Online Appendix

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

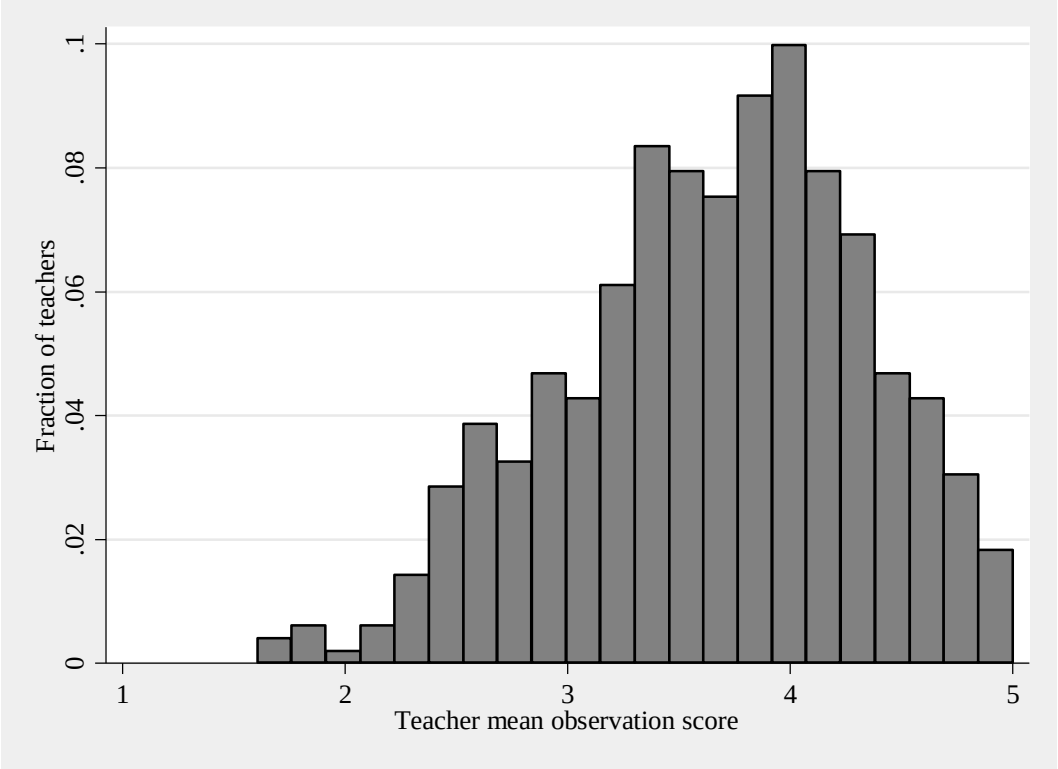
John P. Papay, Brown University
Eric S. Taylor, Harvard University
John H. Tyler, Brown University and NBER
Mary E. Laski, Harvard University

April 2019

Appendix A: Additional Figures and Tables

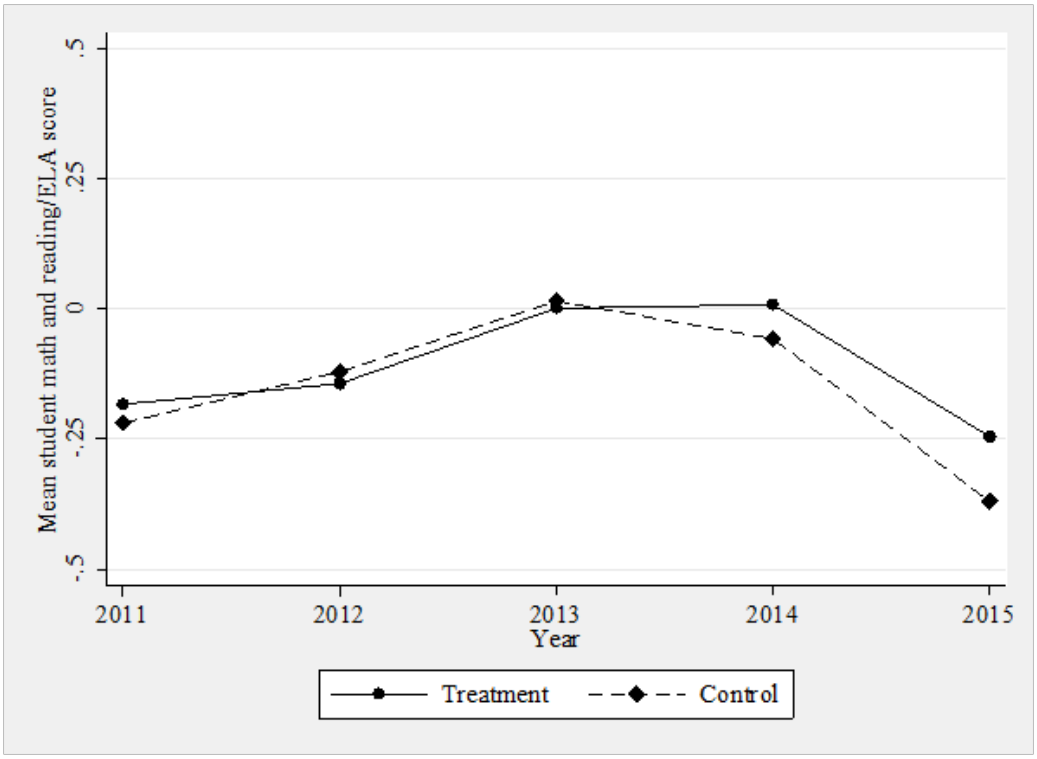
Questioning		
Significantly Above Expectations (5)	**At Expectations (3)**	**Significantly Below Expectations (1)**
• Teacher questions are varied and high-quality, providing a balanced mix of question types: o knowledge and comprehension; o application and analysis; and o creation and evaluation. • Questions require students to regularly cite evidence throughout lesson. • Questions are consistently purposeful and coherent. • A high frequency of questions is asked. • Questions are consistently sequenced with attention to the instructional goals. • Questions regularly require active responses (e.g., whole class signaling, choral responses, written and shared responses, or group and individual answers). • Wait time (3-5 seconds) is consistently provided. • The teacher calls on volunteers and non-volunteers, and a balance of students based on ability and sex. • Students generate questions that lead to further inquiry and self-directed learning. • Questions regularly assess and advance student understanding. • When text is involved, majority of questions are text based.	• Teacher questions are varied and high-quality providing for some, but not all, question types: o knowledge and comprehension; o application and analysis; and o creation and evaluation. • Questions usually require students to cite evidence. • Questions are usually purposeful and coherent. • A moderate frequency of questions asked. • Questions are sometimes sequenced with attention to the instructional goals. • Questions sometimes require active responses (e.g., whole class signaling, choral responses, or group and individual answers). • Wait time is sometimes provided. • The teacher calls on volunteers and non-volunteers, and a balance of students based on ability and sex. • When text is involved, majority of questions are text based	• Teacher questions are inconsistent in quality and include few question types: o knowledge and comprehension; o application and analysis; and o creation and evaluation. • Questions are random and lack coherence. • A low frequency of questions is asked. • Questions are rarely sequenced with attention to the instructional goals. • Questions rarely require active responses (e.g., whole class signaling, choral responses, or group and individual answers). • Wait time is inconsistently provided. • The teacher mostly calls on volunteers and high-ability students.

