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

Eric S. Taylor
Harvard University

March 2017

I study how teachers' assigned job tasks—the 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 job tasks. I examine a randomized trial of different approaches to teaching math, each 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": explaining and modeling rules and procedures. The relationship is weaker, perhaps negative, for newer "student-led" methods.

eric_taylor@harvard.edu Gutman Library 469, 6 Appian Way, Cambridge, MA 02138. I thank Eric Bettinger, Brian Jacob, Susanna Loeb and seminar participants at Stanford University, the New York Federal Reserve, and the NBER Economics of Education Meetings 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.

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), among others, 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 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 productivity is, in comparison,

[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).

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 between two 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 two approaches to teaching, randomly assigned to schools, can be characterized as "direct instruction" and "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,

[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.

however, I show that this apparent zero correlation masks meaningful heterogeneity caused by the different tasks (instructional methods) teachers are assigned. The correlation between teacher skills and productivity is positive when teachers use "direct instruction" methods. By contrast, the correlation is much weaker, perhaps even negative, when teachers use a "student-led" approach to teach 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 student-led methods score 0.11 standard deviations lower at the end of the year than their peers taught with more-conventional direct instruction. By contrast, there is no significant 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.

Causal interpretation of these estimates relies largely on the random assignment of schools and teachers to the instructional-method treatment conditions. 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.

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 2016) 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 the sampled 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, Rockoff, and Staiger 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. 2016).[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, Rowan, and Ball 2005, Rockoff et al. 2011, Hill et al. 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 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 different *instructional methods* for teaching math—the

[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 2015). 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.

treatment conditions—but the math *concepts* teachers were asked to cover during the school year did not vary.[5] 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.[6]

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 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.[7]

[5] To be more specific, the *state math curriculum (content) standards* did not vary across treatment conditions (randomization was at the school level within districts). Curriculum standards define what math concepts teachers are expected (asked) to teach.

While the expectations did not vary, there may have been variation in the sequence of concepts or the amount of time spent on a concept. There are no data on sequence, but some data on time spent. In a post-treatment survey teachers were asked how many lessons they taught during the year for 20 different topics, for example, "adding and subtracting, with whole numbers" and "decimals." The first two factors of the 20 items explain 83 percent of the variation in reported time spent on different topics. Measured by these two factors, teachers in the "student-led" condition (see Section 2.1) spent less time on the given math topics than did teachers in the "direct instruction" condition, about one-third of a standard deviation less. However, these differences do not covary with teachers' MKT scores. These results are shown in Appendix Table A1.

[6] 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.

[7] 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

Table 1 describes the study participants, including 80 schools with nearly 600 teachers and their students. 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

In this paper I compare teachers in two experimental conditions: those assigned "direct instruction" tasks and those assigned "student-led" tasks. However, schools were originally randomly assigned, within blocks, to one of four treatment conditions.[8] Each of the four conditions is an approach to teaching elementary mathematics codified in a commercially published set of teacher instructions and materials. As described in the following paragraphs, the first-order differences between the four approaches are differences in the use of "direct instruction" methods and "student-led" methods, and so for this paper's analysis I collapse the four original conditions into two conditions.

The four original conditions are known by the commercial names *Investigations in Number, Data, and Space*; *Math Expressions*; *Saxon Math*; and *Scott Foresman-Addison Wesley Mathematics*. Collectively *Investigations*, *Saxon*,

effects) were estimated separately for teachers in two groups: those in the bottom quintile of the sample MKT distribution, and all other teachers.

[8] Randomization blocks were defined by district and observable characteristics. By rule, each 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.

and *SFAW* are used in about one-third of K-2 classrooms; *Expressions* is a relatively new product (Agodini and Harris 2010).[9]

The contrast between "direct instruction" and "student-led" methods captures the first-order differences between the four original conditions. *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 to teaching. Student-led methods "focus on [students'] understanding, rather than on students answering problems correctly" (Agodini et al. 2010, p. xxi). In student-led classrooms, students are expected to reason-through and articulate math concepts with each other, while 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* is designed to combine both direct instruction and student-led methods, though written descriptions suggest greater weight is given to direct instruction activities.

These differences between conditions can be seen empirically in Table 2 using data collected during classroom observations.[10] Trained observers spent, on average, 1.5 hours in each classroom 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

[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.

[10] 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.

"probes for [a student's] reasoning or justification of a solution." Using factor analysis I reduce these detailed data to two summary measures: teacher behavior characteristic of (i) student-led methods and (ii) direct instruction. 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).[11] Table 2 Column 1 reports results from a simple regression of the student-led factor (mean zero, standard deviation one) on indicators for the four treatment conditions and fixed effects for the randomization blocks; standard errors are clustered at the school level. Column 2 reports the same for the direct instruction factor.

As measured in classroom observations, teachers assigned to the *Saxon* and *SFAW* conditions used direct instruction methods much more often—over one standard deviation more often—than teachers assigned to *Investigations*. In parallel, *Investigations* teachers used student-led methods much more often than *Saxon* or *SFAW* teachers, differences of 0.96 and 0.43 standard deviations respectively. Teachers assigned to *Expressions* use direct instruction methods about as often as *Saxon* and *SFAW* teachers; but are more likely to use student-led methods, though not reaching *Investigations* levels.

In contrast to these differences in teachers' use of direct instruction and student-led methods, there are no (statistically significant) differences on a measure of classroom management or environment also drawn from the observation data. Observers rated classroom environment on 31 characteristics, separate from the frequency of teacher practices data used above. Items including "student behavior disrupts the classroom," "class time is spent on understanding

[11] All the included measures and factor weights are detailed in Appendix Table A2. 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.

or practicing math," and "teacher has techniques for gaining class attention in less than 10 sections" were each scored on a four point scale from "1 = not at all characteristic (almost never)" to "4 = extremely characteristic (almost always evident)". In Column 3 the dependent variable is the first predicted factor from these 31 items.[12] In short, teachers' assigned condition did not affect observed classroom management.

