Learning Job Skills from Colleagues at Work: Evidence from a Field Experiment Using Teacher Performance Data

By JOHN P. PAPAY, ERIC S. TAYLOR, JOHN H. TYLER, AND MARY E. LASKI*

We study a program designed to encourage learning from coworkers among school teachers. In an experiment, we document gains in job performance when high- and low-skilled teachers are paired and asked to work together on improving their skills. Pairs are matched on specific skills measured in prior evaluations. Each pair includes a target teacher who scores low in one or more of nineteen skills, and a partner who scores high in (many of) the target's deficient skills. Student achievement improved 0.12 standard deviations in low-skilled teachers' classrooms. Improvements are likely the result of target teachers learning skills from their partner.

* Papay and Tyler: Brown University, Education Department, Box 1938, 164 Angell St., 2nd Floor, Providence, RI 02912 (emails: john_papay@brown.edu, john_tyler@brown.edu); Taylor and Laski: Harvard University, Graduate School of Education, 13 Appian Way, Cambridge, MA 02138 (emails: eric_taylor@harvard.edu, marylaski@g.harvard.edu). We thank Jonah Rockoff, anonymous reviewers, and seminar participants at Harvard University, Michigan State, NBER Summer Institute, New York Federal Reserve, UC Irvine, and University of Virginia for helpful comments and suggestions. The Bill & Melinda Gates Foundation provided financial support of this research, and we benefited greatly from discussions with our program officer Steve Cantrell. We are equally indebted to the Tennessee Department of Education, and particularly Nate Schwartz, Laura Booker, Tony Pratt, Luke Kohlmoos, and Sara Heyburn for their collaboration throughout. Finally, we thank Verna Ruffin, superintendent in Jackson-Madison County Schools, and the principals and teachers who participated in the program.

Whether and how employees learn job skills from their coworkers, and at what costs, are practical questions for personnel management. Economists' interest in these questions dates to at least Alfred Marshall (1890). In this paper, we report

experimental results for an intervention designed to encourage learning from coworkers among school teachers. Briefly, low-skilled classroom teachers were paired with a higher-skilled teacher working in the same school, pairs were asked to work together to improve their skills, and recommended pair activities were complementary with their day-to-day job responsibilities.[1]

We document meaningful improvements in job performance as a result of teachers working together in skill-matched pairs. The input to our matching algorithm is teachers' scores in nineteen different skills measured in prior evaluations. Each pair includes a "target" teacher who scores low in one or more of the nineteen skills, and a "partner" who scores high in one or more of the target teacher's deficient skills. Our algorithm for matching pairs generates variation in the characteristics of the pairs, and that variation allows us to test empirical predictions consistent with learning from peers and other potential mechanisms. Among other results, we test whether the treatment effects are due to general benefits from working with good peers, broadly defined, or are due to complementarities in peer effects arising from matching on specific skills. The results are consistent with the latter suggesting efficiency gains from active management of coworker pairings.

Our primary focus is classroom teachers and a novel attempt at improving the job performance of existing teachers. We do believe the intervention and results suggest lessons useful for managers in other occupations. Still, effective strategies

[1] The teacher-pairing intervention we study is one (potential) empirical example of learning from coworkers or peers, but there are many other examples which are similar or related. First, a large and growing literature estimates the incidence of and effects of employee training (for a review see Frazis and Loewenstein 2006). Employee training or on-the-job training includes a diverse set of types: (i) formal training, separated from day-to-day work, and often provided by an outside contractor or department specialized in training; (ii) formal apprenticeships; (iii) informal (less-formal) training by a peer or supervisor, often done concurrent with regular job tasks; (iv) learning-by-watching coworkers; and (v) more-solitary forms like learning-by-doing or self-study. Most empirical work in employee training, owing to data limitations, either combines all (most) of these diverse types into "on-the-job training" broadly or focuses on formal training alone. The intervention we study has features of type (iii) and (iv); empirical evidence on these types is rare. Second, coworkers are a particular kind of peer. A large literature examines the role of peers in classroom learning and other formal education settings (for a review see Sacerdote 2010).

for managing the teacher workforce, even if only for the teacher workforce, can have sizable returns for students and economies. Classroom teaching represents a substantial investment of resources: one out of ten college-educated workers in the U.S. is a public school teacher, and public schools spend $285 billion annually on teacher wages and benefits (U.S. Census Bureau 2015, Table 6).[2] Moreover, there is substantial variability in teacher job performance: measured both in the short-run with students' test scores (see Jackson, Rockoff, and Staiger 2014 for a review) and the long-run with students' economic and social success years later as adults (Chetty, Friedman, and Rockoff 2014). Nevertheless, we still know little about what gives rise to these observed between-teacher differences, or how to improve the performance of struggling teachers.[3]

In the study's design, schools were randomly assigned to either treatment—the teacher partnership intervention—or to business-as-usual control. The assignment of teachers to roles and pairs was not random, but rather assigned by an algorithm which we describe in section 1 (and detail fully in Online Appendix B). Importantly, the matching algorithm was carried out in all schools, both treatment and control, prior to random assignment. The treatment effect estimates we report are intent-to-treat estimates based on the roles and pairs as assigned by the algorithm.

We find that the skill-matched partnerships improve teachers' job performance, as measured by their students' test score growth. At the end of the experiment school year, the average student in a treatment school scores 0.06σ (student standard deviations) higher on standardized math and reading/language arts tests

[2] Authors' calculations of workforce share from Current Population Survey 1990-2010.

[3] One exception, and a consistent predictor of differences in teacher performance, is experience on the job (Rockoff 2004, Harris and Sass 2011, Papay and Kraft 2015). The estimated differences due to experience are, notably, much larger than differences due to formal pre-service or formal training (Jackson, Rockoff, and Staiger 2014). Still, even experience predicts relatively little of the total variability in teacher value-added, especially beyond the first few years of a teacher's career.

than she would have in a control school, regardless of whether her teacher participated in a partnership. The gains are concentrated among "target" teachers; in target teachers' classrooms students score 0.12σ higher. These are meaningful gains. One standard deviation in teacher performance is typically estimated to be 0.15-0.20σ (Hanushek and Rivkin 2010, Jackson, Rockoff, and Staiger 2014). These improvements in teacher performance persist, and perhaps grow, in the school year following treatment.[4]

After documenting average treatment effects, we turn to examining mechanisms. In particular we ask: Can the performance improvements be attributed to growth in teachers' skills through peer learning, or are other changes in behavior or effort behind the estimated effects? In section 3, we test a number of empirical predictions motivated by different potential mechanisms. If the underlying mechanism is skill growth, we would predict larger treatment effects for target teachers when the high-performing partner's skill strengths match more of the target teacher's weak areas. We would also predict larger improvements in classroom observation scores for the skills matched, but smaller or no improvements in the skills left unmatched. Both predictions are true empirically. Moreover, while the treatment effects are larger when more skills are matched, the effects do not covary with a broad measure of the target's or partner's prior performance, namely teacher "value added" scores which measure total contributions to test score growth.

The first-order costs of the intervention are the opportunity costs of participating teachers' time. One key potential opportunity cost is that the higher-skilled "partner" teacher might give less effort to teaching her own students. In our

[4] Kane et al. (2011) and Kane and Staiger (2012) do find that classroom observation scores—of the kind we use to match teacher pairs—do predict teachers' contributions to student test scores (teacher value-added scores). That observation scores predict value-added is in contrast to the general lack of good predictors for teacher value-added (beyond experience level, see Jackson, Rockoff, and Staiger 2014). Moreover, other (quasi-)experimental studies, discussed below, show that value added improves as a result of observation based evaluation.

estimates, students of partner teachers were no worse off, if anything they may be benefited slightly. We discuss costs further in the conclusion.

Statistical inference deserves careful attention in this experiment. The study sample includes 14 schools, thus 14 units randomized. We discuss inference in detail in section 2. Practically, we report *p*-values from both a Fisher randomization test and wild cluster bootstrap-*t*. We do find results which are (often only marginally) statistically significant by these inference methods. Our precision is greatly increased by controlling for pre-experiment test scores, i.e., the lagged dependent variable. The precision of our results is consistent with conventional power calculations. Nevertheless, in several places our estimates remain relatively imprecise, and thus sometimes we cannot reject the null hypothesis of no effect even when the point estimate is large.

One contextual feature of the experiment is also important to interpreting these results and their generalizability. The detailed microdata with which teachers were paired are taken from the state's performance evaluation system for public school teachers, which the Tennessee Department of Education introduced in 2011. Among participating teachers and schools the treatment was known as the "Evaluation Partnership Program." These close connections to formal evaluation, and its stakes, likely influenced principals' and teachers' willingness to participate and their effort. We detail the evaluation system and its incentives in section 1. These closely-related incentives are potentially important considerations in the generalizability of our estimates. Prior research, perhaps in settings with weaker incentives, does not consistently find effects of formal mentoring or formal training for teachers (for a review see Jackson, Rockoff, and Staiger 2014).[5] However, teachers in both treatment and control schools in the experiment faced the same

[5] Exceptions include an example of mentoring studied by Rockoff (2008) and an example of training studied by Angrist and Lavy (2001). For a review of (quasi-)experimental evidence on teacher mentoring specifically, but mostly outside of economics, see Kraft, Blazar, and Hogan (2018).

accountability pressures to improve their performance. Thus, the incentives were not a feature of the treatment *per se*, though the incentives may have contributed to teachers' participation and effort in the partnerships.

Among the existing literature on teacher performance, our study is most closely related to Jackson and Bruegmann (2009). Jackson and Bruegmann (2009) also study peer effects on teacher performance, using quasi-experimental variation in coworkers arising naturally when teachers move between schools. A teacher's performance improves when a new higher-performing college arrives at her school; and, consistent with peer learning, the improvements persist after she is no longer working with the same colleague. Our study is complementary but different in two important ways. First, we experimentally generate interactions between teachers. Second, more substantively, we match teachers on specific skill areas, while Jackson and Bruegmann (2009) study low- and high-performing teachers defined very broadly by teachers' total contribution to student test scores (teacher value added scores). As we show below, the benefits of the current intervention arise from skill-matched pairs working together not from broadly low- and high-performing teachers working together.

This paper is also close in topic to a small literature on how evaluation programs affect teacher performance (e.g., Taylor and Tyler 2012, Dee and Wyckoff 2015, Steinberg and Sartain 2015, Bergman and Hill 2018). Taylor and Tyler (2012) find improvements in teacher performance as result being evaluated—evaluation based on classroom observations by specialized evaluators who were recently classroom teachers themselves—and those improvements persist after the evaluation ends. In the present study by contrast, all teachers, treatment and control, are evaluated based on classroom observations; the experiment's teacher pairings are in addition to formal evaluation but in the context of formal evaluation.

