *Scott Foresman-Addison Wesley Mathematics (SFAW)*.[9]

[6] The original reports include just one analysis using the MKT scores. Among a list of several effect heterogeneity analyses, differences in student test scores across conditions (i.e., treatment effects) were estimated separately for teachers in two groups: those in the bottom quintile of the sample MKT distribution, and all other teachers.
[7] Throughout the paper sample sizes have been rounded to the nearest 10 following NCES restricted data reporting rules.
[8] By rule, each randomization block includes at least four schools and at most seven. If a district has four schools there is one block, if eight schools then two blocks, etc; and if five schools then one block, if nine schools then one block of four and one of five, etc.
[9] The four products were selected in a competitive process conducted by IES. Investigations and SFAW are published by Pearson Scott Foresman. Expressions and Saxon are published by Houghton Mifflin Harcourt.

The first-order differences between the four approaches can be summarized by two dimensions: the use of "direct-instruction" methods and the use of "student-led" methods. Both *Saxon* and *SFAW* make extensive use of direct-instruction or teacher-directed methods, but few, if any, student-led methods. In direct-instruction, teachers explicitly describe and model math concepts and procedures, sometimes following a provided script; and students practice skills frequently. By contrast, *Investigations* is a strongly student-led or constructivist approach. Student-led methods "focus on [students'] understanding, rather than on students answering problems correctly" (Agodini et al. 2010, p. xxi). In student-led classrooms, teachers "spend much of their time facilitating conversations among students, helping students express their thoughts, and guiding students to a deeper understanding of math" (Agodini et al. 2010, p. 6-7). *Expressions* uses both direct-instruction and student-led methods, though written descriptions suggest greater weight is given to direct-instruction activities.

Empirical evidence on the relative effectiveness of these four approaches is scarce. The original analysis of this experiment, reported in Agodini et al. (2010), compared mean test scores in each condition, separately for first and second grade (8 mean estimates, and 12 pair-wise mean differences). There were four statistically significant differences: *Expressions* test scores were higher than both *Investigations* and *SWAF* in 1st grade classrooms, 0.11 student standard deviations (σ); and both *Expressions* and *Saxon* scores were higher than *SWAF* in 2nd grade, 0.12σ and 0.17σ respectively.[10] Other evaluations have examined *Investigations*, *Saxon*, and

Collectively Investigations, Saxon, and SFAW are used in about one-third of K-2 classrooms (Agodini and Harris 2010).

[10] I have replicated these results. Additionally, using this paper's estimation methods, and pooling grade levels, I estimate positive test score effects for *Expressions* and *Saxon* compared to *SFAW*, of 0.08σ and 0.10σ respectively (see Table 2 Column 6).

SWAF individually; they generally find no effects, but the counterfactuals are difficult to define (see the review in Agodini and Harris 2010).

2.2 The MKT test—a measure of teachers' math skills

Notably, among the data collected, each teacher's own math skills were measured with a pre-experiment test. The Mathematical Knowledge for Teaching (MKT) test is designed to measure both teachers' knowledge of mathematics per se and knowledge of pedagogy specific to teaching math (Hill, Shilling, and Ball 2004, Hill, Rowan, and Ball 2005). The latter skill includes, for example, "providing grade-level-appropriate but precise mathematical definitions, interpreting and/or predicting student errors, and representing mathematical ideas and procedures in ways learners can grasp" (Hill, Kapitula, and Umland 2011, p. 804).[11] An example MKT item is shown in Figure 1.[12]

2.3 Student test scores and other data

Students were tested pre- and post-experiment using math tests developed for the ECLS-K study.[13] Both the 1st grade and 2nd grade forms include questions in several areas—number sense, properties, and operations; measurement; geometry and spatial sense; data analysis, statistics, and probability; and patterns, algebra, and functions—though the weights differ by grade. Beginning with the scale scores, I standardize both pre and post scores (mean zero, standard deviation one) within grade.

The available data also include, as listed in Table 1, the several traditional student demographic characteristics, as well as many details on teacher demographics, experience, and

[11] While conceptually distinct, the math ability and math teaching skills sub-test scores are correlated 0.97, at least in the sample for this study. Throughout the paper I use the univariate overall MKT score.
[12] Additional information on the MKT and many example items are available at www.sitemaker.umich.edu/lmt.
[13] Not all students in were tested. Given cost constraints, a sample of students was randomly selected within each class for testing. In the results presented in the paper I weight student-level analyses by the inverse probability of selection. However, the results are robust to equal weighting, and those results are provided in the appendix

education. There are also data from classroom observations and teacher surveys, both designed to measure what activities occurred in study classrooms and what math content was covered during the school year. I describe these data as they arise in the analysis.

2.4 Baseline covariate balance and attrition

Causal interpretations of many results in this paper rely on the success of the original randomization to treatment conditions. Using the traditional test of random assignment, in Table 1 I compare the average pre-treatment characteristics of students, teachers, classes, and schools across the four conditions. The samples are well balanced. There is some evidence of differences in the proportion of teachers with a master's degree, but this is one of more than twenty characteristics tested.

My measurement of teacher productivity requires student observations with both pre- and post-experiment test scores. Thus, even if samples were balanced at baseline, differential attrition across conditions could bias my estimates. Since, as I describe shortly, teacher productivity is measured with student test score growth, attrition correlated with baseline is of particular concern. As shown in Table 2, there is little evidence of differences in attrition patterns: no differences in attrition rates in the four conditions (Columns 1-2), and no differences in the relationship between baseline test scores and likelihood of attrition (Column 3). Similarly, there is little evidence of differences in teacher attrition: no differences in response rates to the MKT test (Column 4), and no differences in attrition before the end of the experimental school year (Column 5).

3. Math skills, teacher productivity, and the effect of instructional methods

My first two empirical objectives are to (i) estimate the relationship between teachers' math skills, as measured by their MKT scores, and student test scores; and (ii) test whether the instructional methods teachers' are assigned to follow affect that relationship. As stated in these objectives, I focus on one aspect of teacher productivity: a teacher's contribution to student academic achievement as measured by test score growth. A large literature documents substantial variability in this aspect of productivity (Hanushek and Rivkin 2010), and recent evidence suggests that variability is predictive of teacher productivity differences measured with students' long-run economic and social outcomes (Chetty, Friedman, and Rockoff 2014b). In general, the data and my analysis cannot cover all aspects of teachers' job skills, job tasks, and job responsibilities; but the foci in this paper are first-order aspects of each category.

3.1 Teachers' MKT scores and student test scores

Teachers' MKT scores and their students' math test scores are positively correlated. This is apparent in Table 3 Column 1 which reports the result of a simple bivariate regression: student i's post-experiment math test score regressed on her teacher j's MKT score (both variables are standardized, and additional estimation details are described in the next section). Students assigned to a teacher with top-quartile math skills score about 0.055σ (student standard deviations) higher at the end of the school year than do their peers assigned to an average-skilled teacher.

However, that positive correlation may be explained by non-random assignment of students to teachers.[14] The regression reported in Table 3 Column 2 is identical to Column 1 except that I have added controls for student i's pre-experiment test score; specifically, a

[14] While the instructional methods treatment conditions were randomly assigned, students were not randomly assigned to teachers or classes.

quadratic in baseline test score where the parameters are allowed to differ in each grade-by-year cell. With this one control the correlation shrinks toward zero by more than half. Column 3 adds school fixed effects, and the point estimate is then essentially zero—0.009σ.[15] Empirical evidence from other settings suggests these two controls, prior test score and school fixed effects, are critical in accounting for between-school and within-school sorting of students to teachers (Kane and Staiger 2008, Chetty, Friedman, and Rockoff 2014a). Adding additional student, peer, and teacher characteristics as controls does not change the result, see Column 5.

The non-random assignment of students to teachers can also be seen by regressing student *i*'s *pre*-experiment math test score on her teacher *j*'s MKT score, as shown in Table 4 Column 1. As mentioned already, some student-teacher sorting is between-schools; accordingly the coefficient falls when school fixed effects are included in Column 2, but there is still evidence of within-school sorting. In short, better students are assigned to better, at least on the MKT dimension, teachers. This kind of within-school sorting has been documented elsewhere using on other observable teacher characteristics (Clotfelter, Ladd, and Vigdor 2005, Kalogrides, Loeb, and Beteille 2013).