To streamline the analysis and discussion in this paper, I collapse the four original conditions into two conditions: (i) "direct instruction" approaches combining *Saxon* and *SFAW*, and (ii) the "student-led" approach *Investigations*. Given the intermediate nature of *Expressions*, I do not use the *Expressions* observations in this paper's main analysis. As shown in the appendix, however, the pattern of results is quite robust to adding *Expressions* into the "direct instruction" group.

There are two notable limitations to using these experimental treatments to study the teacher tasks that characterize direct instruction and student-led methods. First, *Investigations* is the only example of student-led methods, though it was selected for the original study in part to represent student-led approaches more generally. Second, the data were collected during a single school year; for most teachers in the student-led condition this was their first year using student-led methods.[13] A teacher's use of any approach or set of materials is likely to improve in the second year and beyond.

One important final note about the treatment conditions: there are differences across conditions in post-experiment student math test scores. I estimate that student test scores are highest, on average, with *Saxon*, followed by

[12] 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 Appendix Table A3. The results are quite similar using the simple average of the 31 items, after reversing the scale of negatively framed items.

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

Expressions (0.02σ, student standard deviations, lower than *Saxon*), *Investigations* (0.05σ lower than *Saxon*), and *SFAW* (0.10σ lower than *Saxon*). However the only statistically significant differences are that *Saxon* and *Expressions* scores are higher than *SFAW* scores.[14] This pattern is the same basic pattern of results reported in the original analysis of the experiment by Agodini and coauthors (2010).[15,16] These effects are meaningful. The 0.10σ difference between *Saxon* and *SFAW* is at least as large as the difference in test scores between the classrooms of novice teachers and teachers with five year experience (Rockoff 2004, Papay and Kraft 2015), and is half as large as the effect of reducing class size by 30 percent in early elementary grades (Kruger 1999). However, the rank ordering by test scores does not align with the direct instruction versus student-led dimension. As reported later in Table 5, there is essentially no difference in average test scores on the binary comparison of direct instruction and student-led methods.

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,

[14] These results are detailed further in Appendix Table A4. The estimation follows the conventional approach to estimating treatment effects as described in Section 3.

[15] Agodini et al. (2010) estimate effects separately for 1st and 2nd grades (Table III.2). Averaging the 1st and 2nd grade estimates from Table III.2, the differences from *Saxon* are 0σ for *Expressions*, -0.08σ for *Investigations*, and -0.12σ for *SFAW*. The Agodini et al. point estimates differ some between grades, but those differences are not statistically significant. Additionally, I have (very nearly) replicated the results in Table III.2 of Agodini et al. (2010) following their description of the original methods.

[16] Other evaluations have examined *Investigations*, *Saxon*, and *SWAF* individually; they generally find no effects, but the counterfactuals are difficult to define (see the review in Agodini and Harris 2010).

interpreting and/or predicting student errors, and representing mathematical ideas and procedures in ways learners can grasp" (Hill, Kapitula, and Umland 2011, p. 804).[17] An example MKT item is shown in Figure 1. The test's reliability is quite high, for example, Hill and coauthors (2012) estimate an IRT reliability of 0.94-0.97.

The MKT test is a widely used measure in education research, appearing in dozens of studies over the past decade.[18] For example, and relevant to judging the test's validity for this study, MKT scores predict teaching skills measured in classroom observations, both observations of general teaching skills and math-specific skills (Hill, Kapitula, and Umland 2011, Hill et al. 2012). Use in economics is rare, however. Rockoff et al. (2011) and Jacob et al. (2016) study teacher hiring decisions and what information on applicants might improve those decisions, including MKT test scores and many other measures.

2.3 Student test scores and other data

Students were tested pre- and post-experiment using math tests developed for the ECLS-K study.[19] 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. Consistent with available data from the original ECLS-K, test score reliabilities in this study's

[17] 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.

[18] Additional information on the MKT test, including many example items and a list of research studies using MKT, is available from the Learning Mathematics for Teaching Project at the University of Michigan: http://www.umich.edu/~lmtweb/.

[19] Not all students in participating schools and classrooms 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.

sample fall between 0.88-0.93 (Rock and Pollack 2002, Agodini et al. 2010). Beginning with the scale scores, I standardize both pre and post scores (mean zero, standard deviation one) within grade. ECLS-K test scores are a common outcome measure in empirical research. In economics ECLS-K scores have been used to study, in the most common examples, race and gender gaps (Fryer and Levitt 2004, 2010, Bond and Lang 2013) and education in early childhood (Bedard and Dhuey 2006, Classens, Duncan, and Engel 2009, Neidell and Waldfogel 2010).

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 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 direct instruction and student-led conditions. The samples are well balanced. There is some evidence of differences in teachers with a master's degree, but these are two of more than twenty characteristics tested.[20]

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

[20] The results are very similar if I test for covariate balance across the four separate original treatment conditions. The results for the attrition tests in the next paragraph are also very similar. These alternative results are provided in the Appendix Tables A5 and A6.

estimates. Since, as I describe shortly, teacher productivity is measured with student test score growth, attrition correlated with baseline score is of particular concern. As shown in Table 3, there is little evidence of differences in attrition patterns: no differences in attrition rates for direct instruction and student-led 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 (Jackson, Rockoff, and Staiger 2014 provide a review), 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' math skills and student test scores

Teachers' MKT scores and their students' math test scores are positively correlated. This is apparent in Table 4 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 0.073σ (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.[21] The regression reported in Table 4 Column 2 is identical to Column 1 except that I have added controls for student i's pre-experiment test score; specifically, a 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.[22] 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, class, 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 6 (bottom panel). The coefficient falls when randomization block fixed effects are included in Column 8, but there is still evidence of within-school sorting. In short, better students are assigned to better

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

[22] 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 4 Columns 3 and 4. Estimates using school fixed effects are provided in Appendix Table A7. 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.

teachers, at least better on the MKT dimension. This kind of within-school sorting has been documented elsewhere using 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 learning. 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 Table 5, are obtained by fitting variations of the following model:

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

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. $T_{s(i)}$ is an indicator $= 1$ for the "student-led" treatment condition, which was randomly assigned at the school level. The "direct instruction" condition is the omitted group. 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, class peer, teacher, and school observable characteristics, notably among them teacher experience and an indicator for having experience with the assigned product previously.[23] Last, $\tau_{b(s)}$ is a series of fixed effects 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. The pattern of results is the same in estimates without weighting.