Across other occupations and sectors, while there is limited evidence on learning from coworkers specifically, there is a growing literature on productivity spillovers

among coworkers generally. Several papers, each focusing on a specific firm or occupation as we do, also find spillovers; in contrast to our results, however, the apparent mechanisms are shared production opportunities or peer influence on effort (Ichino and Maggi 2000, Hamilton, Nickerson and Owan 2003, Bandiera, Barankay and Rasul 2005, Mas and Moretti 2009, Azoulay, Graff Zivin, and Wang 2010).[6]

Finally, more practically, the impact of this pair matching treatment suggests promising ideas for managing peer learning in the workplace, especially the idea of actively matching on specific skills. In that regard, this paper contributes to the still small literature on the causal effects of management practices (Bloom et al. 2013, Bloom et al. 2014).

Next, in section 1, we describe the treatment in detail, along with other features of the experimental setting and data. In section 2 we describe the average treatment effects and treatment effects by teachers' assigned partnership role. Section 3 discusses potential mechanisms and presents tests of empirical predictions related to those mechanisms. We conclude in section 4 with some further discussion of the benefits and costs.

I. Treatment, Setting, and Data

At each treatment school, several low-performing teachers were each paired with a high-performing coworker, and each pair was encouraged to work together on improving each other's teaching skills. Pairs were matched using microdata from prior performance evaluations, which include separate scores for 19 specific teaching skills (e.g., "asking questions" and "managing student behavior"). Each low-performing "target" teacher was identified as a target because he had low

[6] Morreti (2004) and Battu, Belfield, and Sloane (2003) document human capital spillovers broadly, using variation between firms, but without insight to mechanisms. Moreover, these spillovers may be substantial. Lucas (1988) suggests human capital spillovers, broadly speaking, could explain between-country differences in income.

scores in one or more specific skill areas; his matched "partner" was selected because she had high scores in (many of) the target teacher's deficient skill areas.

In this section we describe the inputs and procedures of the experiment, summarized briefly in the prior paragraph. Additional details are provided in Online Appendix B.

A. Teacher Evaluation in Tennessee

One critical input to the treatment—data measuring each teacher's skills in 19 specific areas—came from Tennessee's formal performance evaluation system for public school teachers. And this formal evaluation system is, more generally, an important feature of the empirical setting. Teachers' and schools' effort in the experiment was likely influenced by incentives, even if weak, created by the evaluation system.[7]

Classroom Observation Scores on 19 Skills.—All public-school teachers in Tennessee are evaluated annually; the performance evaluation includes classroom observations to assess 19 teaching skills. Each teacher is observed and scored multiple times during the course of the school year, typically by the school principal or vice-principal. In our data, and consistent with policy, the average teacher is observed roughly 3½ times a year total (mean 3.6, standard deviation 1.5) by two different observers (mean 2.1, standard deviation 0.83).[8]

[7] The evaluation policies described here, and relevant at the time of the experiment, began in the 2011-12 school year when the state introduced new, more-intensive requirements for teacher evaluation.

[8] Typically, the observer scores only a subset of the 19 skills during an observation, e.g., only scoring "Instruction" during a visit. In our data, the average teacher is scored 1.8 times per year for a given skill (standard deviation 0.64). This is consistent with the state requirements. For the typical teacher—tenured, prior-year overall score of 2-4—the requirement is two ratings for each of the 12 "Instruction" skills, and one rating for each of the 7 "Planning" and "Environment" skills. For non-tenured teachers and tenured teachers who scored 1 the prior year, the requirement is three for the "instruction" skills and two for the other skills. Teachers who scored 5 the prior year are rated just once in each of the 19 skills.

Scoring is guided by a detailed rubric, based in part of the work of Charlotte Danielson (1996). For each skill teachers are given an integer score 1-5 with the labels: (1) "Significantly below expectations," (2) "Below expectations," (3) "At expectations," (4) "Above expectations," and (5) "Significantly above expectations." For each skill the rubric describes specific behaviors and decisions that must be demonstrated to receive a given score. As an example, Appendix Figure A1 reproduces the rubric for "Questioning", one of the 19 skills.[9]

All 19 skills are listed in Table 1 along with summary statistics for teachers' scores in the year prior to the experiment.[10] Teachers score lowest in "problem solving" and highest in "content knowledge." Scores for classroom environment skills are most variable, especially "managing student behavior." The distribution of scores is not centered at (3) "At expectations." The probability of a teacher scoring below 3 ranges from 0.05 to 0.23 across skills, though 41 percent of teachers score below 3 in at least one skill. This left-skew is typical of classroom observation scores (Weisberg et al. 2009, Kraft and Gilmour 2017), and typical of many other performance evaluations by supervisors, sometimes labeled "leniency bias."[11]

[Insert Table 1 About Here]

Another relevant characteristic of the classroom observation skill scores is their reliability. The skill scores we use to match teacher pairs partly reflect teacher's true skill level (signal), but also partly idiosyncratic differences due to the observer,

[9] The full rubric is available at: http://team-tn.org/evaluation/teacher-evaluation/.

[10] Table 1 is based on teachers in the treatment and control schools, but these statistics are similar when calculated using all Tennessee schools.

[11] While left-skewed, there remains useful variation as shown in Table 1. Additionally, Appendix Figure A2 is a histogram of mean teacher observation scores, i.e., each teacher's average of the 19 skill scores. It is roughly Gaussian, though skewed and truncated at 5. This useful variation shown in the table and histogram contrast characterizations in the press, for example, New York Times, March 30, 2013.

the specific lesson and day, the students in the room, etc. (noise). Ho and Kane (2013) estimate the reliability (proportion of total variation that is signal) of similar scores at about 0.45. In our study sample, we estimate a quite similar reliability of 0.51.[12] This reliability is "low" judged against general heuristic rules.

However, we can be more specific about the consequences of "low" reliability for the present experiment. Reliability of 0.50 means that we will make some mistakes in the teacher pair matching: first incorrectly identifying some teachers as weak (strong) in one of the 19 skills, and thus incorrectly pairing some teachers given the goal of weak-to-strong skill matches. To the extent we make such mistakes, the true quality of the matches will be lower than any observed measures of match quality. Thus, first, the potential benefits of teacher pairs will be reduced, at least the benefits that arise from weak-to-strong matches. Second, results which make use of observed measures of match quality will be attenuated. Both of these consequences bias against finding any treatment effect.

Other Features of Evaluation in Tennessee.—Each teacher receives a final annual evaluation score on the same integer 1-5 "expectations" scale described above. The final score combines (i) the classroom observation scores described above, and (ii) measures of teacher contributions to student test scores, each of the two given equal weight.[13] Student test score data, while certainly used in the state's formal evaluation scores, were not used in the matching of teachers for the experiment, nor in our communication with teachers about the goals of the teacher partnerships. By contrast, the experiment's partnerships were designed and described as a way to

[12] The comparison is not apples to apples. Ho and Kane (2013) have videos of lessons, and so they can more cleanly estimate the variance components by having multiple observers score the same lesson video. Thus, our 0.51 estimate may understate the reliability as defined in Ho and Kane (2013). Moreover, as shown in Ho and Kane (2013), reliability rises with multiple observations and multiple observers, both features of the Tennessee system and our data; that added reliability is not reflected in our estimate of 0.51.

[13] For the grade 4-8 math and reading/language arts teachers in our study, the 0.50 weight to (ii) is divided: 0.35 to a conventional teacher value-added score, and 0.15 to another teacher self-selected student achievement measure.

improve observation scores—scores salient to teachers because they contribute to formal performance evaluations.

What consequences or incentives are attached to these scores? The clearest consequences are in rules for granting tenure, dismissing teachers, and future evaluation requirements. New teachers must score 4 or 5 in two consecutive years to receive tenure.[14] Scoring 1 or 2 can be used as grounds for dismissal, even among tenured teachers, though empirically dismissal rates are quite low. Regarding future evaluation, tenured teachers who score 1 in a given year are observed more often the following year (2-3 times instead of 1-2); they are, in effect, treated as non-tenured teachers for the class observation requirements of evaluation. While teachers, tenured or not, who score 5 are observed just once the following year. For some teachers the threat of additional (fewer) observations may be motivating.

Other potential consequences or incentives are less well defined or implicit. State policy regulations allow evaluations to be used "to inform human capital decisions, including, but not limited to individual and group professional development plans, hiring, assignment and promotion, tenure and dismissal, and compensation"; however, these uses are not required.[15] Notably, we are aware of no compensation related to evaluation in our study district. Additionally, even absent formal immediate consequences, teachers likely understand that evaluation reveals (potentially) new information about their performance to others, and thus teachers may have career concerns incentives.[16]

[14] Tenure in Tennessee, a right to work state, is less of a benefit then the traditional characterization of teacher tenure. Additionally, new teachers cannot qualify for tenure until after their fifth year teaching.

[15] Tennessee State Board of Education Policy 5.201, Teacher and Administrator Evaluation Policy. This quotation appears in the policy dated April 19, 2013, but has not changed and appears in the policy dated April 20, 2018.

[16] There is some indirect evidence that teachers in Tennessee recognize the consequences of their evaluation scores. Using a regression discontinuity design, Koedel, Li, Springer, and Tan (2017) show that teacher self-reported job satisfaction jumps discontinuously at the cut scores which define Tennessee's 1-5 integer rating bins. The researchers have the underlying continuous scores which are converted to the 1-5 integer scores.

B. Sample

The experiment was conducted at 14 elementary and middle schools (7 treatment, 7 control) in a medium-sized district in Tennessee. Jackson-Madison County School System is the 12th largest in the state enrolling approximately 13,000 students. The district's typical student is economically disadvantaged (77 percent of students) and African-American (61 percent, 32 percent are white and 7 percent are Hispanic). The district spends about $9,750 per pupil annually. In 2013-14, the state of Tennessee's measure of student test score growth placed Jackson-Madison in the lowest-performing category for districts.

In the summer of 2013, principals from all 21 of the district's elementary and middle schools were briefed on the program and 14 volunteered to participate in the study. During the experiment year 2013-14, those 14 schools employed 136 teachers for math or reading/language arts in grades 4 through 8.[17] These 136 teachers are the focus of our paper because these are the grades and subjects where we observe pre- and post-experiment student achievement measured by state tests. Descriptive information on the students and teachers is provided in Table 2.

C. Experimental Procedures and Treatment

Identifying and Matching High- and Low-Performing Teachers.—In October 2013, prior to random assignment, the research team created at list of recommended teacher pairings for each of the 14 participating schools. The matching algorithm which created those pairings is summarized here and described in detail in Online Appendix B. To simplify exposition we describe the process for one school; the same process is simply repeated for each school.

[17] We discuss attrition below. We observe student test scores for 136 teachers at the end of the experiment school year. At the time of random assignment, our sample included 141 teachers.