3.2 The effects of assigned instructional methods

The evidence presented so far offers little support for the hypothesis that differences in teachers' knowledge of math content and pedagogy contribute to differences in student math

[15] While school fixed effects may be preferable to account for non-random sorting, most estimates presented in the paper use randomization block fixed effects instead of school fixed effects, primarily to permit the estimation of treatment condition main effects. In general the patterns of results presented in the paper are robust to using school fixed effects, for example compare Table 3 Columns 3 and 4. Estimates using school fixed effects are provided in the appendix. The robustness is likely due in part of the relatively small size of the randomization blocks. Moreover, compared to the large urban districts common in empirical research, this study's districts have fewer schools on average and thus less scope for sorting across-schools within a district.

learning.[16] However, the null average relationship described above may mask meaningful, but heterogeneous, relationships that depend on how teachers are asked to teach. For example, as suggested in the introduction, the day-to-day tasks of direct-instruction likely rely on teachers own math knowledge more frequently than the tasks of student-led methods.

The focus of this section is on estimating the effect of assigned instructional methods, the treatment conditions, on the relationship between teachers' skills and student learning. The treatment effect estimates, reported in Tables 3 and 5, are obtained by fitting variations of the following model:

$$A_{i,t} = f\left(MKT_{j(i)}\right) + C_{s(i)}\delta + h_{g(i),y(i)}\left(A_{i,t-1}\right) + X\beta + \tau_{b(s)} + \varepsilon_{i,t}, \tag{1}$$

where $A_{i,t}$ is the post-experiment end-of-school-year math test score for student i. Each student is observed in only one school year, in 1st or 2nd grade, assigned to one teacher j at school s. $C_{s(i)}$ is a vector of indicator variables for the treatment conditions, which were randomly assigned at the school level. The function h is a quadratic in pre-experiment beginning-of-school-year test score, $A_{i,t-1}$, interacted with grade-by-year indicators, allowing the quadratic parameters to vary on those dimensions. The vector X includes several student, peer, teacher, and school observable characteristics, notably among them teacher experience and an indicator for having experience with the assigned product previously.[17] Last, $\tau_{b(s)}$ is a series of fixed effects

[16] The coefficients on MKT score reported in Table 3 may be biased by measurement error or omitted variables. The former is likely not large, Hill and coauthors report relatively high reliability for MKT scores (Hill, Schilling, and Ball 2004, Hill, Umland, Litke, and Kapitula 2012).

[17] 11 percent of teachers had used their assigned product previously.

The complete list of covariates is, for student i, indicators for female, African-American, Latino, and English learner; and a quadratic in age. For teacher j, indicators for female, white, master's degree, having taken advanced math courses, having used the assigned product previously, and novice teacher; linear terms for years since MA degree and age; and quadratics in total experience, experience at the school, and professional development hours previous school year. These student and teacher variables are occasionally missing; for each student and teacher covariate replace missing values with zero and include an indicator variable $= 1$ for if the covariate is missing for the observation. The results presented in the paper are robust to excluding observations with missing values. For peers in teacher j's class with student i, linear terms for the mean and standard deviation of pre-experiment test

for the randomization blocks to account for the unequal probabilities of selection into treatment conditions.

Two additional notes on methods before discussing the estimates and their interpretation: First, throughout the paper, I report cluster-corrected standard errors which allow for correlation of $\varepsilon_{i,t}$ within schools (the unit of random assignment). Second, as described in Section 2, not all students in participating schools were tested; students were randomly sampled within classrooms. In all results presented, I weight by the inverse of the probability of selection. Estimates without weighting are shown in the appendix, and the pattern of results is the same.

The current research question can be thought of as comparing different ways to specify $f(MKT_{j(i)})$ in Model 1. The results discussed in the previous section are estimates where $f(MKT_{j(i)})$ is a single constant linear term, i.e., $f = \alpha * MKT_{j(i)}$; recall that $\hat{\alpha}$, the relationship between teachers' MKT scores and their students' test scores, was close to zero and statistically insignificant. Contrast those null results with the results in Table 3 Column 8 which uses the specification in Equation 1, and lets α vary by treatment condition, i.e., $f = MKT_{j(i)} * C_{s(i)} * \alpha$. For teachers in the *SFAW* condition $\hat{\alpha}$ is 0.035 (different from zero p-value = 0.071). For *Saxon* and *Expressions* teachers $\hat{\alpha}$ is also positive (different from zero p-values 0.134 and 0.126, respectively), but not statistically different from the *SFAW* estimate. However, for teachers in the *Investigations* condition $\hat{\alpha}$ is negative, significantly different from the other conditions, and significantly different from zero.

In other words, the methods of instruction teachers are asked to follow do affect whether and how teachers' math skills contribute to their students' learning growth. Among teachers

score. For school s, linear terms for the proportion of students eligible for free or reduced price lunch and title 1 eligible.

using primarily direct-instruction methods—the *SFAW* and *Saxon* conditions—there is an apparent positive relationship: students assigned to a teacher with better math skills do score higher on math tests at the end of the school year. This also holds for teachers using *Expressions*, which combines both direct-instruction and student-led methods. In an important contrast, among teachers using primarily student-led methods—the *Investigations* condition—there is an apparent *negative* relationship: students assigned to a teacher with better math skills score lower in the end.

What of these results can be interpreted causally? In short, the individual slope estimates, the $\hat{\alpha}$ for each treatment condition, should not be given a strong causal interpretation; but the pattern of differences between the slopes can be interpreted as a causal effect of the assigned instructional methods. Regarding the slopes, I have no new identification strategy for the slope estimates. Nevertheless, my estimates do address critical sources of potential bias. First, recent empirical evidence suggests bias arising from non-random student-teacher sorting is well addressed by including controls, as I do in Model 1, for students' prior test score, school fixed effects, and other commonly available covariates (Kane and Staiger 2008, Kane, McCaffrey, Miller, and Staiger 2013, Chetty, Friedman, and Rockoff 2014a). Still, even if students were randomly assigned to teachers, studying any single measure of teacher skill, like MKT scores, remains subject to omitted variable bias. Second, I can control for a much richer set of teacher characteristics than is usually possible. Encouragingly, the point estimates of interest are not substantially different with and without this rich set of controls; Table 3 Column 7 without any controls is quite similar to Column 8, and the same is true for several other permutations of included teacher covariates not reported here.

The pattern of differences between slopes can be interpreted causally under the standard experimental assumption: At the start of the experiment, there was no difference across the treatment conditions in potential outcomes, including student math achievement and teacher productivity. Stated differently: Any source of bias in estimating the α terms is independent of assigned treatment condition. This assumption rests on the random assignment of schools which, as discussed in Section 2, appears to have been successful. Table 4 Column 5 provides some additional evidence of successful random assignment. Notably, while students may be sorted to teachers, so that baseline test scores are correlated with teacher MKT scores, that form of sorting does not appear to be different across schools assigned to different treatment conditions.

But random assignment only identifies the differences in slopes. Thus, for example, it may be that the true relationship between teacher math skills and student math learning is positive for all four treatment conditions, but biased downward by some omitted variable so that the slope for *Investigations* appears to be negative. Even if that example were the case, our causal interpretation of the difference in slopes would not change: when using strongly student-led methods teachers' math skills contribute less to student learning than when using direct-instruction methods.

To summarize, ignoring differences in teachers job tasks—how they are asked to teach—can easily generate the misleading result that teachers' math skills do not contribute to student math achievement growth. That null estimate, however, masks important differences that depend on teachers assigned job tasks. When using direct-instruction methods, teachers with more knowledge of math concepts and pedagogy produce students with more math knowledge, but teachers' knowledge contributes less to student learning when teachers are asked to use student-

led methods. Indeed, the relationship between teacher skills and student learning may be negative in schools that use student-led instructional methods.

3.3 Heterogeneity across the distribution of teacher skills

I next investigate whether the effects of instructional methods on teacher productivity depend on teachers' prior skills. To this point the analysis has assumed linear relationships between teachers' math skills and student learning outcomes, and thus also assumed the treatment effect is a constant shift in the slope. It turns out that, as I describe in this section, there is important heterogeneity in treatment effects across the distribution of teacher skills.

To test for heterogeneity I first (i) divide teachers into three equal groups based on their MKT score rank, and then (ii) estimate treatment effects within those terciles. In the language of Equation 1, I replace the linear term, $MKT_{j(i)}$, in f with three indicator variables for each MKT tercile. The results, all drawn from a single regression, are reported in Table 5 and plotted in Figure 2.[18] Again, to be precise, comparisons between teachers using different instructional methods can be interpreted causally given random assignment; comparisons between teachers of different skill levels should not be given the same strong causal interpretation.

Several important patterns are evident in Figure 2. First, the productivity of both low-skilled and average-skilled teachers evidently does not depend on the instruction method they are asked to follow. The exception to that pattern is the strongly direct-instruction *SFAW*, but the *SFAW* deficit holds for nearly all teachers regardless of math skill level. Moreover, low-skilled teachers appear equally as good as their average-skilled colleagues at producing student achievement growth.