The current research question can be thought of as comparing different ways to specify $f\left(MKT_{j(i)}\right)$ in Model 1. The results discussed in the previous section are estimates where $f\left(MKT_{j(i)}\right)$ 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 5 Column 2 which uses the specification in Equation 1, and lets α vary by treatment condition, i.e., $f = MKT_{j(i)} * T_{s(i)} * \alpha$. For teachers assigned to use direct instruction $\hat{\alpha}$ is positive, 0.024 (different from zero p-value = 0.081), but for teachers assigned to use

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

student-led methods $\hat{\alpha}$ is negative, -0.048. The $\hat{\alpha}$ for direct instruction is relatively less precisely estimated; I cannot reject a slope close to zero or a slope twice as large as the estimate. The $\hat{\alpha}$ for student-led relatively more precise; it is significantly different from the $\hat{\alpha}$ for direct instruction condition (p-value = 0.005), and significantly different from zero (p-value = 0.001).

In other words, for direct instruction 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. But for student-led methods there is an apparent *negative* relationship: students assigned to a teacher with better math skills score lower in the end.

The estimated magnitudes of these differences are educationally and economically meaningful. In schools using student-led methods, a one standard deviation difference in teacher math skills (MKT scores) translates into a 0.05σ difference in student test scores. For comparison, one standard deviation in total teacher contribution to student test scores is typically estimated to be 0.10-0.20σ (Jackson, Rockoff, and Staiger 2014). Thus a 0.05σ difference is one-quarter to one-half of the total between-teacher difference student achievement. More concretely, 0.05σ is half, or more, of the difference between a novice teacher and a teacher with five years of experience (Rockoff 2004, Papay and Kraft 2015). Additionally, 0.05σ is roughly one-quarter the estimated gain from doubling the amount of class time middle and high school students spend in math (Taylor 2014, Cortes, Goodman, and Nomi 2015). In schools using direct instruction, the slope is half as large 0.024σ, but still meaningful.

Additionally, the estimated difference in MKT-slope between direct instruction and student-led conditions is much larger than the estimated difference in mean student test scores between the two conditions (see Table 5 Row 1). The estimated difference in MKT-slope is also large in comparison to the pairwise differences in mean test scores. Recall that the largest difference was that *Saxon* students score 0.10σ higher than *SFAW* students, on average, and that all pairwise differences between the four original conditions were between 0.02-0.10σ (see Appendix Table A4). The comparison of the MKT-slope differences to the mean differences suggests the interaction of teachers math skills and assigned teaching tasks is a first-order consideration for school leaders choosing among different approaches.

Should these results be interpreted as causal relationships? In short, the *slope* of the relationship between MKT scores and student scores—the $\hat{\alpha}$ for each of the two conditions—should not be given a strong causal interpretation. These slopes are potentially biased by omitted variables. But the *difference in slopes* can plausibly be interpreted casually. The difference in slopes is substantially less threatened by omitted variables because the treatment conditions were randomly assigned.

First consider the individual slope estimates. Causal interpretation of either individual slope requires the assumption that $MKT_{j(i)}$ is uncorrelated with the error term, $\varepsilon_{i,t}$; otherwise the slope estimate will be biased by omitted variables. Teachers, and workers generally, bring a bundle of correlated skills to their jobs, and the unobserved skills are a perennial concern in studying how one or a few specific skills affect productivity. I have no new identification strategy to offer in this paper. My estimates do, however, address critical sources of potential bias. Recent empirical evidence suggests that bias arising from the non-random sorting of students to teachers 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 et al. 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 bias from other omitted measures of teacher skill. In this case I can control for a much richer set of teacher characteristics than is usually possible, for example, whether the teacher has taken advanced math courses. Encouragingly, the estimates of interest are not substantially different with and without this rich set of teacher controls; Table 5 Column 3 without any controls is quite similar to Column 2, and the same is true for several other permutations of included teacher covariates not reported here.

One plausible omitted variable not available in this paper's data is a measure of teachers verbal or language skills. There is, however, some relevant evidence from other settings. Rockoff and coauthors (2011) observe teachers' SAT verbal scores, along with MKT scores and student test scores. The estimates in Rockoff et al. (2011) suggest omitting verbal skills may negatively bias the slope on MKT by 16 percent.[24] The potential bias is relatively small in part because MKT scores and SAT verbal scores are only moderately correlated at 0.38. Hill, Rowan, and Ball (2005) report a similar correlation of 0.39 between MKT and a measure of teacher knowledge for teaching reading.

Now consider the difference in slopes. Interpreting this change in slope as a causal effect of switching from direct instruction to student-led methods requires a weaker identifying assumption: that any omitted viable bias is independent of treatment assignment. This assumption rests primarily on the random assignment of schools which, as discussed in Section 2, appears to have been successful. Table 4 Column 10 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. However, random assignment does not justify the identifying assumption necessarily. Omitted viable bias will be the product of the

[24] Rockoff and coauthors (2011) estimate the slope on teacher MKT scores at 0.019 (st.err. 0.011) which is within the confidence interval of the estimate for the direct-instruction condition.

relationship between the omitted variable and MKT score, and the relationship between the omitted variable and student test scores. The latter of these two relationships may be affected by treatment assignment. This is the case, for example, for the (potentially omitted variable) years of teaching experience, as shown in Table 5 Column 5. When teachers are assigned to use student-led methods, instead of direct instruction, the returns to experience fall essentially to zero. Still, accounting for this more-subtle source of potential bias does not change the results for MKT scores; compare Columns 3 and 6.