(Step 1) Make a list of low-performing "target" teachers. Teacher j is said to be "weak" in a given skill s if his prior-year pre-experiment score in that skill $R_{js} < 3$. The threshold of 3 corresponds to the label "At Expectations," as described above; and scoring less than 3 is relatively rare empirically, as reported in Table 1. Between 5 and 23 percent of teachers score below 3 on a given skill. A teacher is a "target" teacher in the experiment if (i) he has one or more "weak" skill areas, and (ii) his average skill score $\bar{R}_j \leq 3$. In our 14 schools, 30 percent of teachers are "target" teachers.

(Step 2) Make a list of high-performing teachers who are *potential* partners for the target teachers. A teacher is a potential "partner" if (i) she was not identified as a target in step 1, and (ii) she is "strong" in one or more skill areas, $R_{js} \geq 4$.

(Step 3) Match "target" and "partner" teachers in pairs. The key constraint is that pairs must be one-to-one, i.e., each target teacher is in (at most) one pair, and each partner is also in (at most) one pair. Even under this constraint, however, there are many possible one-to-one pairings of teachers for a given school. We define the optimal set of pairings, p^*, as follows. Let (j,k) be a single potential pair, target j paired with partner k; and let p be a set of possible pairings under the one-to-one constraint, $\{(j,k),(j',k'),\dots,(j'',k'')\}$. To optimize we need some measure to differentiate between options, $M(p)$. In this experiment we used $M(p) = m(j,k) + m(j',k') + \cdots + m(j'',k'')$, with $m(j,k) = \sum_{s=1}^{19} 1\{R_{js} < 3\} \times 1\{R_{ks} \geq 4\}$. In words, the "match quality", m, of a potential pair, (j,k), is the number of skills where the target, j, is "weak" and the partner, k, is "strong." Then $p^* = \underset{p}{\operatorname{argmax}}\big(M(p)\big)$.[18]

[18] This optimization problem can be solved with algorithm first described by Kuhn (1955), sometimes called the Kuhn-Munkres algorithm or Hungarian algorithm.

This matching process was carried out for all 14 schools and 141 teachers using pre-experiment classroom observation data. In the final pairings, just over half of teachers were in a pair (26.4 percent target teachers, and an equal fraction of partners given the one-to-one constraint). Then after randomization the list of "recommended" one-to-one pairings, p^*, was sent to principals in treatment schools only. In our analysis, we use the p^* pairings, and thus our estimates are in that sense intent-to-treat.

Treatment Inside Schools.—At each treatment school, the principal was responsible for introducing each target-partner pair to each other, explaining why the two had been paired, and encouraging the pair in their work together. In October 2013, we provided each treatment school principal with a spreadsheet listing each target teacher and his recommended matched partner teacher, i.e., the p^* one-to-one pairings described above. An example spreadsheet is shown in Appendix Figure B1. The spreadsheet listed all 19 skill areas indicating, or each pair, the target and partner had a weak-to-strong skill match. Accompanying the pair report was a set of talking points to help principals introduce teachers to the program, and program guides for teachers (see Appendix Figures B2 and B3).[19]

Program guides provided to teachers (by their principal) encouraged teacher pairs to engage in several different activities together. The list of activities included: set goals for the partnership, review each other's prior evaluation results, observe each other teaching in the classroom, review lesson plans, develop strategies for improvement and share advice, and follow-up on each other's commitments. The

[19] Six of the seven treatment principals introduced pairs; and thus, at least at this extensive margin, the school took up the treatment.

program guide also included a recommended timeline and general tips for working and communicating in a partnership.[20]

Control schools continued with business as usual. Both control and treatment schools continued with their existing novice mentoring programs, which pair new hires with an experienced overall high-performing mentor. Among study schools, this is the activity closest to the experiment's skill-matched pairs, though much smaller in scope and pairing teachers on different criteria. Such programs for novices are common in schools (Kraft, Blazar, and Hogan 2018). Control schools did not pair teachers in any new ways, and certainly not using the experiment's matching algorithm.[21]

Random Assignment.—On October 2, 2013, the research team randomly assigned schools to treatment and control. The 14 schools were placed in seven randomization pairs (blocks), and one school randomized to treatment within each pair. Pairs were defined by (i) school level, elementary or middle, and then (ii) within level, by pairing on student enrollment.

Interpreting the results of this experiment as causal effects rests largely on the success of random assignment. Table 2 reports the traditional test of randomization, comparing pre-treatment characteristics of students, teachers, and teacher pairs. These tests include fixed effects for the randomization block pairs. Most differences are relatively small.[22] None of the 13 characteristics show a statistically significant

[20] We have little data on what teacher pairs actually did (or did not) do after being introduced by their principal. What we know is discussed in more detail in Online Appendix B. In brief, self-reports from surveys (and activity logs from some pairs) suggests participating teachers did, broadly speaking, observe each other teaching, discuss their own prior evaluation results, and discuss feedback and strategies for improvement.

[21] All school principals, both treatment and control, already had access to the data used to create the pairings. Thus, a control school principal with some data skills and a little time (or a helpful friend) could create lists of skill-by-skill scores for her school's teachers. However, the matching algorithm was not revealed in detail to any principals. Additionally, when asked, none of the seven control principals described creating such reports or attempting any new matching program during the experiment year.

[22] Note that the teacher value-added scores are in teacher standard deviation units, not student standard deviations as is often the case. In student standard deviations the differences are roughly one-tenth to one-fifth the magnitudes in Table 2 and Appendix Table A1.

difference between treatment and control means. However, as discussed further in section 2, precision is limited in this experiment. While we do find statistically significant treatment effects on outcomes, those estimates are benefited by controlling for lagged dependent variables which we do not have in these balance tests. In Appendix Table A1, we check for covariate balance separately for teachers in each assigned program role: low-performing target teachers, high-performing partners, and teachers not assigned a role. The results are broadly similar. If anything, treatment target teachers have lower prior value-added scores and slightly less experience, but higher prior classroom observation scores. The patterns in Table 2 may be driven by teachers who were not assigned a role by the algorithm.

[Insert Table 2 About Here]

D. Data

The Tennessee Department of Education provided data for this experiment, covering 2010-11 through 2013-15. As described above, we use classroom observation microdata, from pre-treatment years, to match teacher pairs. We also use post-randomization classroom observation scores as an outcome measure. Other data used in this paper are state administrative records on (i) student scores from annual state standardized tests in math and reading/language-arts in grades 3 through 8, (ii) information on student demographics and special educational programs, (iii) records linking each student to her assigned teacher(s) for each subject each year, (iv) and information about teacher experience and prior performance. We standardize all test scores (mean zero, standard deviation one) within year-grade-subject cells using the statewide distribution (as opposed to the moments from the specific study district).

II. Effects on Student Achievement

In this paper, we ask two primary empirical questions: First, did the treatment—pairing classroom teachers to work together on improving skills—benefit (or harm) teacher performance? Second, if there were improvements, is there evidence that those improvements are the result of growth in teachers' skills from peer learning? Our primary measure of teacher job performance is differences in student test scores.

The estimated average treatment effect, across all teachers in treatment schools, is an improvement of 0.06 student standard deviations (σ) on tests covering math and reading/language arts. We estimate this treatment-control difference in means, δ, by fitting the regression specification

$$A_{ijkt} = \delta T_{s(j)} + \boldsymbol{X}_i\beta + \pi_{b(s)} + \varepsilon_{ijkt}, \quad (1)$$

where A_{ijkt} is the end-of-year t (the experiment year) test score in subject k (math or reading/language arts) for student i assigned to teacher j in school s.[23] The treatment indicator, $T_{s(j)}$, varies only at the school level s. All estimates throughout the paper include randomization block fixed effects, $\pi_{b(s)}$. The vector $\boldsymbol{X}_i$ includes additional covariates included to improve precision. The additional covariates include student i's prior achievement, $A_{i,t-1}$, as well as her gender, race/ethnicity, English language learner status, and special education status.[24] The vector $\boldsymbol{X}_i$ also includes a pre-experiment "value added" measure of teacher j's contributions to

[23] The student-teacher link records allow students to be linked to more than one teacher for a given subject, though three-quarters in our sample are linked to just one teacher. When a student has more than one teacher, the state assigns a "percent responsibility" to each teacher. When a student has two or more teachers, we include one observation for each student-by-teacher pairing and weight proportional to the "percent responsibility." But our results are robust to assigning students to the one teacher with the highest weight.

[24] We have very little missing data in X_i, for example, less than 4 percent of students are missing baseline achievement. When baseline achievement or another given covariate is missing, we replace it with a value of zero and include an indicator = 1 for all students missing the given covariate. Our results are robust to excluding these approximately 4 percent of students.

student test scores in subject k.[25] Last, $\boldsymbol{X}_i$ includes grade-by-subject fixed effects and we allow the slope on prior achievement score to differ by subject and grade. Estimates of $\hat{\delta}$ are shown in Table 3.

A. Statistical Inference

Statistical inference deserves careful attention in this experiment. In the simplest terms we have an experiment with 14 observations (schools) randomized. When the number of clusters (units randomized to treatment) is sufficiently large, in the asymptotic sense, the typical approach to inference for specifications like equation 1 is to apply the standard heteroskedasticity-cluster-robust correction to the standard errors with schools as the clusters.[26] However, we have 14 clusters to estimate specification 1, and as Cameron, Gelbach, and Miller (2008) show, cluster-robust standard errors are biased downward when the number of clusters is small.

In response to this potential bias we report two approaches to inference throughout the paper. First, we report *p*-values from the wild cluster bootstrap-*t* (WCBt) method suggested by Cameron, Gelbach, and Miller (2008) which builds on the more-common wild bootstrap by resampling at the cluster level.[27] Second, we also report *p*-values derived from randomization inference, often called the Fisher randomization text (FRT) (Fisher 1935; for modern treatments see Rosenbaum 2002, Imbens and Rubin 2015). In this experiment there were $2^7 = 128$ possible outcomes of the random assignment procedure (seven blocks with two schools in each block). We calculate a "treatment effect" estimate, $\hat{\delta}_r$, for each

[25] Teacher value added scores come from the state's performance evaluation system. The scores are known as "TVAAS scores" for Tennessee Value-Added Assessment System.

[26] Errors may also be correlated within a group of students taught by the same teacher. In our setting teachers are nested within schools. Thus, clustering errors at the school level is equivalent to clustering at both the school and teacher levels simultaneously (Cameron, Gelbach, and Miller 2011).