[18] An alternative approach to analysis is to replace the linear term $MKT_{j(i)}$ in f with a higher-order polynomial. Figure 2 suggests a quadratic where the parameters are allowed to vary by treatment condition. Results using a quadratic are provided in the appendix. Adding higher order terms beyond a quadratic do not improve the model, at least as judged by likelihood ratio tests.

In stark contrast, instructional methods do affect the productivity of high-skilled teachers. High-skilled teachers generate noticeably more student math learning using direct-instruction methods then they would generate using student-led methods. This is clear comparing student scores in either the *Expressions* or *Saxon* conditions to *Investigations*. Even *SFAW* tops *Investigations* among high-skilled teachers, though the difference is not statistically significant. Indeed, while high-skilled teachers apparently out perform their average- and low-skilled colleagues when using direct-instruction, the opposite is true when teachers use the strongly student-led methods of *Investigations.*

For high-skilled teachers the consequences are large. Students of high-skilled teachers using direct-instruction methods can score 0.13-0.16σ higher than their peers assigned to equally high-skilled teachers using student-led methods (Table 5 Column 7). A difference of 0.13-0.16σ is roughly equivalent to the standard deviation in total teacher productivity (Hanushek and Rivkin 2010).

3.4 Potential mechanisms

Higher returns to math skills for teachers using direct-instruction, as reported above, are consistent with the differences in teacher tasks between direct-instruction and student-led methods—in particular differences in the extent to which teachers explicitly teach math concepts and procedures to their students. Direct-instruction methods rely on explicit teaching far more frequently than student-led methods. Thus direct-instruction should be more successful the better the teacher understands math herself, and even more successful if she understands the typical ways students (mis)understand math concepts and procedures. The MKT primarily tests these two kinds of knowledge, as illustrated by the sample question in Figure 1.[19] Put differently, high-

[19] MKT scores also capture variation in teachers' own test-taking skills for standardized math tests, and teachers may be imparting those skills to their students as well.

MKT teachers have the ability to answer math problems correctly using standard procedures; direct-instruction gives them many opportunities to demonstrate and explain those skills to their students, while such opportunities are infrequent with student-led methods.

Teachers' *assigned* tasks are, of course, not necessarily the tasks they actually *do* from day to day. The previous paragraph assumes treatment assignment generated meaningful, empirical differences in the extent to which teachers used direct-instruction or student-led methods. I test for differences in teacher behavior across the four treatment conditions using data collected during classroom observations.[20] Trained observers spent, on average, about 1.5 hours in each teacher's classroom (mean 81 minutes, standard deviation 40), recording the frequency of dozens of specific teacher practices and behaviors. For example, observers tallied the number of times the teacher "tells information [or] models procedures" and "probes for [a student's] reasoning or justification of a solution." I focus here on two summary measures derived from these micro-data using factor analysis: teacher behavior characteristic of (i) student-led methods and (ii) direct-instruction methods. These are the first and second predicted factors, which together explain nearly two-thirds of the variation in the observation data (39 and 24 percent of the variation respectively).[21] All the included measures and factor weights are detailed in the appendix. Among the highest-weighted items in the "student-led" factor are "probes for [a student's] reasoning or justification of a solution," "poses open-ended questions," and "elicits multiple strategies/solutions." For the "direct-instruction" factor the highest-weighted items

[20] Complete details regarding the classroom observations are provided in Agodini et al. (2010). Among those details, first, Agodini and coauthors report evidence of strong inter-rater reliability. Second, teachers were randomly selected for classroom observation from among all study teachers: 82 percent of 1st grade teachers and 90 percent of 2nd were selected for observation, with response rates of 96 and 91 percent respectively.

[21] The original analysis by Agodini et al. (2010) also involved a similar factor analysis; the results are comparable to the factor loadings reported here (see their Table C.2) including a first "student-centered" factor and second "teacher-directed" factor.

include “tells information [or] models procedures,” “guides practice on problems,” and “states if [student answer is] correct or not without elaborating.”

Table 6 summarizes relevant variation in these two measures of student-led and direct-instruction teacher behavior. I estimate least squares regressions, similar to those in Table 3, with the specification

$$Y_j = f\left(MKT_j\right) + C_{s(j)}\delta + X\beta + \tau_{b(s)} + \nu_j, \tag{2}$$

where Y_j is either the (i) student-led factor score for teacher j or the (ii) direct-instruction factor score drawn from the classroom observation data. In both cases Y_j is standardized (mean zero, standard deviation one) within the sample. The right hand terms are as before (see Specification 1), and standard errors are clustered at the school level. When the additional controls, X, include student variables, those variables are the mean characteristic among teacher j’s students.

The extent of direct-instruction and student-led behavior by teachers, as observed in the classroom, depends on how teachers are asked to teach (assigned treatment condition) but not on their math skills (MKT score). First, as reported in the top rows of Table 6 Columns 1-4, there are large and statistically significant differences between treatment conditions in both student-led and direct-instruction behavior. For example, teachers assigned to *Investigations* used direct-instruction methods a full standard deviation less often than *SFAW* teachers, and half a standard deviation less than *Expressions* or *Saxon* teachers. However, there is little evidence that either student-led or direct-instruction behaviors are correlated with teachers’ math skills. Pooling across conditions, the coefficient on MKT score for student-led behavior is 0.016 with a standard error three times as large. The same coefficient for direct-instruction is -0.023. Moreover, even if there is a relationship between skills and these two behaviors, that relationship is not affected by

treatment assignment. I cannot reject the hypothesis that the four slope coefficients are equal, nor the hypothesis that the four are jointly zero.

In short, teachers' assigned tasks did change how they taught day to day, but those changes did not depend on the teachers' differing math skills. Thus, differences in teachers' adherence to assigned tasks cannot explain the productivity effects seen in Table 3 or Figure 2, at least given the evidence from these classroom observation measures.

While the two measures in Table 6 Columns 1-4 are the actions mostly likely affected, treatment assignments may have (unintentionally) affected other aspects of teachers' jobs. Of particular interest are "other aspects" correlated with teachers' math skills. As a contrast with Columns 1-4, in Columns 5-6 I provide a measure of classroom environment also drawn from the classroom observation data. Separate from the teacher practices and behaviors described above, observers also rated the classroom environment on about 30 characteristics. Items including "student behavior disrupts the classroom," "class time is spent on understanding or practicing math," and "teacher has techniques for gaining class attention in less than 10 sections" are each scored from "1 = not at all characteristic (almost never)" to "4 = extremely characteristic (almost always evident)". Columns 5-6 report results using the first predicted factor, standardized, from a factor analysis of these environment items.[22] Treatment assignment did not affect observed classroom environment, neither in levels nor in the relationship with teachers' skills.

[22] This first factor accounts for 62 percent of the variation in these environment items. The full list of items and factor loadings are reported in the appendix. The results are quite similar using the simple average of the 31 items, after reversing the scale of negatively framed items.

4. Total effects on teacher productivity

The instructional methods schools adopt may, plausibly, affect teacher productivity through mechanisms unrelated to teachers' math skills. Section 3 focuses on differences in productivity arising from interactions between skills and methods. In this section I situate those skill-related effects in the broader context. Empirically, I first estimate the effect of treatment—the contrasts between instructional methods—on *total* teacher productivity. I focus on treatment effects on the variance of productivity. Then, second, I examine whether the total effect is explained by the mechanisms related to teachers' math skills.

4.1 Estimating treatment effects on the variance of total teacher productivity

My first empirical objective is to estimate the variance of teacher productivity in each treatment condition, and test for differences between conditions. Throughout the paper I focus on one aspect of productivity: a teacher's contribution to student math achievement as measured by test score growth.

A teacher's contribution to her students' test scores is not directly observable. To isolate the teacher's contribution, I first assume a statistical model of student test scores, similar to the model in Equation 1, where a test score, $A_{i,t}$, for student i at the end of the experiment school year t can be written

$$A_{i,t} = h_{g(i),y(i)}\left(A_{i,t-1}\right) + X\beta + \psi_{s(i)} + \mu_{j(i)} + \nu_{i,t}, \tag{3}$$

The $\mu_{j(i)}$ term represents the effect of teacher j on student i's test score; net of baseline achievement, $h_{g(i),y(i)}\left(A_{i,t-1}\right)$, other student characteristics, X, and school effects, $\psi_{s(i)}$. The vector X includes all of the student variables described above for Model 1, but does not include any teacher or school characteristics. Specification 3, now commonplace in the literature on teachers, is motivated by a dynamic model of education production, suggested by Todd and

Wolpin (2003), in which prior test score, $A_{i,t-1}$, is a sufficient statistic for differences in prior inputs.