The results on years of teaching experience in Table 5 are also relevant more generally.[25] Experience is itself a reasonable proxy for teaching skills, especially skills best learned on the job like how to manage student misbehavior. Indeed, experience is the most consistent observable predictor of teacher productivity (except prior productivity). Thus, experience should be at the top of the list of potential omitted variables for this and any similar examination of specific teacher skills. In this case, the results for MKT scores are quite robust to the inclusion or exclusion of flexible functions of experience. Additionally, the results for experience themselves suggest that teachers' assigned tasks may affect the role of other teaching skills beyond the skills measured by the MKT test.

To summarize the results presented thus far: First, in schools using a direct instruction approach to math, teachers with better math skills are apparently more productive than their lower skilled colleagues, where productivity is measured by contributions to student math achievement. The modifier "apparently" is a reminder that the causal relationship is uncertain, given the potential for omitted variables bias. The productivity returns to teacher skills could plausibly be smaller or larger than estimated here.

[25] The main estimates in Columns 1 and 2 do include a quadratic in experience. Columns 4-6 use a linear term in experience for exposition, but the results are robust to using a quadratic or less-parametric specification.

Second, if schools switched from direct instruction to student-led methods the benefit of employing (being taught by) a teacher with better math skills would shrink or even be reversed. Put differently, student-led methods reduce the return to teacher math skills in the production of student math knowledge, relative to the return using direct instruction. The causal relationship here is much less uncertain because identification is based primarily on the random assignment design.

I cannot econometrically rule out the possibility that, in schools using direct instruction, teachers with better math skills are in fact *less* productive than their lower skilled colleagues. There would, however, need to be substantial omitted variables bias in my estimates to hide this kind of negative return on skills. Even if this was the case, the *loss* from employing (being taught by) a teacher with better math skills would be minimized by choosing direct instruction over student-led 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 plotted in Figure 2.[26] Each point on the graph measures the average student test score for teachers of a given MKT tercile and treatment condition. Test scores are measured relative to the average for middle-tercile direct instruction; the dotted lines mark 95 percent confidence intervals. 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.

A few 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. There is no statistically significant difference between direct instruction and student-led methods in those two terciles. Moreover, low-skilled teachers appear equally as good as their average-skilled colleagues at producing student achievement growth.

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. Students of top-tercile teachers using direct instruction methods score 0.11σ higher, on average, than their peers assigned to equally high-skilled teachers using student-led methods (p-value = 0.012). That difference is consequential; 0.11σ is half to three-quarters of the standard deviation in total teacher productivity (Jackson, Rockoff, and Staiger 2014).

3.4 Potential mechanisms

[26] 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 Appendix Table A7. Adding higher order terms beyond a quadratic do not improve the model, at least as judged by likelihood ratio tests.

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 to their students. Direct instruction relies on the teacher, far more frequently than student-led methods, to explain and model math concepts and procedures. 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.[27] Put differently, high-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. Moreover, which of their assigned tasks teachers do may be a function of math skills. Thus to evaluate the hypothesis in the previous paragraph it would be helpful to have evidence on the covariance of teachers' skills, assigned tasks, and tasks carried out.

The reminder of this section describes evidence available from data collected in this study. Data on what teachers actually did in the classroom come from classroom observations by the research team and from teacher self-reports in a post-experiment survey. These data include dozens of items measuring the frequency of teacher behaviors and practices, including "tells information [or] models procedures" and "probes for [a student's] reasoning or justification of a solution" from classroom observations, and "demonstrate or model math concepts

[27] 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.

or procedures for students" and "invite students to use multiple strategies or solutions to a problem" from the teacher survey.[28]

Analysis of these relatively-rich data shows, first, that the tasks teachers actually do from day to day are related to their math skills; and, second, that the relationships are quite different when teachers are asked to use student-led methods instead of direct instruction. In the next paragraph I describe a few concrete examples of individual observation and survey items. The main results in Table 6, discussed after the examples, combine the dozens of individual items into a single test.

Consider the frequency with which teachers "demonstrate or model math concepts or procedures for students," as an example task. In schools using direct instruction, teachers with better math skills "demonstrate…for students" more frequently than their lower skilled colleagues. A one standard deviation increase in a teacher's MKT score predicts a 0.12 standard deviation increase in the frequency of that tactic (standard error 0.046).[29] By contrast, in schools using student-led methods, teachers with better math skills are apparently no more or less likely to "demonstrate…for students." The estimated slope is negative but not statistically significant; the difference in slopes (treatment effect on the slope) is -0.18 and statistically significant.[30] As a second example, this same pattern of results holds for the number of practice problems teachers give their students

[28] The classroom observation data are described earlier in Section 2.1. In the survey, teachers were asked to report how frequently they use specific teaching tactics. Teachers responded using a six-point ordered scale of "never", "less than once a month,"…, "1-2 times a day", "3 or more times a day." Teachers were also asked in the survey how many lessons they taught during the school year for 20 different topics like "adding and subtracting, with whole numbers" and "money."

[29] This result comes from a regression of the frequency of "demonstrate…for students" (mean zero, standard deviation one) on an indicator for the student-led condition, MKT score, and the interaction of the two. The specification also includes randomization block fixed effects, and standard errors are clustered at the school level. Detailed results are available from the author.

[30] The average teacher in the student-led condition "demonstrate[s]…for students" nearly one-third of a standard deviation less frequently than the average direct instruction teacher (main effect point estimate -0.306, standard error 0.091).

during class, as measured in classroom observations. Teachers with better math skills give students more practice problems when using direct instruction (point estimate 0.14, standard error 0.064), but the relationship turns negative with student-led methods (treatment effect on the slope -0.30, standard error 0.099). A final example contrasts the first two. In direct-instruction schools, teachers with better math skills are less likely to "invite students to use multiple strategies or solutions to a problem" (point estimate -0.17, standard error 0.075). In student-led schools the relationship is still negative, but weaker and not statistically significant (treatment effect on the slope -0.14, standard error 0.096).