[27] Following Cameron, Gelbach, and Miller (2008), we (i) do not impose the null hypothesis; and (ii) use Mammen weights. *p*-values using Rademacher weights and imposing the null are both similar but slightly smaller ("less conservative"). Throughout the paper we use 500 replications.

of the 128 possible configurations, *r*, keeping everything about estimation constant except T_{sr}. We calculate the FRT *p*-value based on the position of our actual estimate in absolute value, $|\hat{\delta}|$, in the distribution of $|\hat{\delta}_r|$.[28]

Each of these two approaches is preferable, at least on theoretical grounds, to the standard cluster-robust approach, but neither is clearly preferable to the other. Randomization inference provides exact *p*-values, even with small samples, under the sharp null hypothesis of no effect at all on outcomes.[29] For inference about this "no effect at all" null hypothesis the FRT *p*-values we report should be used. However, the no effect *at all* null is not the same as the more-conventional null hypothesis of no effect *on average*. The latter is arguably of more interest to managers and policymakers, and it is possible to reject the null of no effect at all even if there is no average effect. The FRT *p*-values can also be used to test the "no average effect" null, but (may) lose their exact property in that test (for a review see Chung and Romano 2013). Indeed, for the null of no average effect and its alternative, the FRT approach can have less power than approaches that instead rely on distributional assumptions, like the WCBt (Ding 2017).[30] In addition to distributional assumptions, the theoretical justification for the WCBt also relies on an asymptotic argument. Nevertheless, in empirical tests Cameron, Gelbach, and Miller (2008) show that the wild cluster bootstrap-*t* approach produces rejection rates quite close to the nominal size of the test even for as few as 5-10 clusters.

At the top of Table 3 we compare the different approaches to inference empirically in our data. Panel B shows estimates from fitting equation 1 with and

[28] The FRT *p*-value = $\sum 1\{|\hat{\delta}_r| \geq |\hat{\delta}|\}/128$.

[29] In the potential outcomes framework, this sharp null of no effect at all means that for every subject the potential outcome under the treatment is equal to the potential outcome under the control. Thus we can form the null distribution by "filling in" the missing potential outcomes based on observed values. This approach requires no distributional assumptions or asymptotic arguments.

[30] As Ding (2017) points out, this difference in statistical power is a plausible explanation for why empirically we might reject with WCBt but fail to reject with FRT. Even though rejecting the "no average effect" null logically implies rejecting the "no effect at all" null.

without additional covariates. For panel A we estimate a regression with school mean test scores,

$$\bar{A}_{st} = \delta T_s + \beta \bar{A}_{s,t-1} + \pi_{b(s)} + \nu_s, \tag{2}$$

to focus on the core variation in the experiment's design. In the simplest case, panel A column 1 with its point estimate of 0.051, the wild bootstrap-*t* *p*-value is 0.30 while the FRT *p*-value is somewhat larger at 0.38.[31] Indeed the randomization inference *p*-value is larger than the wild (cluster) bootstrap-*t* *p*-value in nearly all cases in the paper, and in that sense randomization inference is empirically the more "conservative" approach to inference in this experiment.

The results in Table 3 emphasize the precision gains from including lagged test scores. Compare *p*-values across the columns 1 and 2. In panel A with school-level regressions, column 2 includes a single control for school average baseline score, $\bar{A}_{s,t-1}$, but column 1 does not. When we control for baseline scores, the randomization inference *p*-value is 0.047. While there are only 14 clusters (observations) in the experiment, there is much less residual variation than might be expected in other settings because (school average) lagged test scores are highly correlated year to year.

Table 3, and the tables which follow, show the empirical precision of our estimates. One important benchmark for comparison is the relevant power calculation. We calculate the minimum detectable effect size (MDES) at 80 percent power and α level of 0.05 under the following assumptions: (i) a sample size of 14 schools with 10 teachers per school, (ii) pre-treatment covariates explain 65 percent of the variation in student test scores,[32] (iii) the between-teacher standard deviation

[31] With school observations, in panel A, the wild cluster bootstrap-*t* is simply a wild bootstrap-*t*.

[32] This estimate is the *r*-squared obtained from estimating equation 1 using data from pre-experiment years (i.e., 2010-11 through 2012-13) omitting the treatment indicator. The estimate is only slightly smaller, 63 percent, if the right hand side includes only controls for prior year test score.

in value added (contribution to student test scores) is 0.15, and (iv) the intra-class correlation (ICC) for teacher value-added scores within schools is 0.15. Under these assumptions, the MDES is 0.094σ. For comparison, our average treatment effect estimate is 0.065σ with an FRT *p*-value of 0.250, and for target teachers 0.123σ with an FRT *p*-value of 0.031. Thus our realized results are generally consistent with the power calculations. If we redo the power calculations changing only the α level to 0.25 (0.03), the MDES is 0.063σ (0.103σ). The key benefit to power is that baseline test score controls explain a substantial share of the variation in test score outcomes.

A related concern, arising from the relatively small number of schools, is that the estimated average treatment effect may be due to large benefits for only one school and little or no effect elsewhere. Under the FRT sharp null of no effect at all, for example, we would reject if only one school benefited. To examine this concern we estimate specification 1 (as in Table 3 column 2) iteratively dropping one of the treatment schools, and its paired control, from the estimation sample. The seven point estimates range from 0.037 to 0.071 with an average of 0.056.[33] In a second test we compare each treatment school's individual mean to the overall control mean; we estimate specification 1 without block fixed effects, keeping one treatment school and all control schools in the estimation sample. The seven point estimates range from -0.022 to 0.093 with an average of 0.040.[34] Taken together, these tests suggest the overall average treatment effect is not driven by one or two schools, but the benefits were likely not uniform.[35]

[33] The seven estimates and their FRT *p*-values are: 0.037 (0.469), 0.046 (0.438), 0.048 (0.438), 0.059 (0.313), 0.064 (0.250), 0.069 (0.188), and 0.071 (0.188).

[34] The seven estimates and their FRT *p*-values are: -0.022 (0.734), 0.0001 (1.000), 0.013 (0.828), 0.033 (0.609), 0.076 (0.250), 0.085 (0.300), and 0.093 (0.063).

[35] Additionally, if the overall estimate was caused by just one school benefiting, that would imply an implausibly large benefit for that one school, i.e., 0.06σ / (1/7) = 0.42σ.

[Insert Table 3 About Here]

B. Average Treatment Effects on Student Achievement

Turning now to the substantive result, students in treatment schools score about 0.06σ higher, on tests covering math and reading/language arts, than their peers in control schools. These are intent-to-treat effects, and do not use any variation in assigned teacher roles or variation in treatment take-up.

These positive average treatment effects are educationally and economically meaningful. Gains of 0.06σ represent roughly one-third of a standard deviation in teacher performance, which is typically estimated at 0.15-0.20σ in math and somewhat smaller in reading (Hanushek and Rivkin 2010, Jackson, Rockoff, and Staiger 2014). A difference of 0.06σ is also roughly one-quarter the estimated gain from reducing class size by 30 percent in elementary grades (Krueger 1999), or 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). However, unlike reducing class size or increasing class time, the current treatment—pairing classroom teachers to work together on improving skills—would require a much smaller increase (if any) in teacher salary expenditures.

These gains can be interpreted as causal effects of treatment under the standard experimental design assumption: At the beginning of the experiment, there was no difference in potential outcomes—student achievement growth, teacher or school performance, etc.—between treatment and control samples at expectation. This assumption is reasonable with the random assignment design. The conventional tests of pre-treatment covariate balance are in Table 2. Additional evidence of successful random assignment is, in Table 3, the consistency of point estimates with different pre-experiment controls. Further, we find no difference in the test score

trends, comparing treatment and control schools, for three years leading up to the experiment (see Appendix Figure A3).

Attrition after random assignment can also limit causal inference. While no school attrited *per se*, the set of teachers in a school are a first-order characteristic of the school, and so any teacher attrition is relevant to understanding changes over time in school-level effects. Empirically teacher attrition during the experiment year was quite rare, and well balanced across conditions. Of the 141 teachers at the beginning of the school year, we have end-of-year student test score data for 136 (96.5 percent). As detailed in Online Appendix C, we find no different in attrition between treatment and control schools, and zero target teachers attrited. Because attrition rates are low and balanced, even Manski-style bounds (e.g., Horowitz and Manski 2000) are relatively tight at 0.038σ to 0.088σ for the average treatment effect.

C. Treatment Effects for Target Teachers and Other Teachers

Next we estimate treatment effects separately for teachers with different roles in the partnership program. The experiment was designed first to improve the job performance of low-performing "target" teachers; thus, the estimates in Table 3 may mask important heterogeneity by role. Teachers were assigned to one of three roles: (i) low-performing target teachers, (ii) high-performing potential partner teachers, and (iii) all other teachers who were not assigned a role in partnerships.

To test for differences by assigned teacher role we simply interact role with treatment. As detailed in section 1, teacher roles were assigned by an algorithm before random assignment. Thus for each control school we know who the target and partner teachers *would have been* if the school had been assigned to treatment, though that information was only known to the researchers. These data on (potential) roles allow us to estimate the specification,

(3) $A_{ijkt} = \delta^T\left(T_{s(j)} * Target_{j(i)}\right) + \delta^P\left(T_{s(j)} * Partner_{j(i)}\right) + \delta^N\left(T_{s(j)} * NoRole_{j(i)}\right) + \alpha^T Target_{j(i)} + \alpha^P Partner_{j(i)} + \boldsymbol{X}_i\beta + \pi_{b(s)} + \epsilon_{ijkt}$,

where $Target_j$, $Partner_j$, and $NoRole_j$ are a set of mutually-exclusive and exhaustive indicator variables for teacher *j*'s role. The role main effect and interaction terms (i.e., the first five right-hand-side terms) replace $\delta T_{s(j)}$ in specification 1. All other details of estimation are the same as described earlier for specification 1. The estimates from specification 3 are best interpreted as intent-to-treat because the role indicator variables are based on the algorithmic, pre-randomization assignment of roles; the estimates do not incorporate any post-randomization endogenous decisions by principals or teachers.

Treatment effects are largest for low-performing target teachers. As reported in Table 3 panel C, test scores improved by 0.12σ for the students of target teachers (FRT *p*-value 0.031, WCBt *p*-value <0.001). To be clear about the comparison being made, 0.12σ is the difference in performance between (i) treatment-school teachers who were target teachers, and (ii) control-school teachers who *would have been* target teachers. The comparison is not between (i) and all teachers in control schools; specification 3 includes main effects for role. In contrast to the treatment effect for target teachers, the estimates for high-performing partner teachers are much smaller. The point estimates for partners (and no role teachers as well) are smaller than those for target teachers, but not unimportant educationally; however, we lack the precision to make any strong claims about these potential positive effects.

Again, these improvements for target teachers are meaningful. A gain of 0.12σ is roughly equivalent to the difference between being assigned to a median teacher instead of a bottom quartile teacher. A gain of 0.12σ is also at least as large as the difference in performance between a novice teacher and a 5 to 10 year veteran (Rockoff 2004, Papay and Kraft 2015). Learning from peers may be one

mechanism for the returns to experience. However, the specific treatment gains we study are not necessarily substituting for experience gains. In Appendix Table A3 we test whether the treatment effect for target teachers is heterogeneous by prior experience. If anything the effect is larger for more experienced teachers, though the differences are not statistically significant.