With the model in 3 as a key building block, I take two separate approaches to estimating treatment effects on the variance of teacher productivity. The first approach is a least-squares estimate of the conditional variance function. Specifically, I estimate the pairwise differences in variance between conditions, $\hat{\gamma}^{LS}$, by fitting

$$\left(\mu_j - \mathbb{E}\left[\mu_j \middle| C_{s(j)}, \tau_{b(s)}\right]\right)^2 = C_{s(j)}\gamma^{LS} + \tau_{b(s)} + u_j, \tag{4}$$

where, just as before, $C_{s(i)}$ is a vector of indicator variables for the treatment conditions, which were randomly assigned at the school level; and $\tau_{b(s)}$ represent fixed effects for each randomization block group, b.[23]

My approach to estimating Model 4 has three steps. Step one, estimate μ, as described in the next paragraph. Then follow the common, feasible approach to fitting conditional-variance equations like 4: Step two, estimate $\mathbb{E}\left[\hat{\mu}_j \middle| C_{s(j)}, \tau_{b(s)}\right]$ by ordinary least-squares, i.e., fit $\hat{\mu}_j = C_{s(j)}\tilde{\gamma} + \tilde{\tau}_{b(s)} + \epsilon_j$. Step three, estimate Equation 4 using the squared residual from step two, $\hat{\epsilon}_j^2$, as the dependent variable. I calculate standard errors for $\hat{\gamma}^{LS}$ that allow for clustering within schools.

In step one I estimate the test-score productivity of each teacher, $\hat{\mu}_j$, by fitting Equation 3 treating the $\mu_{j(i)}$ as teacher fixed effects.[24] The $\psi_{s(i)}$ terms are school fixed effects, and $h_{g(i),y(i)}$ is, as before, a quadratic in pre-experiment test score. The parameters of $h_{g(i),y(i)}$ are allowed to be different for each combination grade, g, and experiment year, y. Note that this teacher-fixed-

[23] Model 4 follows from the general observation that $var(Y|X) = \mathbb{E}[(Y - \mathbb{E}[Y|X])^2|X]$.
[24] The teacher fixed effects are parameterized to be deviations from the school average, rather than deviations from an arbitrary hold out teacher, using the approach suggested by Mihaly, McCaffrey, Lockwood, and Sass (2010).

effects approach does not require a distributional assumption about $\mu_{j(i)}$, and identifies other model parameters using only within-teacher variation.

The second approach to estimating γ is a maximum likelihood estimate, $\hat{\gamma}^{ML}$, obtained by treating $\mu_{j(i)}$ as teacher random effects. I fit a slightly re-parameterized version of Equation 3,

$$A_{i,t} = h_{g(i),y(i)}\left(A_{i,t-1}\right) + X\beta + \psi_{s(i)} + \mu_{j(i)}^{INV} INV_{s(i)} + \mu_{j(i)}^{EXPR} EXPR_{s(i)} + \mu_{j(i)}^{SAX} SAX_{s(i)} + \mu_{j(i)}^{SFAW} SFAW_{j(i)} + \nu_{i,t}, \quad (5)$$

where $INV_{s(i)}$, $EXPR_{s(i)}$, $SAX_{s(i)}$, and $SFAW_{j(i)}$ are treatment condition indicators, and the four μ_j terms are random effects with the assumed distribution

$$\begin{bmatrix} \mu_{j(i)}^{INV} \\ \mu_{j(i)}^{EXPR} \\ \mu_{j(i)}^{SAX} \\ \mu_{j(i)}^{SFAW} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu^{INV} \\ \mu^{EXPR} \\ \mu^{SAX} \\ \mu^{SFAW} \end{bmatrix}, \begin{bmatrix} \sigma_{INV}^2 & & & \\ 0 & \sigma_{EXPR}^2 & & \\ 0 & 0 & \sigma_{SAX}^2 & \\ 0 & 0 & 0 & \sigma_{SFAW}^2 \end{bmatrix} \right).$$

That is, the model allows the estimated variance of the teacher-specific random intercepts to differ for each treatment condition. All other covariates are identical to the least squares approach. Maximum likelihood estimation of this linear mixed model provides estimates of the σ^2 terms, and thus $\hat{\gamma}^{ML} = \{(\hat{\sigma}_{INV}^2 - \hat{\sigma}_{EXPR}^2), (\hat{\sigma}_{INV}^2 - \hat{\sigma}_{SAX}^2), \ldots, (\hat{\sigma}_{SAX}^2 - \hat{\sigma}_{SFAW}^2)\}$.

To interpret estimates from either approach, $\hat{\gamma}^{ML}$ or $\hat{\gamma}^{LS}$, as causal effects—the effect of adopting one adopting instructional method instead of another—requires two assumptions.

Assumption 1: At the start of the experiment, there were no differences across treatment conditions in teachers' potential productivity during the experiment school year. This assumption should be satisfied by the random assignment study designs.

Assumption 2: Students were not assigned to teachers based on unobserved (i.e., omitted from Specifications 3 or 5) determinants of potential for test score growth: $\mathbb{E}\left[\nu_{i,t} \mid j\right] = \mathbb{E}\left[\nu_{i,t}\right]$.

This assumption is necessary for obtaining consistent estimates of $\hat{\mu}_j$, and parameters like it throughout the teacher effects literature. Empirical tests of this assumption by Chetty, Friedman, and Rockoff (2014a) and Kane and Staiger (2008) find little residual bias in $\hat{\mu}_j$ if the estimating equation includes, as I do, flexible controls for students' prior achievement, and controls for teacher and student sorting between schools.[25]

Assumption 2 is, strictly speaking, only needed to identify the *levels* of variance. A weaker alternative is sufficient for causal estimates of the *relative difference* in variance, and thus the sign of $\hat{\gamma}^{ML}$ or $\hat{\gamma}^{LS}$. *Assumption 2 Alternative:* Any source of (residual) bias in estimating $\hat{\mu}_j$ is independent of assigned treatment condition. Like Assumption 1, this alternative assumption should be satisfied by random assignment.

4.2 Measuring the influence of skill-related mechanisms

My follow-up empirical question is: To what extent are the total productivity effects explained by mechanisms related to teachers' math skills? To answer that question I repeat the estimation methods described in the previous section with one modification: in Equations 4 and 5 I add controls for a quadratic in MKT_j allowing the parameters of the quadratic to differ for each of the four treatment conditions. As an alternative modification, I replace the quadratic terms with indicators for MKT tercile.

4.3 Results and discussion

Asking teachers to use student-led instruction methods, instead of more-conventional direct-instruction methods, reduces the differences between teachers in student learning. That is,

[25] For detailed discussions of the theoretical and econometric issues in isolating teacher contributions to student test score growth see Todd and Wolpin (2003), Kane and Staiger (2008), Rothstein (2010), and Chetty, Friedman, and Rockoff (2014a). Rothstein (2010), in particular, provides a skeptical analysis, and raises some concerns not yet resolved or tested empirically.

as reported in Table 7 Column 1, the variance of teacher productivity is smallest in the student-led *Investigations* condition. Focusing on the maximum-likelihood random-effects estimates (top panel), the estimated standard deviation of teacher effects is 0.08σ for teachers using *Investigations*, which is half or less as large as the between-teacher standard deviation for *Expressions* 0.19σ or *Saxon* 0.16σ. Both of those differences are statistically significant. The difference between *Investigations* and *SFAW* is smaller and not as precisely estimated, but is in the same direction.

This pattern of results is robust to alternative approaches to estimation. First, the pattern is the same using the conditional-variance of teacher fixed-effects method (bottom panel), though the estimated differences between conditions are smaller. Second, the pattern is also robust to replacing the school fixed effects, $\psi_{s(i)}$ in Specification 3 or 5, with controls for treatment condition main effects and the available school characteristics listed in Section 3. Results for this second alternative are provided in the appendix. Broadly speaking the estimates in Table 7 and the appendix span the range of teacher effect variances commonly estimated in other settings (Hanushek and Rivkin 2010).

The magnitudes of differences across instructional methods conditions are educationally substantial. Consider the MLE point estimates. In the widely-used direct-instruction *Saxon* classrooms, students assigned to a teacher at the 75th percentile of the teacher performance distribution will score approximately 0.11σ higher on math achievement tests than their peers assigned to a median teacher. By contrast, in the student-led *Investigations* classrooms a student's teacher assignment is much less consequential. The median to 75th percentile difference is just 0.05σ. These results can be read as greater "equity" of outcomes across teachers

and classrooms, in a sense; but, as suggested in Figure 2, that greater equity comes in part at the expense of the students assigned to high-skilled teachers.