The main results for this section are shown in Table 6. While the examples in the previous paragraph are illustrative, they are only three examples out of dozens of observation and survey items that measured what teachers actually did in the classroom. In Table 6 I combine these dozens of items into a single index of teachers' classroom practices, and then examine how that practices index vary with teachers' math skills and treatment condition. To construct the index I estimate an auxiliary regression of student test scores on the dozens of different teacher practices variables in the survey and observation data; the fitted value from this regression is the practices index.[31] This index is essentially a weighted average of teacher practices where the weights are the extent to which the practice predicts higher student test scores. Teachers with higher values of the index are teachers who more often engage in practices which promote student achievement.

In direct instruction schools, the setting typical in public schools, teachers with better math skills are more likely to use practices which promote student

[31] This auxiliary regression is analogous to the regressions in the top panel of Table 4, except that "MKT score" is replaced with the vector of teacher practices variables. The teacher practices variables are: (a) Item level data from classroom observations. These data are described in Section 2.1 and were used to construct the dependent variables in Table 2. (b) Item level data from the teacher survey on coverage of 20 math topics. These were used to construct the dependent variable in Appendix Table A1. (c) Item level data from the teacher survey on classroom practices, described in this section, and listed in Appendix Table A8.

achievement. As reported in Table 6 Column 1, a one standard deviation increase in a teacher's MKT score predicts a 0.13 standard deviation increase in the practices index.[32] However, in student-led methods schools, the relationship is reversed or at least diminished. Teachers with better math skills are less likely to use practices which promote achievement. The MKT slope for student-led teachers is -0.07 (= 0.131-0.205), though not statistically significant (standard error 0.101). Teachers' assigned tasks change what teachers actually do in the classroom day-to-day.

The results in Table 6 Column 1 are robust to how the practices index is estimated. One potential concern is that estimated auxiliary-regression coefficients used to construct the index may be biased by selection, especially the non-random sorting of students to teachers. With this concern in mind, the index dependent variable in Column 1 comes from an auxiliary regression which also includes the student, class, and school controls used in Table 4 and elsewhere in the paper. I fit the regression specification:

$$A_{i,t} = P_{j(i)}\theta + h_{g(i),y(i)}\left(A_{i,t-1}\right) + X\beta + \tau_{b(s)} + \varepsilon_{i,t}, \tag{2}$$

where P_j is the vector of teacher practices variables drawn from observations and surveys, and all other terms and notation are as described for Equation 1. While the regression includes controls, the dependent variable for Column 1 is the fitted value $P_j\hat{\theta}$ which only uses variation in P_j. For comparison Column 2 shows results where $h\left(A_{i,t-1}\right)$ and $X\beta$ are left out of the auxiliary regression; the pattern of results is essentially unchanged. A second potential concern is that if treatment condition changes the returns to practices, θ, then the results in Columns 1 and 2

[32] The units in Table 6 are teacher standard deviations of the index measure. One teacher standard deviation of the index is roughly equivalent to 0.15 student test score standard deviations for the index used in Columns 1-2, and 0.19-0.20 student standard deviations for the index used in Columns 3-4.

could (partly) reflect misspecification of the auxiliary regression. For the results in Columns 3 and 4 I modify the auxiliary regression to allow the teacher practices coefficients to differ by treatment condition (in Equation 2 I interact P_j with T_s). If anything, this makes the pattern of results stronger. The MKT-practices slope for direct instruction teachers remains positive and not very different in magnitude, but the same slope for student-led methods teachers is more steeply negative at -0.23 (= 0.131-0.362, standard error 0.111).

4. 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 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 0.11σ 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.

Two features of the empirical setting potentially limit the generalizability of these results. First, the results for student-led methods are based on a single example of the approach, *Investigations*, while the results for direct instruction include both *Saxon* and *SFAW*. *Investigations* is strongly student-led and typical of the approach generally, but *Investigations* may have idiosyncratic features that make the differences estimated here larger or smaller than they would be with other student-led products. Second, the data are from a single school year—often the first school year teachers had used student-led methods. A teacher’s success with student-led methods, or any specific approach and materials, is likely to improve in the second year of use or beyond. It is less clear weather or how the relative differences between student-led and direct instruction will evolve as teachers’ gain experience with each approach.

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, Daron, and David H. Autor. 2011. Skills, tasks and technologies: Implications for employment and earnings. *Handbook of Labor Economics* 4, ed. Orley Ashenfelter and David Card, 1043-1171. Amsterdam: Elsevier.

Acemoglu, Daron, and David H. Autor. 2012. What does human capital do? A review of Goldin and Katz's *The race between education and technology*. *Journal of Economic Literature* 50, no. 2:426-463.

Acemoglu, Daron, and Fabrizio Zilibotti. 2001. Productivity differences. *Quarterly Journal of Economics* 116, no. 2:563-606.

Agodini, Robert, and Barbara Harris. 2010. An experimental evaluation of four elementary school math curricula. *Journal of Research on Educational Effectiveness* 3, no. 3:199-253.

Agodini, Robert, Barbara Harris, Melissa Thomas, Robert Murphy, and Lawrence Gallagher. 2010. *Achievement effects of four early elementary school math curricula: Findings for first and second graders.* Publication no. 2011-4001. Washington, D.C.: Institute for Education Sciences, U.S. Department of Education.

Autor, David, Frank Levy, and Richard J. Murnane. 2003. The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics* 118, no. 4:1279-1333.

Bond, Timothy N., and Kevin Lang. 2013. The evolution of the black-white test score gap in grades K–3: The fragility of results. *Review of Economics and Statistics* 95, no. 5:1468-1479.

Becker, Gary. 1964. *Human capital: A theoretical and empirical analysis, with special reference to education.* Chicago: University of Chicago Press.

Bedard, Kelly, and Elizabeth Dhuey. 2006. The persistence of early childhood maturity: International evidence of long-run age effects. *Quarterly Journal of Economics* 121, no. 4:1437-1472.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014a. Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review* 104, no. 9:2593-2632.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014b. Measuring the impacts of teachers II: Teacher value-added and outcomes in adulthood. *American Economic Review* 104, no. 9:2633-2679.