Moreover, 0.12σ may underestimate the effect of treatment on teachers who actually participate in the program. Appendix Table A2 reports estimates of program take-up by teacher role (a "first stage" by role).[36] Among treatment teachers assigned by the algorithm to the target role, 61 percent participated in the program in the target role, suggesting a treatment-on-the-treated estimate of about 0.20σ $(= 0.12/0.61)$. These improvements are large but similar to the gains documented by Taylor and Tyler (2012) for the ex-ante lowest performing teachers who are similar to the current study's low-performing target teachers.

One potential explanation for the gains in target teacher classrooms, 0.12σ, is that treatment affected how students were assigned to teachers; a sorting explanation as opposed to a true change in teacher job performance. Treatment school principals knew who the target teachers were, and perhaps tried to help target teachers by giving them more high-scoring students. While we cannot test this explanation directly, we argue that it is unlikely. First, treatment assignment occurred after the start of the school year, leaving limited opportunity for changes in assignments. Empirically we find no treatment effect on the probability that student *i* switched teachers during the year.[37] Second, student assignments are likely zero-sum: assigning more high-scoring (fewer low-scoring) students to target teachers likely

[36] The sample and specification are identical to Table 3 panel C following specification 3, except that the dependent variables are indicators for participation in a specific role.

[37] The dependent variable is an indicator = 1 if student *i* was assigned to more than one teacher during the year for subject *k*, which we regress on the treatment indicator and block FE. The estimated coefficient for treatment is -0.007 with a randomization inference *p*-value of 0.84. We also test for heterogeneity by interacting treatment with baseline test score; the main treatment effect is -0.001 (*p*-value 0.89) and the interaction coefficient is -0.05 (*p*-value 0.23).

requires assigning more low-scoring (fewer high-scoring) students to other teachers. If student sorting were the explanation we might expect negative treatment effects for partner and no-role teachers, but, as reported in Table 3, this is not the case.[38]

D. Treatment Effects After Two Years

Last, we estimate treatment effects on the performance of teachers in the year following treatment. Table 3 column 3 shows estimates of specifications 1 and 3 using test score data for students taught by treatment and control teachers in the next school year, 2014-15.[39] In other words, we follow teachers to the next group of students they teach and observe their performance. The improvements in teacher job performance persist, and perhaps grow, in the year after treatment. However, the estimates are more imprecise than elsewhere in the paper. The estimated effect for low-performing target teachers is 0.25σ (FRT *p*-value 0.422, WCBt *p*-value 0.068, column 3 panel B). The estimated average treatment effect is also larger at 0.11σ, though not statistically significant (FRT *p*-value 0.375, WCBt *p*-value 0.220, column 3 panel A).

Teacher attrition is an important consideration in interpreting these second-year results, more important than in the experiment year results. The sample for Table 3 column 3 includes 96 of the original 141 treatment and control teachers (68.1 percent). Some teachers has stopped working in Tennessee public schools entirely; others had switched to teaching non-tested grades or subjects. As reported in Appendix Table C2, treatment teachers were 4.6 percentage points more likely to

[38] This is an imperfect test; partner and no-role teachers may be better at teaching otherwise low-scoring students.

[39] Some of the 136 teachers in the sample switched schools between 2013-14 and 2014-15. As long as a teacher is working in a Tennessee public school teaching grades 4-8 we include her in the sample. In these cases "school", *s*, in specifications 1 and 3 for purposes of estimation is her school during 2013-14. For all teachers "role" is her program role during 2013-14.

attrit, and target treatment teachers 6.7 points more likely. These differences are not statistically significant, though as elsewhere we have limited precision.

To assess the potential influence of this teacher attrition, we estimated bounds for the treatment effects following the two approaches described by Manski and Horowitz (2000) and Lee (2009). Details are provided in Online Appendix C. Given the roughly one-third attrition rate, the Manski-style bounds are quite wide, for example, ranging from -0.132σ to 0.277σ around the school-level average effect estimate of 0.106σ. The Lee-style bounds are 0.051σ to 0.141σ, though even for the upper bound we cannot reject the null of no effect.

III. Growth in Teachers' Skills and Other Potential Mechanisms

The previous section documents educationally meaningful and economically significant effects on teacher performance, but we would also like to understand the mechanisms through which the peer pairings influence performance. In this section we take up our second empirical question: Can the improvements in student learning—the positive average treatment effects—be attributed to growth in teachers' skills from peer learning? Or are other changes in behavior or effort behind the treatment effects?

In treatment schools, low-performing teachers were paired with a high-performing partner, and each pair was explicitly asked to work together on improving teaching skills. Thus our first hypothesized mechanism is changes in teachers' skills. Nevertheless, there are two other categories of potential mechanisms: changes in teachers' motivation or effort, and changes in shared resources or tasks (joint production).[40] These are not mutually exclusive; all three could be contributing, in varying degrees, to the average treatment effects. Jackson

[40] Learning new skills requires effort, as Jackson and Bruegmann (2009) point out. The reverse is not true, however, workers can increase (decrease) effort even if skills remain constant. Thus "changes in motivation and effort" is a potential explanation for the treatment effects even if there were no changes in skills, and so we list it separately.

and Bruegmann (2009) describe how these three categories of teacher spillovers likely affect performance in a typical school context. We briefly discuss how these mechanisms might operate in our setting, and then turn to some empirical tests.

The second category of potential mechanisms is changes in teachers' motivation or effort. Asking a low-performing teacher to spend more time with a high-performing colleague and talk together about evaluation results and performance explicitly, may have made her more optimistic or enthusiastic about work or made her more embarrassed about her poor performance. Similarly, treatment teachers may have felt more accountability to their new partner. These interactions may, in turn, lead to increased effort: either (i) transitory increases in effort, for example, motivated by specific accountability to one's partner or direct monitoring by ones' partner; or (ii) lasting increases in effort, for example, finding a new preferred equilibrium level of effort as a result of interacting with one's partner. There is evidence for coworker effects on effort outside the education sector (for example Mas and Moretti 2009, even if in a very different kind of job).

A related but distinct hypothesis is that the treatment gives target teachers new information about their performance; and that information alone, without any peer effect, causes target teachers to increase effort or invest in skill development. Individual teachers, in treatment and control, knew their prior evaluation scores, including both the various components and the final formal overall 1-5 integer (see section 1). What then might be new information? Individual teachers know their own scores, but teachers do not know their own rank nor do they know other teachers' scores. The only public comparison information is the proportion of teachers in each of the 1-5 integer categories. A treatment target teacher would learn, from his principal, that his matched partner is higher-performing in specific skill areas; that might or might not be a surprise. The partner might then reveal more details once the pair starts working together. But the target would not learn about other teachers beyond the partner. In short, the scope for new information is

limited to one other comparator, but still might induce more effort from some teachers.

For any effort mechanism hypothesis, it is important to note that both treatment and control teachers—and particularly the low-performing target teachers—already face the same first-order incentives to increase their performance. This includes incentives to improve observation scores—the focus of the intervention—and to improve student test scores on which evaluations are also partly based (see section 1). The treatment may, however, have made a target teacher's low skill scores more salient to the teacher by giving more attention to observation scores than otherwise would have been given.

The third category of potential mechanisms is changes in teachers' opportunities to share resources or production tasks. Teacher partnerships formed by the treatment program may have expanded to activities outside the original program scope. For example, teachers paired by the treatment may have been more likely to share existing lesson plans or cooperate in creating new lessons in ways that benefited productivity. Of course, teachers in control schools could exchange lessons or collaborate in other ways; thus, to explain treatment effects, the shared production must be a specific result of the new pairing.

A. Empirical Tests of Mechanisms

Our main purpose in this section is to present empirical evidence, where available, to help discriminate among these three potential mechanisms. We have already seen two important empirical results that are consistent with teacher skill growth. First, improvements in performance are largest for the low-performing target teachers; the average effect may have been entirely driven by target teacher improvements. Second, the performance improvements appear to have persisted in the year after treatment ended. However, the pattern of larger effects for target

teachers could result from an asymmetric changes motivation or effort, or asymmetric benefits of new resource or task sharing.

In the remainder of this section, we present several empirical tests. In short, we examine whether treatment effects for target teachers covary with the characteristics of teacher pairings in ways that would be consistent with skill growth or the other potential mechanisms.

First, if low-performing target teachers did learn new skills from their high-performing partner, we would expect larger treatment effects when the partner teacher's specific skill strengths matched the target teacher's specific weaknesses. We test this prediction in Table 4 column 1 by interacting the treatment indicator with the proportion of target teacher j's weak skill areas matched by her recommended partner's strong skill areas.[41] The specification for column 1 is

(4)

$$A_{ijkt} = \delta^T\left(T_{s(j)} * Target_{j(i)}\right) + \delta^P\left(T_{s(j)} * Partner_{j(i)}\right) + \delta^N\left(T_{s(j)} * NoRole_{j(i)}\right) + \alpha^T Target_{j(i)} + \alpha^P Partner_{j(i)} + \gamma\left(T_{s(j)} * Target_{j(i)} * C_{j(i)}\right) + \theta\left(Target_{j(i)} * C_{j(i)}\right) + X_i\beta + \pi_{b(s)} + u_{ijkt}.$$

Specification 4 is identical to specification 3 except that we have added the two terms with $C_{j(i)}$, which is the proportion of skills matched. Table 4 reports the estimate of γ. We also control for θ, the "main effect" of the proportion of skilled matched; recall we observe $C_{j(i)}$ for control target teachers since pair assignments were determined algorithmically before random assignment. In the other columns

[41] Recall that proposed pairings were determined algorithmically (for both treatment and control school teachers) based on matches in 19 specific skill areas measured in each teacher's prior evaluation microdata. We count up the number of skill areas in which there is a match: the target teacher has a score less than 3 and the recommended partner has a score of 4 or greater (see section 1). Then, we divide the number of matches by the number of areas in which the target teacher scored less than 3. This "proportion skills matched" measure is based on the one-to-one pairings recommended in the original principal reports in the spirit, again, of intent-to-treat estimates.

of Table 4 we examine other characteristics, $C_{j(i)}$, of the partnership using specification 4. We have standardized the "proportion skills matched" (mean 0, s.d. 1) for comparison with other characteristics in Table 4.

[Insert Table 4 About Here]

Consistent with the peer learning prediction (in the previous paragraph), the interaction coefficient on the "proportion skills matched" is positive ($\hat{\gamma} = 0.056\sigma$). The treatment effect is larger in pairs where the high-performing teacher is better suited to teach new skills to her low-performing partner. This estimate is marginally statistically significant: the FRT *p*-value is 0.125 and the WCBt *p*-value is 0.052. As shown below, this "proportion skills matched" coefficient is more precisely estimated when we include additional characteristics of the target and partner teachers as controls.