The estimated variances are not substantively different when I add controls for teachers' skills. Comparing Table 7 Column 1 to either Columns 6 or 7, in some cases adding MKT controls reduces the variance estimate, in other cases the estimate increases, and in most cases the changes are only a few percent in magnitude. Additionally, the pattern of differences in variance between treatment conditions does not change. First, these results are a reminder that teachers' assigned instructional methods can affect productivity through mechanisms unrelated to their math knowledge, as measured by MKT. Student-led methods, for example, rely on teachers' verbal and communication skills more, perhaps, than direct-instruction. As a second example, for some tasks *SFAW* provides a script for teachers to read, which may contribute to the lower variance for *SFAW* relative to the other direct-instruction conditions. Second, though the estimated variance does not change much after controlling for MKT scores, nevertheless the interaction between teachers' math skills and assigned instructional methods does affect the estimated rank ordering of teachers, as shown in Section 3.

5. Conclusion

In this paper I show that the job tasks teachers are assigned—the instructional methods they are asked to use in the classroom—partly determine the returns to teacher skills in education production, and partly determine the variability in teacher productivity more generally. These results are one example of the value in making a distinction between workers' skills and the job tasks to which those skills are applied, as in Acemoglu and Autor (2011, 2012).

Using data from a field-experiment in 1st and 2nd grade classes, I first examine the relationship between teachers' math skills, measured by the Mathematical Knowledge for Teaching (MKT) test, and teacher productivity, measured by teachers' contributions to their students' test score growth. That relationship is positive and educationally meaningful when teachers are asked (by random assignment) to use conventional "direct-instruction" methods. But, in stark contrast, the relationship is much weaker, perhaps even negative, when teachers are asked to use "student-led" instructional methods. This difference in the returns to skills is largely driven by high-skilled teachers. In the classrooms of top-tercile MKT teachers, students' math scores grow 0.13-0.16σ faster when the teacher uses direct-instruction instead of student-led methods. However, there is little or no difference between instructional methods in the classrooms of bottom- or middle-tercile MKT teachers. In short, whether and how a teacher's math skills contribute to her productivity depends on how she is asked to teach math.

Second, I show that teaching tasks or methods can substantially shrink (expand) the variation in teacher productivity. The standard deviation of productivity among teachers using direct-instruction is 0.12-0.19 student standard deviations, but at least one-third smaller, 0.08, for teachers using student-led methods. These are differences in total productivity, not just differences that arise through math-skill-related mechanisms; thus these differences suggest teachers assigned tasks operate on productivity through other skills or mechanisms.

Understanding how skills and job tasks translate into productivity is especially relevant and timely in public schools. In recent years, differences in teacher productivity have become central to political and managerial efforts to improve public schools. This paper's results suggest that decisions about *how* teachers are asked to teach can be as important as decisions about *who* is hired to teach.

References

Acemoglu, D. & Autor, D. (2011). Skills, Tasks and Technologies: Implications for Employment and Earnings. *Handbook of Labor Economics, 4,* 1043-1171.

Acemoglu, D. & Autor, D. (2012). What Does Human Capital Do? A Review of Goldin and Katz's *The Race between Education and Technology. Journal of Economic Literature, 50*(2): 426-463.

Acemoglu, D. & Zilibotti, F. (2001). Productivity Differences. *Quarterly Journal of Economics, 116*(2): 563-606.

Agodini, A., & Harris, B. (2010). An Experimental Evaluation of Four Elementary School Math Curricula. *Journal of Research on Educational Effectiveness, 3*(3), 199-253.

Agodini, A., Harris, B., Thomas, M., Murphy, R., & Gallagher, L. (2010). *Achievement Effects of Four Early Elementary School Math Curricula: Findings for First and Second Graders.* (NCEE 2011-4001). Washington, D.C.: Institute for Education Sciences, U.S. Department of Education.

Autor, D., Levy, F., & Murnane, R. J. (2003). The Skill Content of Recent Technological Change: An Empirical Exploration. *The Quarterly Journal of Economics, 118*(4), 1279-1333.

Becker, G. S. (1964). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education.* Chicago: University of Chicago Press.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review, 104*(9), 2593-2632.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the Impacts of Teachers II: Teacher Value-Added and Outcomes in Adulthood. *American Economic Review, 104*(9), 2633-2679.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. (2005). Who Teaches Whom? Race and the Distribution of Novice Teachers. *Economics of Education Review 24,* 377-392.

Cook, J. B., & Mansfield, R. K. (2014). Task-Specific Experience and Task-Specific Talent: Decomposing the Productivity of High School Teachers. Cornell University, ILR working paper 173.

Costinot, A. & Vogel, J. (2010). Matching and Inequality in the World Economy. *Journal of Political Economy, 118*(4), 747-786.

Dobbie, W. (2011). Teacher Characteristics and Student Achievement: Evidence from Teach for America. Harvard University Working Paper.

Hanushek, E. A. (1971). Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data. *American Economic Review*, 61(2), 280–288

Hanushek, E. A. (1997). Assessing the Effects of School Resources on Student Performance: An Update. *Educational Evaluation and Policy Analysis*, *19*(2), 141-164.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations About Using Value-added Measures of Teacher Quality. *The American Economic Review*, *100*(2), 267-271.

Hill, H. C., Kapitula, L. R., & Umland, K. (2011). A Validity Argument Approach to Evaluating Teacher Value-Added Scores. *American Educational Research Journal*, *48*(3), 794-831.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement. *American Educational Research Journal*, *42*(2), 371-406.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing Measures of Teachers' Mathematics Knowledge for Teaching. *Elementary School Journal*, *105*(1), 11-30.

Hill, H.C., Umland, K., Litke, E., & Kapitula, L. R. (2012). Teacher Quality and Quality Teaching: Examining the Relationship of a Teacher Assessment to Practice. *American Journal of Education*, *118*(4), 489-519.

Jackson, C. K. (2013). Non-Cognitive Ability, Test Scores, and Teacher Quality: Evidence from 9th Grade Teachers in North Carolina. National Bureau of Economic Research 18624.

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

Jacob, B. A., Rockoff, J. E., Taylor, E. S., Lindy, B., & Rosen, R. (2015). Teacher Applicant Hiring and Teacher Performance: Evidence from DC Public Schools. Working paper.

Kalogrides, D., Loeb, S., & Beteille, T. (2013). Systematic Sorting: Teacher Characteristics and Class Assignments. *Sociology of Education*, *86*(2), 103-123.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment.* Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T. J. & Staiger, D. O. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. National Bureau of Economic Research 14607.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying Effective Classroom Practices Using Student Achievement Aata. *Journal of Human Resources*, *46*(3), 587-613.

Mihaly, K., McCaffrey, D. F., Lockwood, J. R., & Sass, T. R. (2010). Centering and Reference Groups for Estimates of Fixed Effects: Modifications to felsdvreg. *Stata Journal*, *10*(1), 82-103.

Murnane, R. J. (1975). *The Impact of School Resources on the Learning of Inner City Children.* Cambridge, MA: Ballinger.

Ost, B. (2014). How Do Teachers Improve? The Relative Importance of Specific and General Human Capital. *American Economic Journal: Applied Economics*, *6*(2), 127-151.

Papay, J. P. & Kraft, M. A. (forthcoming). Productivity Returns to Experience in the Teacher Labor Market: Methodological Challenges and New Evidence on Long-Term Career Improvement. *Journal of Public Economics.*

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools and Academic Achievement. *Econometrica, 73*(2), 417–458.

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review, 94*(2), 247–252.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can You Recognize an Effective Teacher When You Recruit One? *Education Finance and Policy, 6*(1), 43-74.

Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics*, *125*(1), 175-214.

Stein, M. K., & Kaufman, J. H. (2010). Selecting and Supporting the Use of Matehematics Curricula at Scale. *American Educational Research Journal, 47*(3), 663-693.

Taylor, E. S. (2015). New Technology and Teacher Productivity. Working paper.

Tinbergen, J. (1974). Substitution of Graduate by Other Labour. *Kyklos, 27*(2), 217-226.

Todd, P. E. & Wolpin, K. I. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *The Economic Journal*, *113*(485), F3-F33.

2. Imagine that you are working with your class on multiplying large numbers. Among your students' papers, you notice that some have displayed their work in the following ways:

Student A	*Student B*	*Student C*
35 ×25 125 +75 875	35 ×25 175 +700 875	35 ×25 25 150 100 +600 875

Which of these students would you judge to be using a method that could be used to multiply any two whole numbers?

	Method would work for all whole numbers	Method would NOT work for all whole numbers	I'm not sure
a) Method A	1	2	3
b) Method B	1	2	3
c) Method C	1	2	3

FIGURE 1—EXAMPLE MKT TEST ITEM

Note: Reproduced from Hill, Shilling, and Ball (2004, p. 28). This item has been publically released, but it was not necessarily included in the MKT test form used in this experiment.