Claessens, Amy, Greg Duncan, and Mimi Engel. 2009. Kindergarten skills and fifth-grade achievement: Evidence from the ECLS-K. *Economics of Education Review* 28, no. 4:415-427.

Clotfelter, Charles T., Helen F. Ladd, and Jacob Vigdor. 2005. Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review* 24, no. 4:377-392.

Cortes, Kalena E., Joshua S. Goodman, and Takako Nomi. 2015. Intensive math instruction and educational attainment: Long-run impacts of double-dose algebra. *Journal of Human Resources* 50, no. 1:108-158.

Costinot, Arnaud, and Johnathan Vogel. 2010. Matching and inequality in the world economy. *Journal of Political Economy* 118, no. 4:747-786.

Dobbie, Willie. 2011. Teacher characteristics and student achievement: Evidence from Teach for America. Harvard University Working Paper.

Fryer, Ronald G., and Steven D. Levitt. 2004. Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics* 86, no. 2:447-464.

Fryer, Ronald G., and Steven D. Levitt. 2010. An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics* 2, no. 2:210-240.

Hanushek, Eric A. 1971. Teacher characteristics and gains in student achievement: Estimation using micro data. *American Economic Review* 61, no. 2:280–288.

Hanushek, Eric A. 1997. Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis* 19, no. 2:141-164.

Hill, Heather C., Laura R. Kapitula, and Kristin Umland. 2011. A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal* 48, no. 3:794-831.

Hill, Heather C., Brian Rowan, and Deborah Loewenberg Ball. 2005. Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal* 42, no. 2:371-406.

Hill, Heather C., Stephen Schilling, and Deborah Loewenberg Ball. 2004. Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal* 105, no. 1:11-30.

Hill, Heather C., Kristin Umland, Eric Litke, and Laura R. Kapitula. 2012. Teacher quality and quality teaching: Examining the relationship of a teacher assessment to practice. *American Journal of Education* 118, no. 4:489-519.

Jackson, C. Kirabo. 2016. What do test scores miss? The importance of teacher effects on non-test score outcomes. NBER Working Paper no. 22226.

Jackson, C. Kirabo., Jonah E. Rockoff, and Douglas O. Staiger. 2014. Teacher effects and teacher related policies. *Annual Review of Economics* 6, 801-825.

Jacob, Brian A., Jonah E. Rockoff, Eric S. Taylor, Benjamin Lindy, and Rachel Rosen. 2016. Teacher applicant hiring and teacher performance: Evidence from D.C. public schools. NBER Working Paper no. 22054.

Kalogrides, Demetra, Susanna Loeb, and Tara Beteille. 2013. Systematic sorting: Teacher characteristics and class assignments. *Sociology of Education* 86, no. 2:103-123.

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. *Have we identified effective teachers? Validating measures of effective* *teaching using random assignment.* Seattle, WA: Bill & Melinda Gates Foundation.

Kane, Thomas J., and Douglas O. Staiger. 2008. Estimating teacher impacts on student achievement: An experimental evaluation. NBER Working Paper no. 14607.

Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten. 2011. Identifying effective classroom practices using student achievement data. *Journal of Human Resources* 46, no. 3:587-613.

Murnane, Richard J. 1975. *The impact of school resources on the learning of inner city children.* Cambridge, MA: Ballinger.

Neidell, Matthew, and Jane Waldfogel. 2010. Cognitive and noncognitive peer effects in early education. *Review of Economics and Statistics* 92, no. 3:562-576.

Papay, John P., and Matthew A. Kraft. 2015. Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics* 130, 105-119.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. Teachers, schools and academic achievement. *Econometrica* 73, no. 2:417–458.

Rock, Donald A., and Judith M. Pollack. 2002. *Early childhood longitudinal study—kindergarten class of 1998-99 (ECLS-K), psychometric report for kindergarten through first grade.* Publication no. NCES 2002-05. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Rockoff, Jonah E. 2004. The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review* 94, no. 2:247–252.

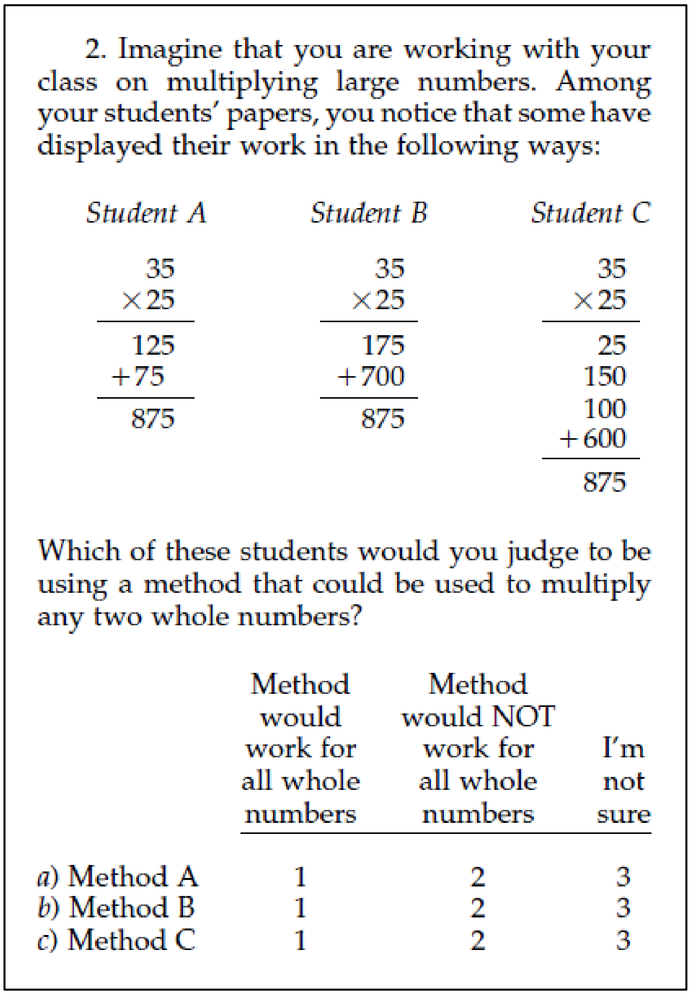
Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger. 2011. Can you recognize an effective teacher when you recruit one? *Education Finance and Policy* 6, no. 1:43-74.