An alternative specification of skill match is a simple indicator = 1 if the proportion of skills matched is above the sample median (the median is about half of skills matched) which we report in column 2. The treatment effects appear to be concentrated among target teachers who were better matched to partners with relevant skills to share. When the match is above median the estimated effect is 0.19 (= 0.036 + 0.156). When the match is below median the effect is positive but not statistically significant (0.036σ, FRT *p*-value = 0.45, WCBt *p*-value = 0.34, see column 2 row 1), and similar to the point estimates for partner teachers and teachers with no role. This pattern is most consistent with the skill development mechanism. If the primary mechanism was a change in effort or a change in joint production behavior, we would have expected similar effects for both above and below median matches.

Moreover, the result for "proportion skills matched" is not driven by target teachers who simply have more weak skills to match on. In column 3 we replace

"proportion skills matched" with "number of skills attempted to match," and in column 4 we include both simultaneously. The target teacher's number of weak skills does not itself predict a larger treatment effect; the effect is larger only when weak skills are matched. By contrast, if extra attention to low skill scores generated a change in teacher effort, we would have expected a positive coefficient on "number of skills attempted to match" but the point estimate is negative.

Focusing on the "proportion skills matched" measure implies a specific form of matching complementarities in peer effects. A simpler alternative explanation is that low-performing teachers benefit from working with high-performing peers, where low- and high-performing are broad characteristics; and that the "proportion skills matched" is only correlated with these more general pair characteristics. In the remaining columns of Table 4 we test this general peer effects explanation against the more-specific skill matching explanation. In column 5 we allow the treatment effect for target teachers to vary with the target teacher's own prior performance and with her partner teacher's prior performance. We measure broad prior performance with "value-added"—a teacher's contribution to student test score growth.[42] In column 6 we test for broad-based complementarities by adding the interaction between target prior performance and partner prior performance. In both specifications we find little evidence of a relationship between general performance levels and the treatment effect, at least little evidence given our statistical power. If anything the point estimates suggest the treatment effect, for the average target teacher, may be decreasing in the performance of the matched partner teacher, with that negative relationship shrinking as the target's own prior performance rises.

[42] We use value added scores provided by the state's TVAAS evaluation system (Tennessee Value-Added Assessment System). We calculate an overall value added score for each teacher by averaging all her subject-by-year value added scores from 2011-12 and 2012-13.

Columns 7 and 8 provide a head-to-head comparison of the more-specific skill matching explanation and more-general broad peer effects explanation. The empirical results are most consistent with complementarities exploited by matching on skills specifically, not simply exposure to a peer who is "good," broadly defined. Still, our limited precision is important to remember. The head-to-head comparison is important for statistical inference reasons as well. The point estimates for the relationships between broad "value added" and pair effects are not small substantively, and comparable in magnitude to the "proportion skills matched" point estimates. Still, the latter are much more precisely estimated. Indeed, the "proportion skills matched" coefficient is more precisely estimated when we include the broad "value added" measures of prior performance.

Evidence for skill growth is also apparent when we examine other aspects of teacher performance. To this point we have focused on performance as measured by student test score growth. Table 5 reports treatment effects for a second measure of teacher performance: evaluation scores from direct, in-class observations of teaching practices. These are observation scores from the teachers' formal evaluations described in section 1. In column 1 the outcome is teacher *j*'s average score across all 19 skills and all classroom observations.[43] All outcome variables are standardized (mean 0, s.d. 1). As measured in direct observations, target teachers' skills improve on average by 0.05 teacher standard deviations. The coefficients for partner teachers are similar in magnitude, but estimates for both partner and no-role teachers have much wider (implicit) confidence intervals.

Consistent with the skill growth mechanism, notably, target teachers' improvements are concentrated in skills where there was a match—a match

[43] As described in section 1, teachers are scored on 19 skills multiple times per year. We first calculate an average score for each skill, standardize the skill scores, and then average the skill scores to obtain the overall average. We use all post-randomization scores from 2013-14 and 2014-15. We regress each outcome on a treatment indicator, randomization block fixed effects, and controls for teacher experience to improve precision.

between the target's weak skills and her partner's strong skills. We examine treatment effects separately (column 2) for the subset of skills where was a match—the target scored less than 3 and her partner scored 4 or higher—and (column 3) the subset of skills where the target was weak—she scored less than 3—but there was no match. We also report (column 4) the effect for skill areas where we did not attempt to match because the target scored 3 or higher.[44] In matched skills, treatment target teachers improved 0.27 teacher standard deviations more than control target teachers did in their would-have-been matched skills. Yet treatment target teachers improved much less, if at all, in weak skill areas where there was no match with a partner strength. This comparison suggests learning from peers. By contrast, imagine that the results in Table 5 had showed increases in both matched and unmatched skills. That pattern could be seen as consistent with a change in effort mechanism, where target teachers work harder to improve their low observation scores to simply avoid being target teachers again next year.

[Insert Table 5 About Here]

One important note of caution in interpreting these classroom observation results in Table 5. Observations are conducted by the school (assistant-)principals who were aware of the experiment. Principals may have, consciously or unconsciously, inflated observations scores to recognize participation in the program; or, alternatively, been more critical as a result of the program.

The final two results, reported in Appendix Table A3, test the hypothesis that treatment improved performance by creating new opportunities to share productive resources or job tasks. Under this hypothesis we would expect the treatment effect

[44] For any individual teacher these three categories are mutually exclusive. Each of the 19 skills must belong to one and only one of matched, not matched, or not attempted to match.

to be greater when teacher pairs teach the same grade-level or subject area.[45] In this case $C_{j(i)}$ in specification 4 is an indicator $= 1$ if the target and partner teachers are both teaching the subject k. We also test an indicator $C_{j(i)} = 1$ if both teach the same grade level(s). In both cases the point estimate is positive, which would be consistent with shared production or resources. But in both cases the estimates are small relative to the overall impact and the (implicit) confidence intervals are quite wide.

IV. Discussion and Conclusion

In this paper, we document improvements in teachers' job performance resulting from an intervention designed to encourage learning from coworkers, in particular learning skills from coworkers with complementary skills. The relatively low-performing teachers targeted by our intervention—and ultimately their students—benefited substantially from partnering with a higher-performing colleague at their school. Target teachers' performance improved 0.12σ in the year of treatment, and the improvement appears to have persisted into at least the next year. These performance improvements were larger when teacher partnerships were better matched on strong and weak skills specifically, not simply when "good" and "bad" teachers were paired.

The skill-pairing intervention we study is different from most formal "mentoring" efforts in teaching or other occupations. Mentoring is not uncommon in schools. The typical mentoring relationship is between (i) an experienced teacher who is high-performing broadly defined and (ii) a novice (newly hired) teacher (Kraft, Blazar, and Hogan 2018). The mentor is sometimes a coworker, but often a former

[45] The assumption motivating this prediction is that, even absent the treatment, shared production activities are easier or higher-return when teachers teach the same grade-level or subject.

classroom teacher specializing in and trained in mentoring novices. Similar mentoring of new hires (novices) occurs in other occupations as well.[46]

Our intervention is different in two key ways. First, it differs on who participates. The "target" teachers in our pairings are not exclusively novices (new hires). The "partner" teachers are not specialized mentors, nor are they necessarily the highest performing teachers overall. The marginal potential partner identified by the algorithm is at about the 40th percentile of the given school's overall performance distribution. Similarly, the marginal target teacher is at about the 30th percentile. Thus, the matching algorithm might pair two "average" teachers, where average is defined by overall performance. Second, pairs are matched based on specific skills. As discussed in section 3, this particular feature seems central to the success of the intervention. Matching on specific skills is a way to include overall "average" teachers who could benefit from help in (help coach others on) specific skill areas. Similarly, "average" teachers who are good at specific skills can help coworkers in a more direct way without being specialized mentors.

The current experiment and results suggest practical alternatives to or complements to existing employee training or mentoring programs. Perhaps all teachers with some skill deficiency, novice or veteran, would benefit from a specialized, trained, formal mentor. Perhaps the coworker pairs in our experiment are a less-effective alternative to the formal approach. We certainly did not test any comparison to formal mentoring in the current experiment. However, it seems plausible that a coworker partnership focused on specific skills would be less costly than hiring more specialized mentors; we turn to the question of costs next. Our pairing design draws more teachers into the "mentor" role in a less-formal, perhaps lower-stakes, way. Our sense is that, at least in public schools, this resource of

[46] Such formal mentoring programs are more common in empirical literature, partly because formal features are easier to record in data (for a review of evidence in teaching see Jackson, Rockoff, and Staiger 2014 and Kraft, Blazar, and Hogan 2018; for other occupations Frazis and Loewenstein 2006).

coworker skills is underutilized, and that is likely true in some other occupations. Last, there are many informal mentoring relationships between coworkers today, formed by employees themselves or by supervisors. In that sense pairing coworkers is nothing new. However, the data-informed skill-specific matching approach may identify beneficial pairings which would not have otherwise been found by coworkers or supervisors.

We now turn to costs. The first-order costs of the intervention are the opportunity costs of participating teachers' time. Potential forgone uses of teachers' time and effort include (i) attention to their current students (or other job responsibilities), (ii) other investments in developing new or existing skills, or (iii) teachers' own leisure. By contrast, other costs are small. The matching algorithm is a computer program—a small fixed cost. The school principal introduced each pair, and likely provided some encouragement over the course of the school year, but these per-pair costs are likely small relative to the paired teachers' own time and effort. The marginal cost per target teacher (or per teacher pair) is unlikely to rise or fall at a larger scale.

One specific and important potential opportunity cost is that the higher-skilled "partner" teacher might give less effort to teaching her own students. In our estimates, students of partner teachers were no worse off, if anything they may be benefited slightly.

Given the opportunity costs, for a scaled-up ongoing version of the partnership program administrators would likely need to be more explicit about what other existing job responsibilities teachers and principals could give up.[47] One intuitive potential place to cut back is on teachers' time in formal training courses, often called "professional development" or "PD" in schools. Today "PD" is the primary

[47] It is possible that the teachers and principals participating in the experiment were willing to pay the opportunity costs but only for the experiment year to support the research project. If the program were added on top of existing responsibilities teachers, especially partner teachers, might want to be compensated explicitly.

approach to formal in-service training for public school teachers. Collectively K-12 schools spend about $18 billion per year on professional development courses, of which $3 billion is paid to external providers (Gates Foundation 2014); the average teacher spends at least 20 hours each year in "PD."[48] Despite the substantial commitment of resources, the empirical evidence suggests little effect on teacher performance (see Jackson, Rockoff, and Staiger 2014 for a review). Similarly, public school systems spend tremendous resources paying for teachers' graduate tuition, and paying higher salaries once teachers' obtain their graduate degree. There is limited evidence that such degrees significantly improve teacher effectiveness (Jackson, Rockoff, and Staiger 2014).