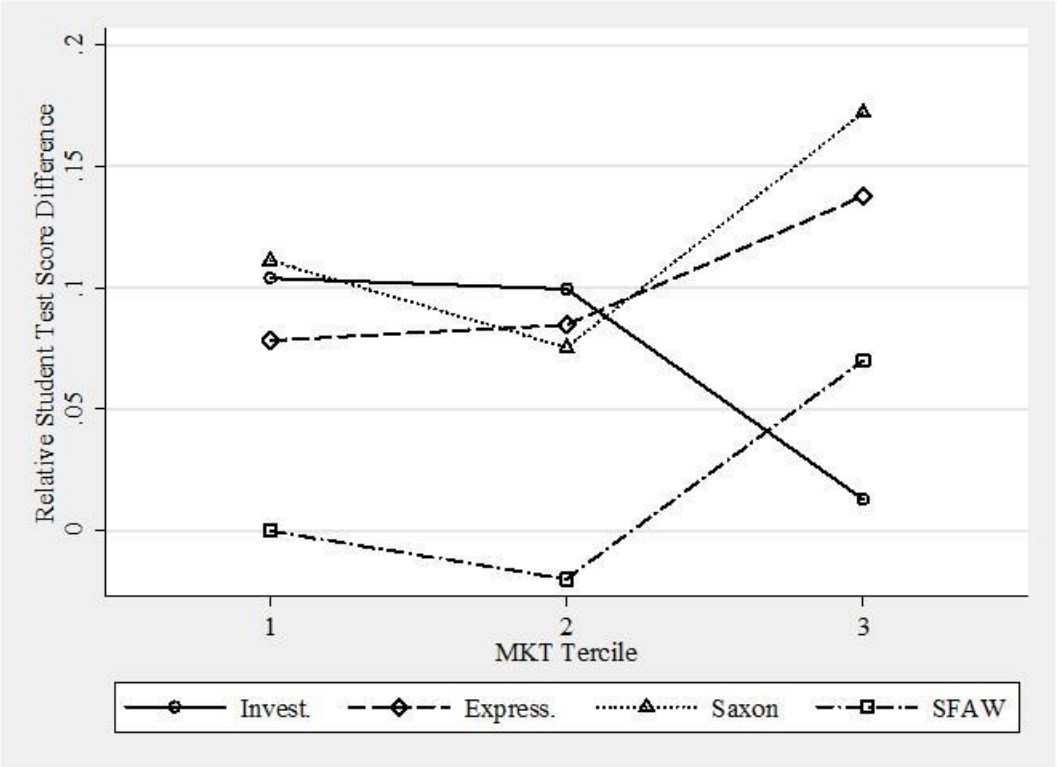

FIGURE 2—STUDENT TEST SCORE DIFFERENCES ACROSS TREATMENT CONDITIONS AND TEACHER MKT TEST SCORE TERCILES

NOTE: Each point is the estimated mean post-experiment math test-score for students in the given MKT-tercile-by-curriculum cell, relative to the mean score of students in the *SFAW* condition assigned to bottom-tercile teachers. Lines connect estimates in the same curriculum treatment condition. All points are estimated in a single regression (the same regression reported in Table 5). The dependent variable is the student's post-experiment standardized ECLS-K math test score. The key independent variables are a vector of indicators, one indicator for each MKT-tercile-by-curriculum combination. Independent variables also include the full set of student, teacher, and peer pre-experiment controls as in Table 5, and randomization block fixed effects. Students were randomly sampled within classrooms, and these results are weighted by the inverse probability of selection. The estimation sample includes 7,650 students, 750 teachers, and 110 schools. Sample sizes have been rounded to nearest 10 following NCES restricted data reporting procedures.

TABLE 1—STUDENT, TEACHER, CLASS, AND SCHOOL CHARACTERISTICS

	Assigned curricula (experimental condition)				Joint test		Category joint test
	Invest.	Express.	Saxon	SFAW	p-value	Obs.	p-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Student characteristics							0.704
Baseline test score, mean	0.009	-0.045	0.012	0.021	0.587	8,850	
	(0.968)	(0.967)	(0.972)	(0.959)			
Baseline test score, variance	0.943	0.941	0.943	0.907	0.760	8,850	
Days between pre- and post-experiment tests	238.0	236.6	236.5	236.9	0.868	8,010	
	(9.055)	(8.385)	(6.480)	(6.609)			
Age	7.016	7.012	7.022	7.009	0.957	7,690	
	(0.626)	(0.654)	(0.622)	(0.631)			
Female	0.498	0.479	0.486	0.493	0.587	8,600	
Latino	0.294	0.280	0.311	0.324	0.632	8,110	
African-American	0.262	0.338	0.278	0.250	0.161	8,110	
English language learner	0.100	0.132	0.140	0.151	0.196	7,510	
Teacher and class characteristics							0.171
MKT score, mean	-0.503	-0.602	-0.604	-0.524	0.148	750	
	(0.477)	(0.483)	(0.462)	(0.490)			
MKT score, variance	0.224	0.231	0.212	0.243	0.746	750	
Total years experience	12.930	12.278	12.078	11.548	0.456	760	
	(8.723)	(9.872)	(9.628)	(9.313)			
Years experience current school	7.969	7.606	7.442	8.171	0.816	720	
	(7.004)	(7.679)	(6.719)	(7.694)			
Female	0.948	0.971	0.969	0.961	0.685	770	
White	0.611	0.616	0.602	0.588	0.880	770	
Master's degree	0.492	0.423	0.474	0.393	0.028	740	
Years with master's degree	4.544	3.821	3.935	3.314	0.312	770	
	(7.367)	(7.312)	(6.051)	(6.189)			
Training hours previous school year	9.501	9.826	8.000	8.753	0.866	730	
	(20.52)	(21.00)	(20.90)	(19.16)			
One or more adv. math courses	0.491	0.514	0.516	0.558	0.654	770	
Class mean baseline test score	0.003	-0.057	0.010	0.015	0.503	790	
	(0.390)	(0.402)	(0.447)	(0.384)			
Class st. dev. baseline test score	0.885	0.891	0.878	0.879	0.929	790	
School characteristics							0.255
Proportion eligible for free or reduced price lunch	0.525	0.471	0.466	0.513	0.222	110	
	(0.132)	(0.102)	(0.109)	(0.120)			
Proportion Title I	0.756	0.756	0.749	0.730	0.992	110	
	(0.265)	(0.232)	(0.315)	(0.358)			

Note: Means (standard deviations) adjusted for randomization block fixed effects. Column 5 tests the null hypothesis that the four curriculum condition means are equivalent for the given pre-treatment characteristic. Column 7 tests the null hypothesis that each set of four means is equivalent for all pre-treatment characteristics within the category. Sample sizes have been rounded to nearest 10 following NCES restricted data reporting procedures.

TABLE 2—STUDENT AND TEACHER ATTRITION

	Student attrited before…			Teacher attrited before…	
	pre-test	post-test		MKT	post-test
	(1)	(2)	(3)	(4)	(5)
Curricula (relative to SFAW)					
Investigations	-0.001	0.005	0.004	0.000	-0.019
	(0.010)	(0.012)	(0.012)	(0.026)	(0.015)
Math Expressions	-0.003	0.006	0.004	0.004	0.028
	(0.007)	(0.013)	(0.013)	(0.024)	(0.022)
Saxon	0.004	0.008	0.008	0.031	0.005
	(0.010)	(0.011)	(0.011)	(0.025)	(0.017)
Baseline test score (main effect)			-0.021**		
			(0.006)		
Baseline score * Investigations			0.007		
			(0.009)		
Baseline score * Expressions			0.005		
			(0.008)		
Baseline score * Saxon			0.001		
			(0.009)		
Observations	8,990	8,820	8,820	790	790
Dependent variable sample mean	0.019	0.092	0.092	0.056	0.028

Note: Each column represents a separate LPM regression with student (Columns 1-3) or teacher (Columns 4-5) observations. Dependent variables are indicators as described in the column headers. All independent variables are as shown above, plus fixed effects for randomization blocks. Standard errors in parentheses. Sample sizes have been rounded to nearest 10 following NCES restricted data reporting procedures.
+ indicates $p<0.10$, * $p<0.05$, and ** $p<0.01$