Stein, Mary Kay, and Julia H. Kaufman. 2010. Selecting and supporting the use of mathematics curricula at scale. *American Educational Research Journal* 47, no. 3:663-693.

Eric S. Taylor. 2014. Spending more of the school day in math class: Evidence from a regression discontinuity in middle school. *Journal of Public Economics* 117, 162-181.

Eric S. Taylor. 2015. New technology and teacher productivity. Harvard University Working Paper.

Jan Tinbergen. 1974. Substitution of graduate by other labour. *Kyklos* 27, no. 2:217-226.



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	35	35
×25	×25	×25
125	175	25
+75	+700	150
875	875	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.

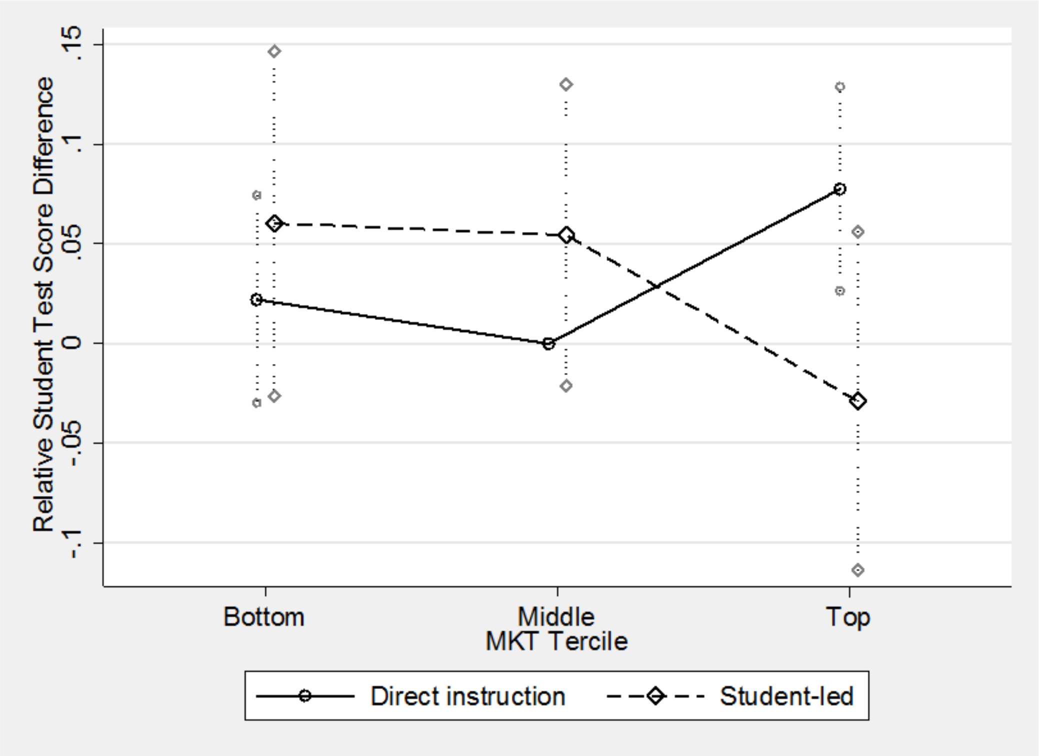


FIGURE 2—STUDENT TEST SCORE DIFFERENCES ACROSS TREATMENT CONDITIONS AND TEACHER MKT TEST SCORE TERCILES

NOTE: Each point on the dashed line and solid line is the estimated mean test score for students in the given MKT-tercile-by-treatment-condition cell, relative to the mean score of students in direct instruction condition assigned to middle-tercile teachers. The vertical dotted lines show 95 percent confidence intervals. All points are estimated in 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-condition 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. Standard errors allow for clustering within schools, the unit of random assignment. The estimation sample includes 5,720 students, 560 teachers, and 80 schools. Sample sizes have been rounded to nearest 10 following NCES restricted data reporting procedures.

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

	Direct instruction	Student-led	Diff. test p-value	Obs.	Joint test p-value
	(1)	(2)	(3)	(4)	(5)
Student characteristics					0.350
Baseline test score, mean	0.017	0.009	0.821	6,600	
	(0.965)	(0.968)			
Baseline test score, variance	0.928	0.934	0.863	6,600	
Days between pre- and post-experiment tests	236.7	238.0	0.494	5,990	
	(6.549)	(9.055)			
Age	7.016	7.016	0.923	5,840	
	(0.626)	(0.626)			
Female	0.490	0.498	0.518	6,420	
Latino	0.317	0.294	0.498	6,060	
African-American	0.264	0.262	0.940	6,060	
English language learner	0.146	0.100	0.066	5,670	
Teacher and class characteristics					0.239
MKT score, mean	-0.565	-0.503	0.241	560	
	(0.477)	(0.477)			
MKT score, variance	0.229	0.214	0.583	560	
Total years experience	11.822	12.930	0.137	570	
	(9.468)	(8.723)			
Years experience current school	7.801	7.969	0.594	540	
	(7.215)	(7.004)			
Female	0.965	0.948	0.266	580	
White	0.596	0.611	0.616	580	
Master's degree	0.435	0.492	0.023	560	
Years with master's degree	3.637	4.544	0.041	580	
	(6.117)	(7.367)			
Training hours previous school year	8.369	9.501	0.673	550	
	(20.04)	(20.52)			
One or more adv. math courses	0.536	0.491	0.326	580	
Class mean baseline test score	0.012	0.003	0.807	590	
	(0.417)	(0.390)			
Class st. dev. baseline test score	0.879	0.885	0.799	590	
School characteristics					0.508
Proportion eligible for free or reduced price lunch	0.490	0.525	0.333	80	
	(0.116)	(0.132)			
Proportion Title I	0.739	0.756	0.825	80	
	(0.335)	(0.265)			