Our estimates are consistent with prior evidence of learning from coworkers among teachers. Two prior quasi-experimental papers are closest to our present paper. Jackson and Bruegmann (2009) find that when a teacher begins working with higher-performing colleagues her own performance improves as a result. A one standard deviation increase in peer performance, as measured by prior value added, generates a 0.03-0.04σ improvement in own performance. Importantly, in the Jackson and Bruegmann (2009) case "working with" is defined as teaching in the same grade and school as the peer; the peer interactions we study are more direct and the effects are larger. Taylor and Tyler (2012) find that a teacher's performance improves 0.05σ during a school year in which the teacher is evaluated by a "peer" based on classroom observations. In the Taylor and Tyler (2012) case the "peer" evaluator is from a different school, and has left teaching herself to specialize in evaluation. Both Jackson and Bruegmann (2009) and Taylor and Tyler (2012) report that the gains in performance are sustained into the future.

Finally, two notes on the context and design of the experiment which are important to interpreting the results. First, statistical inference is limited by the

[48] Author's calculation from the Schools and Staffing Survey 2011-12.

small number of schools, 14, which were randomized to treatment and control. We do report inference statistics suited to the small number of clusters, or at least better suited than conventional cluster-robust standard errors. And we do find several key results are significant at conventional levels, partly because precision is aided by controlling for lagged dependent variables. Still, in several places our results remain relatively imprecise, preventing us from rejecting on even seemingly large point estimates.

Second, the experiment was conducted in the context of a formal teacher evaluation system. All teachers in our study—treatment and control, target and partner and no role—are subject to Tennessee's formal performance evaluation system. Principals are evaluators. Teacher pairs were identified based on prior evaluation results, and teacher pairs were encouraged, in part, to work on improving evaluation results. These connections to formal evaluation, and its incentives, likely influenced principals' and teachers' willingness to participate at the extensive margin, and the effort they gave on the intensive margin. To be sure, estimated treatment effects in any setting are a function of the empirical intensive margin of effort by participants. More specifically, our results may not generalize to settings where the pairing skills are not themselves incentivized by formal evaluation. Still, many public school systems now have evaluation programs similar to Tennessee's (Steinberg and Donaldson 2016).

The teacher job performance improvements documented in this paper suggest learning from colleagues is at least as valuable as formal training ("professional development") or the gains from experience in developing teaching skills. Most practically, the treatment and results suggest promising ideas for managing peer learning in the sizable teacher workforce.

REFERENCES

Angrist, Joshua and Victor Lavy. 2001. “Does Teacher Training Affect Pupil Learning? Evidence from Matched Comparisons in Jerusalem Public Schools.” *Journal of Labor Economics* 19 (2): 343–369.

Azoulay, Pierre, Joshua S. Graff Zivin, and Jialan Wang. 2010. “Superstar Extinction.” *Quarterly Journal of Economics* 125 (2): 549-589.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2005. “Social Preferences and the Response to Incentives: Evidence from Personnel Data.” *Quarterly Journal of Economics* 120 (3): 917-962.

Battu, Harminder, Clive R. Belfield, and Peter J. Sloane. 2003. “Human Capital Spillovers within the Workplace: Evidence for Great Britain.” *Oxford Bulletin of Economics and Statistics* 65 (5): 575-594.

Bergman, Peter, and Matthew Hill. 2018. “The Effects of Making Performance Information Public: Evidence from Los Angeles Teachers and a Regression Discontinuity Design.” *Economics of Education Review* 66: 104-113.

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts. 2013. “Does Management Matter? Evidence from India.” *Quarterly Journal of Economics* 128 (1): 1-51.

Bloom, Nicholas, James Liang, John Roberts, and Zhichun Jenny Ying. 2015. “Does Working from Home Work? Evidence from a Chinese Experiment.” *Quarterly Journal of Economics* 130 (1): 165-218.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. “Bootstrap-Based Improvements for Inference with Clustered Errors.” *Review of Economics and Statistics* 90 (3): 414-427.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2011. “Robust Inference with Multiway Clustering.” *Journal of Business and Economic Statistics* 29 (2): 238-249.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. “Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood.” *American Economic Review* 104 (9): 2633-2679.

Chung, EunYi. and Joseph P. Romano. 2013. “Exact and Asymptotically Robust Permutation Tests.” *The Annals of Statistics* 41 (2): 484–507.

Cortes, Kalena, Joshua Goodman, Takako Nomi. 2015. “Intensive Math Instruction and Educational Attainment: Long-Run Impacts of Double-Dose Algebra.” *Journal of Human Resources* 50 (1): 108-158.

Danielson, Charlotte. 1996. *Enhancing Professional Practice: A Framework for Teaching.* Alexandria, VA: ASCD.

Dee, Thomas S., and James Wyckoff. 2015. “Incentives, Selection, and Teacher Performance: Evidence from IMPACT.” *Journal of Policy Analysis and Management* 34 (2): 267-297.

Ding, Peng. 2017. “A Paradox from Randomization-Based Causal Inference.” *Statistical Science* 32 (3): 331-345.

Fisher, Ronald. 1935. *Design of Experiments.* Edinburgh: Oliver and Boyd.

Frazis, Harley, and Mark A. Loewenstein. 2007. “On-the-Job-Training.” *Foundations and Trends in Microeconomics* 2 (5): 363-440.

Gates Foundation. 2014. *Teachers Know Best: Teachers Views on Professional Development.* Seattle: Bill & Melinda Gates Foundation.

Hamilton, Barton H., Jack A. Nickerson, and Hideo Owan. 2003. “Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation.” *Journal of Political Economy* 111 (3): 465-497.

Hanushek, Eric A., and Steven G. Rivkin. 2010. “Generalizations about Using Value-Added Measures of Teacher Quality.” *American Economic Review* 100 (2): 267-271.

Harris, Douglas N., and Tim R. Sass. 2011. “Teacher Training, Teacher Quality, and Student Achievement.” *Journal of Public Economics* 95: 798-812.

Ho, Andrew D., and Thomas J. Kane. 2013. *The Reliability of Classroom Observations by School Personnel.* Seattle: Bill & Melinda Gates Foundation.

Horowitz, Joel L., and Charles F. Manski. 2000. "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data." *Journal of the American Statistical Association* 95 (449): 77–84.

Ichino, Andrea, and Giovanni Maggi. 2000. "Work Environment and Individual Background: Explaining Regional Shirking Differentials in a Large Italian Firm." *Quarterly Journal of Economics* 115 (3): 1057-1090.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences.* New York: Cambridge University Press.

Jackson, C. Kirabo, and Elias Bruegmann. 2009. "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers." *American Economic Journal: Applied Economics* 1 (4): 85-108.

Jackson, C. Kirabo, Jonah E. Rockoff, and Douglas O. Staiger. 2014. "Teacher Effects and Teacher-Related Policies." *Annual Review of Economics* 6 (1): 801-825.

Kane, Thomas J., and Douglas O. Staiger. 2012. *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.* Seattle, WA: Bill & Melinda Gates Foundation.

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 (3): 587-613.

Kraft, Matthew. A., David Blazar, and Dylan Hogan. 2018. "The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence." *Review of Educational Research* 88 (4): 547-588.

Kraft, Matthew A., and Allison F. Gilmour. 2017. "Revisiting the Widget Effect: Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness." *Educational Researcher* 46 (5): 234-249.

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

Kuhn, Harold W. 1955. “The Hungarian Method for the Assignment Problem.” *Naval Research Logistics Quarterly* 2: 83–97.

Lee, David S. 2009. “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects.” *Review of Economic Studies* 76: 1071-1102.

Marshall, Alfred. 1890. *Principles of Economics.* London: Macmillan.

Mas, Alexandre, and Enrico Moretti. 2009. “Peers at Work.” *American Economic Review* 99 (1): 112-145.

Moretti, Enrico. 2004. “Workers' Education, Spillovers, and Productivity: Evidence from Plant-Level Production Functions.” *American Economic Review* 94 (3): 656-690.

New York Times. 2013, March 30. “Curious Grade for Teachers: Nearly All Pass.”

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.

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

Rockoff, Jonah E. 2008. “Does Mentoring Reduce Turnover and Improve Skills of New Employees? Evidence from Teachers in New York City.” National Bureau of Economic Research Working Paper 13868.

Rosenbaum, Paul R. 2002. *Observational Studies.* New York: Springer.

Sacerdote, Bruce I. 2011. “Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?” In *Handbook of Economics of Education Volume 3*, edited by Hanushek, Eric A., Stephen Machin, and Ludger Woessmann. Amsterdam: North Holland.

Steinberg, Matthew P., and Morgaen L. Donaldson. 2016. “The New Educational

Accountability: Understanding the Landscape of Teacher Evaluation in the Post-NCLB Era." *Education Finance and Policy* 11 (3): 340-359.

Steinberg, Matthew P., and Lauren Sartain. 2015. "Does Teacher Evaluation Improve School Performance? Experimental Evidence from Chicago's Excellence in Teaching Project." *Education Finance and Policy* 10 (4): 535-572.

Taylor, Eric S. 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.

Taylor, Eric S., and John H. Tyler. 2012. "The Effect of Evaluation on Teacher Performance." *American Economic Review* 102 (7): 3628-3651.

U.S. Census Bureau. 2015. *Public Education Finances: 2013.* Educational Finance Branch G13-ASPEF. Washington, D.C.: United States Census Bureau.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling. 2009. *The Widget Effect: Our National Failure to Acknowledge and Act on Teacher Effectiveness.* New York: The New Teacher Project.