TABLE 3—TEACHER MATH SKILLS, INSTRUCTION PRACTICES, AND STUDENT ACHIEVEMENT
(dep. var. = post-experiment ECLS-K math test score, standardized)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
MKT score	0.082**	0.036*	0.009	0.006	0.008	0.011	0.029+	0.035+
	(0.021)	(0.014)	(0.010)	(0.010)	(0.010)	(0.010)	(0.017)	(0.019)
Curricula main effects (relative to SFAW)								
Investigations						0.054	0.053	0.063+
						(0.037)	(0.036)	(0.035)
Expressions						0.080*	0.085*	0.085**
						(0.031)	(0.039)	(0.031)
Saxon						0.104**	0.122**	0.110**
						(0.034)	(0.034)	(0.033)
MKT score * Investigations							-0.082**	-0.081**
							(0.027)	(0.025)
MKT score * Expressions							-0.000	-0.003
							(0.029)	(0.028)
MKT score * Saxon							0.012	-0.004
							(0.029)	(0.027)
Baseline test score controls		√	√	√	√	√	√	√
Student, teacher, peer covariates					√	√		√
School covariates					√	√		√
Rand. block fixed effects				√	√	√	√	√
School fixed effects			√					
Adjusted R-squared	0.007	0.581	0.619	0.603	0.618	0.619	0.606	0.620

Note: Each column represents a separate regression with student observations. The dependent variable is the student's post-experiment standardized ECLS-K math test score. The independent variables are as shown above. "Baseline test score controls" include a quadratic in pre-experiment test score, which is allowed to differ in each year-by-grade cell. "Student covariates" include a quadratic in age, and indicators for female, Black, Hispanic, and English language learner. "Teacher covariates" include indicators for female, white, MA degree, having taken advanced math courses, having used the assigned curriculum previously, and novice teacher; linear terms for years since MA degree, and age; and quadratics in total experience, experience at the school, and professional development hours previously. "Peer covariates" include the mean and standard deviation of pre-experiment test score calculated among the student's classmates. "School covariates" include the proportion of students eligible for free or reduced price lunch, and title 1 eligible. Students were randomly sampled within classrooms, and these results are weighted by the inverse probability of selection. Standard errors allow for clustering within schools, the unit of random assignment. The estimation sample includes 7,650 students, 750 teachers, and 110 schools. Sample sizes have been rounded to nearest 10 following NCES restricted data reporting procedures.
+ indicates $p<0.10$, * $p<0.05$, and ** $p<0.01$

TABLE 4—TEACHER MATH SKILLS, INSTRUCTION PRACTICES, AND STUDENT SORTING
(dep. var. = pre-experiment ECLS-K math test score, standardized)

	(1)	(2)	(3)	(4)	(5)
MKT score	0.057**	0.032*	0.026+	0.026+	0.014
	(0.019)	(0.014)	(0.015)	(0.015)	(0.031)
Curricula main effects (relative to SFAW)					
Investigations				0.010	0.005
				(0.069)	(0.069)
Math Expressions				-0.035	-0.039
				(0.058)	(0.059)
Saxon				0.029	0.028
				(0.061)	(0.061)
MKT score * Investigations					0.051
					(0.039)
MKT score * Math Expressions					-0.021
					(0.040)
MKT score * Saxon					0.018
					(0.053)
Rand. block fixed effects			√	√	√
School fixed effects		√			
Adjusted R-squared	0.003	0.094	0.054	0.055	0.055

Note: Each column represents a separate regression with student observations. The dependent variable is the student's pre-experiment standardized ECLS-K math test score. The independent variables are as shown above. Students were randomly sampled within classrooms, and these results are weighted by the inverse probability of selection. Standard errors allow for clustering within schools, the unit of random assignment. The estimation sample includes 7,650 students, 750 teachers, and 110 schools. Sample sizes have been rounded to nearest 10 following NCES restricted data reporting procedures.
+ indicates $p<0.10$, * $p<0.05$, and ** $p<0.01$

TABLE 5—PAIRWISE TEST SCORE DIFFERENCES WITHIN MKT SCORE TERCILES
(dep. var. = post-experiment ECLS-K math test score, standardized)

	Bottom tercile			Middle tercile			Top tercile		
	Invest.	Express.	Saxon	Invest.	Express.	Saxon	Invest.	Express.	Saxon
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Express.	-0.026			-0.015			0.125*		
	(0.050)			(0.044)			(0.052)		
Saxon	0.007	0.033		-0.024	-0.009		0.159**	0.034	
	(0.043)	(0.044)		(0.050)	(0.050)		(0.056)	(0.054)	
SFAW	-0.104*	-0.078+	-0.111*	-0.119*	-0.105*	-0.095+	0.057	-0.068	-0.102+
	(0.052)	(0.047)	(0.044)	(0.046)	(0.044)	(0.049)	(0.050)	(0.046)	(0.053)

Note: Each cell is an estimated pairwise test-score difference between students in two curriculum conditions, conditional on their teacher's MKT tercile. For example, among students assigned to top-tercile MKT teachers, students whose teachers used "Math Expressions" (row 1) scored 0.125 standard deviations higher than students whose teachers used "Investigations" (Column 7). All estimates in this table come from a single regression. The dependent variable is the student's post-experiment standardized ECLS-K math test score. The key independent variables are a vector of indicators, one indicator for each MKT-tercile-by-curriculum combination. Independent variables also include the full set of pre-experiment controls as in Table 3 Column 5, 6, and 8, and randomization block fixed effects. Students were randomly sampled within classrooms, and these results are weighted by the inverse probability of selection. Standard errors allow for clustering within schools, the unit of random assignment. The estimation sample includes 7,650 students, 750 teachers, and 110 schools. Sample sizes have been rounded to nearest 10 following NCES restricted data reporting procedures.
+ indicates different from zero at $p<0.10$, * $p<0.05$, and ** $p<0.01$

TABLE 6--OBSERVED TEACHER BEHAVIORS AND CLASSROOM ENVIRONMENT

	Observed teacher behavior characteristic of…					
	student-led methods		direct-instruction methods		Classroom environment	
	(1)	(2)	(3)	(4)	(5)	(6)
Curricula main effects (relative to SFAW)						
Investigations	0.393**	0.397**	-1.016**	-1.009**	-0.056	-0.054
	(0.116)	(0.119)	(0.116)	(0.117)	(0.104)	(0.107)
Expressions	-0.374**	-0.374**	-0.443**	-0.442**	-0.166	-0.170
	(0.104)	(0.104)	(0.098)	(0.100)	(0.115)	(0.115)
Saxon	-0.549**	-0.543**	0.085	0.090	-0.127	-0.124
	(0.100)	(0.100)	(0.098)	(0.100)	(0.112)	(0.112)
MKT score	0.016		-0.023		0.031	
	(0.048)		(0.044)		(0.041)	
MKT score * Investigations		0.051		-0.030		0.027
		(0.141)		(0.086)		(0.081)
MKT score * Expressions		-0.103+		-0.072		-0.008
		(0.055)		(0.082)		(0.100)
MKT score * Saxon		-0.025		-0.020		0.064
		(0.055)		(0.058)		(0.070)
MKT score * SFAW		0.129+		0.025		0.039
		(0.074)		(0.077)		(0.082)
Adjusted R-squared	0.147	0.150	0.248	0.245	0.084	0.080
MKT * method slopes equal (p-value)		0.078		0.833		0.941
MKT * method slopes jointly zero (p-value)		0.130		0.882		0.883

Note: Each column represents a separate regression with teacher observations. Dependent variables are listed in the column headers. Each dependent variable is a predicted factor score derived from a factor analysis of classroom observation micro-data (see the text for complete details), and then standardized within the sample (mean zero, standard deviation one). In addition to the independent variables shown above, all specifications include randomization block fixed effects and several additional covariates. The additional covariates are the "teacher covariates" and "peer covariates" described in the notes for Table 3, as well as the teacher-/class-level means of all "student covariates" described in Table 3. Standard errors allow for clustering within schools, the unit of random assignment. The estimation sample includes 610 teachers and 110 schools. Sample sizes have been rounded to nearest 10 following NCES restricted data reporting procedures.
+ indicates $p<0.10$, * $p<0.05$, and ** $p<0.01$

TABLE 7—TREATMENT EFFECTS ON THE VARIANCE OF TEACHER PRODUCTIVITY

	A: Main estimates					B: Controlling for MKT	
						Quadratic	Terciles
	St. dev. teacher effects	Test of pairwise difference from… (p-value)			Joint test (p-value)	St. dev. teacher effects	St. dev. teacher effects
		Invest.	Express.	Saxon			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Maximum likelihood estimate of teacher st. dev. (teacher random effects)							
					0.001		
Investigations	0.077					0.072	0.076
Expressions	0.189	0.000				0.187	0.187
Saxon	0.161	0.007	0.307			0.161	0.159
SFAW	0.124	0.135	0.015	0.187		0.120	0.121
Estimated conditional st. dev. of teacher fixed effects							
					0.001		
Investigations	0.204					0.205	0.176
Expressions	0.264	0.001				0.265	0.268
Saxon	0.257	0.001	0.677			0.262	0.269
SFAW	0.227	0.154	0.024	0.033		0.237	0.185

Note: Columns 1, 6, and 7 report estimated between-teacher standard deviations in student standard deviation units.