Note: Means (standard deviations) adjusted for randomization block fixed effects. Column 1 "Direct instruction" combines the *Saxon* and *SFAW* treatment conditions. Column 2 "Student-led" is the *Investigations* treatment condition. Column 5 tests the null hypothesis that means are 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—OBSERVED TEACHER BEHAVIORS AND CLASSROOM ENVIRONMENT

	Observed teacher behavior characteristic of…		
	student-led methods	direct instruction	Classroom environment
	(1)	(2)	(3)
Curricula main effects (relative to *Investigations*)			
Expressions	-0.779**	0.577**	-0.173
	(0.119)	(0.122)	(0.109)
Saxon	-0.961**	1.184**	-0.099
	(0.115)	(0.125)	(0.109)
SFAW	-0.430**	1.070**	0.011
	(0.119)	(0.109)	(0.119)
Adjusted R-squared	0.153	0.244	0.033

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. 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 3—STUDENT AND TEACHER ATTRITION

	Student attrited before…			Teacher attrited before…	
	pre-test	post-test		MKT	post-test
	(1)	(2)	(3)	(4)	(5)
Student-led (relative to direct inst.)	-0.002	0.001	0.001	-0.013	-0.002
	(0.008)	(0.009)	(0.009)	(0.023)	(0.026)
Baseline test score (main effect)			-0.020**		
			(0.004)		
Baseline score * Student-led			0.006		
			(0.008)		
Observations	6,740	6,600	6,600	590	590
Dependent variable sample mean	0.021	0.093	0.093	0.059	0.073

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 allow for clustering within schools, the unit of random assignment. 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 AND STUDENT TEST SCORES

	Dep. var. = post-experiment math test score				
	(1)	(2)	(3)	(4)	(5)
MKT score	0.073**	0.030+	0.000	-0.001	-0.002
	(0.024)	(0.016)	(0.011)	(0.011)	(0.011)
Baseline test score controls		√	√	√	√
Student, teacher, class, school covariates					√
Rand. block fixed effects				√	√
School fixed effects			√		
Adjusted R-squared	0.005	0.582	0.622	0.607	0.621
	Dep. var. = pre-experiment math test score				
	(6)	(7)	(8)	(9)	(10)
MKT score	0.056*	0.055**	0.038*	0.038*	0.022
	(0.023)	(0.016)	(0.019)	(0.019)	(0.026)
Student-led (relative to direct inst.)				-0.006	-0.009
				(0.057)	(0.056)
MKT score * Student-led					0.046
					(0.036)
Rand. block fixed effects			√	√	√
School fixed effects		√			
Adjusted R-squared	0.003	0.102	0.060	0.059	0.060

Note: Each column within panels represents a separate regression with student observations. In the top panel, the dependent variable is the student's post-experiment standardized ECLS-K math test score. In the bottom panel, the dependent variable is the pre-experiment or baseline 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. "Class 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 include 5,720 students, 560 teachers, and 80 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—TEACHER MATH SKILLS, ASSIGNED INSTRUCTIONAL PRACTICES, AND STUDENT ACHIEVEMENT

	Dep. var. = post-experiment math test score					
	(1)	(2)	(3)	(4)	(5)	(6)
Student-led (relative to direct inst.)	-0.006	-0.001	-0.017	-0.019	-0.024	-0.020
	(0.031)	(0.030)	(0.031)	(0.031)	(0.031)	(0.030)
MKT score	-0.002	0.024+	0.027+	0.028*		0.027*
	(0.011)	(0.014)	(0.014)	(0.013)		(0.013)
MKT score * Student-led		-0.073**	-0.076**	-0.081**		-0.076**
		(0.022)	(0.025)	(0.024)		(0.024)
Years of experience (standardized)				0.038**	0.057**	0.057**
				(0.012)	(0.014)	(0.014)
Years of experience * Student-led					-0.071**	-0.065**
					(0.021)	(0.021)
Baseline test score controls	√	√	√	√	√	√
Quadratic in teacher years of experience	√	√				
Student, teacher, class, school covariates	√	√				
Rand. block fixed effects	√	√	√	√	√	√

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. "Class 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 include 5,720 students, 560 teachers, and 80 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 6—TEACHER MATH SKILLS AND CLASSROOM PRACTICES

	Dep. var. = practices index			
	(1)	(2)	(3)	(4)
Student-led (relative to direct inst.)	-0.094	-0.071	-0.063	-0.086
	(0.135)	(0.132)	(0.123)	(0.133)
MKT score	0.131*	0.114+	0.131*	0.124*
	(0.060)	(0.060)	(0.054)	(0.052)
MKT score * Student-led	-0.205+	-0.190	-0.362**	-0.351*
	(0.118)	(0.125)	(0.127)	(0.134)
Auxiliary regression to construct practices index				
Baseline test score, student, class, school controls	√		√	
Interact practices vector with treatment condition			√	√

Note: Each column represents a separate regression with teacher observations. The dependent variable is an index of teacher practices. The index is the fitted value after an auxiliary regression of student test scores on a vector of teacher classroom practices variables from classroom observations and teacher survey data (see text for details of these data). The index is then standardized (mean zero, standard deviation one) at the teacher level. For Columns 1 and 4 the auxiliary regression also includes additional controls listed above and described in the note for Tables 4 and 5. The auxiliary regression always includes randomization block fixed effects. However, in all cases the fitted value index (dependent variable) is equal to the practices vector * the estimated practices coefficients. For Columns 3 and 4 the auxiliary regression includes interactions between the practices vector and student-led (treatment condition) indicator. For the regressions reported in this table, all independent variables and coefficients are shown above. All specifications also include randomization block fixed effects. Standard errors allow for clustering within schools, the unit of random assignment. The estimation sample includes 5,720 students, 560 teachers, and 80 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