TABLE 1— CLASSROOM OBSERVATION EVALUATION SCORES, YEAR PRIOR TO EXPERIMENT

(Panel A) Scores in 19 skill areas, and scores averaging across skill areas

			Proportion scoring	
	Mean	Standard Deviation	< 3	≥ 4
	(1)	(2)	(3)	(4)
Overall average	3.66	(0.68)	0.18	0.34
Instruction average	3.56	(0.69)	0.21	0.31
Problem Solving	3.19	(0.78)	0.23	0.22
Thinking	3.34	(0.75)	0.16	0.30
Questioning	3.39	(0.74)	0.17	0.34
Grouping Students	3.43	(0.79)	0.16	0.35
Lesson Structure and Pacing	3.43	(0.89)	0.22	0.37
Academic Feedback	3.49	(0.79)	0.14	0.39
Activities and Materials	3.56	(0.80)	0.13	0.42
Standards and Objectives	3.59	(0.81)	0.15	0.48
Presenting Instructional Content	3.64	(0.86)	0.15	0.48
Teacher Knowledge of Students	3.75	(0.88)	0.13	0.53
Motivating Students	3.85	(0.81)	0.08	0.57
Teacher Content Knowledge	4.12	(0.78)	0.05	0.70
Planning average	3.62	(0.80)	0.17	0.43
Assessment	3.38	(0.91)	0.18	0.45
Student Work	3.66	(0.89)	0.10	0.55
Instructional Plans	3.83	(0.88)	0.09	0.66
Environment average	3.97	(0.82)	0.11	0.57
Expectations	3.81	(0.92)	0.10	0.60
Managing Student Behavior	3.94	(0.96)	0.09	0.68
Respectful Culture	4.06	(0.89)	0.07	0.72
Environment	4.07	(0.92)	0.06	0.70

(B) Teachers with skill scores < 3 or ≥ 4

	Proportion	Mean
	(5)	(6)
One or more skill areas < 3	0.41	
Number of skills with score < 3		2.37
Number of skills with score < 3 | at least one < 3		5.84
One or more skill areas ≥ 4	0.89	
Number of skills with score ≥ 4		9.09
Number of skills with score ≥ 4 | at least one ≥ 4		10.17

Notes: Authors' calculations using 507 teacher observations from 2012-13 for treatment and control schools. Data for Tennessee as a whole are similar. Observation scores in natural units (1-5 scale).

TABLE 2— STUDENT AND TEACHER CHARACTERISTICS, AND PRE-TREATMENT BALANCE

	Mean (standard deviation)		Difference = 0 *p*-value	
	Control	Treatment	WCBt	FRT
	(1)	(2)	(3)	(4)
Student characteristics				
Baseline test scores				
Mathematics	0.078	0.036	0.280	0.313
	(0.531)	(0.658)		
Reading/language arts	0.065	0.038	0.516	0.609
	(0.562)	(0.668)		
Female	0.490	0.488	0.952	0.938
Race/ethnicity				
White	0.333	0.300	0.728	0.781
African-American	0.610	0.594	0.828	0.875
Latino(a)	0.047	0.087	0.176	0.422
Other	0.010	0.018	0.160	0.250
English language learner	0.015	0.038	0.004	0.250
Special education	0.118	0.126	0.704	0.609
Retained in grade	0.001	0.002	0.300	0.500
Teacher characteristics				
Years of experience	12.151	13.800	0.284	0.313
	(11.92)	(10.18)		
Baseline value-added	-0.023	0.112	0.416	0.391
	(0.720)	(0.783)		
Baseline classroom observation score	3.694	3.869	0.472	0.438
	(0.580)	(0.550)		

Notes: Means and standard deviations net of randomization block fixed effects. Baseline student test scores and baseline teacher value-added standardized (mean 0 standard deviation 1) using the statewide distributions for students and teachers respectively. Observation scores in natural units (1-5 scale). The sample has 2,947 student and 141 teacher observations. Wild cluster (school) bootstrap-*t* *p*-values in column 3, and Fisher randomization test *p*-values in column 4. See text for details of the two approaches to inference.

TABLE 3— TREATMENT EFFECTS (INTENT TO TREAT)

	Dependent variable = student math and reading/ELA test scores in		
	Experiment year, t		$t+1$
	(1)	(2)	(3)
(A) Average treatment effect	0.051	0.065	
using school-level regression	[0.300]	[0.000]	
	(0.375)	(0.047)	
School mean $t-1$ score control		√	
School observations	14	14	
(B) Average treatment effect	0.055	0.056	0.106
using student-level regression	[0.224]	[0.080]	[0.220]
	(0.328)	(0.250)	(0.375)
Pre-experiment covariates		√	√
Student-by-subject observations	5,511	5,511	4,392
Teacher observations	136	136	96
(C) Treatment effect by teacher role			
Low-performing target teachers	0.082	0.123	0.252
	[0.024]	[0.000]	[0.068]
	(0.297)	(0.031)	(0.422)
High-performing partner teachers	0.039	0.029	0.056
	[0.532]	[0.252]	[0.684]
	(0.813)	(0.547)	(0.641)
No assigned role	0.013	0.029	0.013
	[0.824]	[0.468]	[0.908]
	(0.844)	(0.625)	(0.891)
Pre-experiment covariates		√	√
Student-by-subject observations	5,511	5,511	4,392
Teacher observations	136	136	96

Notes: Each column within panels reports estimates from a separate regression. Wild cluster (school) bootstrap-*t* *p*-values in brackets, and Fisher randomization test *p*-values in parentheses. See text for details of the two approaches to inference. All regressions include randomization block fixed effects. Throughout the table, the outcome variable is (based on) student test scores in math and reading/language arts, which have been standardized (mean 0, s.d. 1) within subject by grade-level cells using the statewide distribution.

Panel A columns 1-2: The regressions use school level observations (14 observations) as in specification 2. The dependent variable is the school's mean student test score at the end of the experiment year (year t). Column 2 includes the scalar control: mean student test score from year $t-1$. Note that with school level observations the wild cluster bootstrap-*t* reduces to the simpler wild bootstrap-*t*.

Panel B columns 1-2: The regressions use student-by-subject level observations as in specification 1. The dependent variable is a student test score at the end of the experiment year. Pre-experiment covariates includes: (i) Baseline student achievement: the average of each student's prior year math and reading/ELA scores. The prior test-score slope is allowed to vary by outcome subject and grade-level. (ii) Teacher pre-experiment value-added in the outcome subject: the average of 2012-13 and 2011-12 scores. (iii) Indicators for student gender, race/ethnicity, English language learner status, special education status, and whether the student is repeating the grade. When baseline test scores or value-added are missing we set the value to zero and include an indicator = 1 for missing. If the student had two or more teachers in a given subject, we include one observation per teacher and weight each observation by the proportion of responsibility allocated by the state to the teacher. Three quarters of students had one teacher in a given subject.

Panel C columns 1-2: Estimation details are identical to panel B, except that the single treatment indicator is replaced with a series of indicators formed by the interaction of treatment and teacher *j*'s assigned role. Specifically, the five indicators are: treatment * target teacher, treatment * partner teacher, treatment * no assigned role, target teacher, and partner teacher (with "no role" teacher the omitted category).

Column 3: Estimation details are identical to column 2, within panels, except as follows: The dependent variable is test scores at the end of 2014-15, the year following the experiment year in 2013-14. We hold the set of teachers from the experiment year fixed, and use observations on the new students they are assigned in 2014-15. Even if the teacher changes schools, the randomization blocks and treatment condition are based on her 2013-14 school.

TABLE 4— TREATMENT EFFECT HETEROGENEITY BY PAIR CHARACTERISTICS
DEP. VAR. = STUDENT MATH AND READING/ELA TEST SCORES END OF EXPERIMENT YEAR *T*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment main effects by role								
Low-performing target	0.133	0.036	0.121	0.128	0.094	0.111	0.135	0.130
	[0.000]	[0.340]	[0.000]	[0.000]	[0.100]	[0.104]	[0.000]	[0.124]
	(0.031)	(0.453)	(0.055)	(0.031)	(0.070)	(0.055)	(0.008)	(0.023)
Pair-characteristics interacted with target treatment								
Treatment * Low-performing target								
* proportion skills matched	0.056			0.066			0.075	0.056
	[0.052]			[0.020]			[0.028]	[0.208]
	(0.125)			(0.102)			(0.016)	(0.078)
* proportion skills matched above median (binary)		0.156						
		[0.000]						
		(0.063)						
* number of skills attempted to match			-0.016	-0.028				
			[0.632]	[0.344]				
			(0.680)	(0.766)				
* own prior value-added					-0.025	0.042	0.040	0.082
					[0.576]	[0.512]	[0.720]	[0.580]
					(0.563)	(0.430)	(0.328)	(0.203)
* partner prior value-added					-0.126	-0.154	-0.105	-0.184
					[0.312]	[0.360]	[0.224]	[0.316]
					(0.711)	(0.539)	(0.711)	(0.586)
* own prior value-added * partner prior value-added						0.032		-0.067
						[0.772]		[0.728]
						(0.422)		(0.523)

Notes: Each column reports estimates from a separate regression. The sample is 5,511 student-by-subject observations and 136 teachers. The details of estimation are identical to Table 3 panel C column 2, except that we add various characteristics of the pair as additional right-hand-side variables, and interact those characteristics with the treatment indicator for target teachers. The interaction coefficients are shown above. In each case the regression also includes a "main effect" for characteristic shown above, specifically the interaction of the characteristic and an indicator for target teacher main effect. A "skill match" occurs when the target teacher has a score below 3 in the skill and the assigned partner has a score of 4 or higher (19 skills possible). The denominator in "proportion skills matched" is the number of skills where the target has a score below 3. The "proportion skills matched", "number of skilled attempted to match", and prior value-added variables have been standardized (mean 0, s.d. 1) within the sample. All regressions do include treatment effects for mentor and no role teachers, as in Table 3 panel C, but those coefficients are suppressed above to simplify the table. The point estimates and *p*-values do not change meaningfully from the estimates in Table 3 panel C for any of the different specifications reported in this table; those estimates are available from the authors. Wild cluster (school) bootstrap-*t* *p*-values in brackets, and Fisher randomization test *p*-values in parentheses. See text for details of the two approaches to inference.

TABLE 5— TREATMENT EFFECT ON TEACHING SKILLS SCORED IN CLASSROOM OBSERVATIONS

	Dependent variable = Average of evaluation scores on…			
	All 19 skills	Skills matched	Skills *not* matched	Skills where no attempt was made to match
	(1)	(2)	(3)	(4)
Low-performing target teachers	0.050	0.267	0.034	-0.071
	[0.060]	[0.000]	[0.500]	[0.348]
	(0.516)	(0.016)	(0.516)	(0.984)
High-performing partner teachers	0.048			
	[0.836]			
	(0.891)			
No assigned role	-0.062			
	[0.892]			
	(0.797)			

Notes: Each cell reports an estimate from a separate regression. The sample is 136 teachers. The dependent variable is a measure of observed teaching practices from formal classroom observations conducted as part of the teacher's performance evaluation (see text for more details). In column 1 the outcome is the teacher's average score across all 19 skills. In columns 2 through 4 the outcome is the target teacher's average score across a specific subset of skills described in the column headers. The skills which contribute to each subset differ from teacher to teacher. But for any given teacher the three columns are mutually exclusive and exhaustive. Evaluation scores contributing to the dependent variables are from the 2013-14 and 2014-15 school years. Dependent variables are standardized (mean 0, s.d. 1), and each of the 19 skill scores was standardized before taking averages. All regressions include randomization block fixed effects, and controls for teacher experience (indicators for quartiles of teacher experience). Wild cluster (school) bootstrap-*t* *p*-values in brackets, and Fisher randomization test *p*-values in parentheses. See text for details of the two approaches to inference.