Top panel, column group A from a linear mixed model estimated by maximum likelihood. The dependent variable is the student's post-experiment standardized ECLS-K math test score. The fixed effects potion includes "baseline test score controls," a quadratic in pre-experiment test score, which is allowed to differ in each year-by-grade cell; "student covariates," a quadratic in age, and indicators for female, Black, Hispanic, and English language learner; "peer covariates," the mean and standard deviation of pre-experiment test score calculated among the student's classmates; and randomization block fixed effects. The random effects portion includes four between-teacher variance parameters, one for each curriculum condition. The joint test in Column 5 is a likelihood-ratio test where the constrained model sets all four teacher variance parameters equal.

Bottom panel, column group A estimated in two steps: (i) estimate teacher fixed effects in a model with the same dependent variable and fixed portion independent variables as in the top panel MLE model; then (ii) estimate the conditional variance of the estimated teacher fixed effects. The latter step is a least-squares regression of squared residuals on the treatment indicators and randomization block fixed effects; the residuals are obtained from a regression of teacher fixed effects on the same right hand side variables. Standard errors allow for clustering within schools. The joint test in Column 5 is an F-test with the null that each of the treatment indicator coefficients in zero.

Column group B estimated just as in column group A but with controls added for MKT score. Top panel, Column 6 adds a a quadratic in MKT score where the parameters are allowed to be different for each of the four treatment conditions. Bottom panel, Column 6 adds the same quadratic terms to both regressions in step (ii). Column 7 replaces the quadratic terms with indicators for MKT tercile, again interacted with treatment condition.

The estimation sample includes 7,650 students, 750 teachers, and 110 schools. Sample sizes have been rounded to nearest 10 following NCES restricted data reporting procedures.

Appendix Table A1—Factor loadings for teacher behavior items from classroom observations

	Factor 1	Factor 2
	Observed teacher behavior characteristic of…	
	student-led	direct-instruction
	(1)	(2)
A. Teacher initiated instructional behaviors (frequency observed)		
Asks closed-ended questions	0.15	**0.79**
Poses open-ended questions	**0.58**	**-0.21**
Tells information, models procedures, or shows students how to represent concepts	0.20	**0.33**
Guides practice on problems	0.22	**0.38**
Elicits multiple strategies/solutions	**0.48**	**-0.20**
Uses representations	0.19	**0.26**
B. Teacher response to student answers (frequency observed)		
States if correct or not without elaborating or repeats what child said with indication of right or wrong	0.16	**0.66**
Calls on other students until the "correct" answer is given	0.15	0.18
Provides correct answer right away	0.06	0.11
Asks class if they agree or disagree with student's response	0.14	0.14
takes student through step-by-step procedure	0.26	0.11
Tells student strategy to use	0.30	-0.01
Elicits other student' questions about the student's response	0.30	-0.09
Labels math strategy, problem, or concept	**0.32**	0.02
Repeats student answer in a neutral way	**0.46**	**-0.24**
C. Teacher guidance and follow up questions (frequency observed)		
Probes for reasoning or justification of solution	**0.65**	-0.16
Provides hint to students	**0.49**	0.05
Clarifies what student says or does	**0.60**	-0.15
Extends what student says of does	**0.41**	-0.06
D. Teacher praise (frequency observed)		
Uses praise or makes positive comments focused on content	**0.32**	0.09
Highlights student work of solution to class	**0.39**	-0.15
Praises effort or behavior	0.28	0.10
E. Evidence of instructional behaviors (binary yes, no)		
States lesson objective at the beginning of class	0.07	0.09
Connects lesson to prior knowledge/instruction	0.19	0.15
Demonstrates how to play game	0.17	-0.16
Guides children in acting out a problem	0.15	0.07
Leads children in a rap, song, or finger play to illustrate math concept or practice	0.01	0.12
Uses children's book to make connections to math concept	0.12	0.07
Connects math to real life problems of situations	0.13	0.13
Directs or encourages students to help one another with math	0.27	-0.08
Prompts child to guide practice or lead class in a routine	0.01	0.07
Leads summary of what was learned or asks student to lead/share summary	0.07	**0.22**
Administered a written assessment	0.01	0.03

Note: Factor loadings obtained from a principal factor analysis of the items listed. Bold loading values represent the ten largest loadings, in absolute value, for the factor. The estimation sample includes 610 teachers. Sample sizes have been rounded to nearest 10 following NCES restricted data reporting procedures.

APPENDIX TABLE A2—FACTOR LOADINGS FOR CLASS ENVIRONMENT ITEMS FROM CLASSROOM OBSERVATIONS

	Factor 1
	(1)
Students are cooperative and attentive to the lesson	0.777
Student behavior disrupts the classroom	-0.750
Teacher spends a lot of time managing behavior	-0.749
Class runs without disruption from student behavior	0.736
Students are perfectly behaved	0.721
Teacher has techniques for gaining class attention in less than 10 seconds	0.713
Class time is spent on understanding or practicing math	0.692
Students are off-task	-0.678
Students spend little time waiting or transitioning	0.651
Students are actively engaged	0.642
Transitions are smooth and students get to work quickly	0.637
Teacher is fluid in presentation	0.625
Students appear excited by the lesson	0.593
Teacher and students have a warm, positive relationship	0.591
Teacher used nonverbal methods to manage misbehaviors	0.529
Teacher spends a lot of time giving directions	-0.468
Students appear familiar with the materials and procedures used	0.446
In monitoring student work, teacher followed through to ensure understanding	0.435
Students are given the opportunity to think and respond	0.433
Students attended to the lesson in a passive way	-0.371
Teacher has materials prepared and ready for students	0.352
During independent work time the teacher monitored student work	0.318
Students had easy access and permission to use manipulative when working	0.283
Teacher used praise or rewards to maintain positive behavior	0.276
Peer to peer interaction about math occurs	0.239
Teacher encourages students to help one another understand the math	0.233
Students help one another to understand math concepts or procedures	0.224
Students need to wait for the teacher to begin or for other students to finish working before they work on next problem or activity	-0.165
Teacher differentiated curriculum for children who were English language learners	0.094
Teacher differentiated curriculum for children who were above level	0.085
Teacher differentiated curriculum for children who were below level	0.080

Note: Factor loadings obtained from a principal factor analysis of the items listed. Items listed in order from largest to smallest loading, in absolute value. The estimation sample includes 610 teachers. Sample sizes have been rounded to nearest 10 following NCES restricted data reporting procedures.

APPENDIX TABLE A3—ADDITIONAL SPECIFICATIONS: TEACHER MATH SKILLS, INSTRUCTION PRACTICES, AND STUDENT ACHIEVEMENT
(dep. var. = post-experiment ECLS-K math test score, standardized)

	Quad.	School fixed effects			Equally weighted							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
MKT score	0.024	0.010	0.034*	0.039*	0.089**	0.039**	0.009	0.007	0.004	0.007	0.018	0.018
	(0.017)	(0.011)	(0.016)	(0.017)	(0.022)	(0.014)	(0.010)	(0.010)	(0.009)	(0.009)	(0.017)	(0.019)
MKT score ^2	0.031*											
	(0.013)											
Curricula main effects (relative to SFAW)												
Investigations	0.090*									0.030	0.036	0.039
	(0.037)									(0.032)	(0.034)	(0.032)
Expressions	0.130**									0.070*	0.081*	0.073*
	(0.034)									(0.030)	(0.036)	(0.030)
Saxon	0.121**									0.101**	0.119**	0.105**
	(0.032)									(0.030)	(0.033)	(0.030)
MKT score * Investigations	-0.075**		-0.065**	-0.065**							-0.056+	-0.054*
	(0.024)		(0.023)	(0.023)							(0.030)	(0.027)
MKT score * Expressions	0.007		-0.002	-0.006							0.001	0.004
	(0.027)		(0.024)	(0.027)							(0.028)	(0.029)
MKT score * Saxon	0.013		-0.039	-0.048							0.024	0.011
	(0.027)		(0.028)	(0.029)							(0.030)	(0.029)
MKT score ^2 * Investigations	-0.025											
	(0.017)											
MKT score ^2 * Expressions	-0.045*											
	(0.020)											
MKT score ^2 * Saxon	-0.012											
	(0.018)											
Baseline test score controls	√	√	√	√		√	√	√	√	√	√	√
Student, teacher, peer covariates	√	√		√					√	√		√
School covariates	√								√	√		√
Rand. block fixed effects	√							√	√	√	√	√
School fixed effects		√	√	√			√					
Adjusted R-squared	0.620	0.629	0.619	0.629	0.008	0.593	0.630	0.615	0.629	0.630	0.618	0.631

Note: Estimation details are identical to Table 3, except that Columns 5-12 are not weighted. + indicates p<0.10, * p<0.05, and ** p<0.01