Appendix Figure A1—Example from TEAM rubric, “Questioning” skills



Appendix Figure A2—Histogram of teacher mean observation scores

Note: Teacher observations. Mean observation score is the teacher's average of 19 skill scores.



Appendix Figure A3—Test scores before, during, and after the treatment year

Note: Each circle is the mean of standardized math and reading/language arts test scores in treatment schools for the given school year (2014 = 2013-14). Each diamond is the same mean for control schools. The treatment year is 2014. Test scores are standardized (mean 0, s.d. 1) using the statewide mean and standard deviation, and are net of randomization block fixed effects. Because the standardization is relative to the state distribution, the overall trends are changes relative to other schools in the state, not necessarily changes in absolute levels of performance.

Appendix Table A1—Pre-treatment balance by teacher role

	Treat. - cont. difference by teacher's assigned role		
	Target	Partner	No role
	(1)	(2)	(3)
Teacher characteristics			
Years of experience	-0.073	1.497	2.462
	[0.892]	[0.552]	[0.532]
	(0.984)	(0.703)	(0.500)
Baseline job performance			
Value-added	-0.355	-0.025	0.472
	[0.248]	[0.976]	[0.068]
	(0.484)	(0.922)	(0.078)
Classroom observation score	0.189	0.217	-0.294
	[0.172]	[0.384]	[0.304]
	(0.328)	(0.438)	(0.391)

Note: Each cell reports a treatment minus control difference in means. The sample is 141 teachers. The three estimates in each row come from a single regression. The dependent variable described by the row label. All regressions include randomization block fixed effects, and main effects for teacher role (i.e., "target" and "partner" with "no role" the omitted category). 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.

Appendix Table A2—Teacher participation

	Dep. var. = 1 if *participated* as a... Target (1)	Partner (2)
Treatment * *assigned* role:		
low-performing target	0.608	0.154
	[0.000]	[0.344]
	(0.078)	(0.422)
high-performing partner	0.070	0.366
	[0.140]	[0.000]
	(0.453)	(0.047)
no assignment	0.073	0.014
	[0.176]	[0.744]
	(0.313)	(0.719)
F-statistic excluded instruments jointly zero	18.5	31.5

Note: Each column reports estimates from a separate LPM regression; specifically, the first stage regressions from 2SLS estimation where actual role is instrumented with assigned role. Estimation is identical to Table 3 panel C column 1, except that the dependent variables are indicators = 1 if we observe participation in the target or partner roles respectively. The sample is 5,511 student-by-subject observations and 136 teachers. 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.

Appendix Table A3—Additional pair characteristics
dep. var. = student math and reading/ELA test scores

	(1)	(2)	(3)	(4)	(5)
Treatment main effects by role					
Low-performing target	0.112	0.113	0.126	0.136	0.095
	[0.024]	[0.000]	[0.344]	[0.000]	[0.060]
	(0.016)	(0.125)	(0.047)	(0.156)	(0.016)
Pair-characteristics interacted with target treatment					
Treatment * Low-performing target					
* teaching the same subject (binary)	0.036				
	[0.612]				
	(0.563)				
* teaching the same grade (binary)		0.045			
		[0.308]			
		(0.656)			
* years of experience			-0.009		
			[0.728]		
			(0.391)		
* years of experience ^ 2			0.000		
			[0.428]		
			(0.219)		
* fewer than 10 years of experience (binary)				-0.029	
				[0.692]	
				(0.656)	
* partner's years of experience					0.003
					[0.340]
					(0.891)

Note: 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 4, but with different target and pair characteristics. In practice teachers often teach more than one grade level, especially in middle. We calculate the student-weighted average of grade level for each teacher. The indicator is =1 if the difference in that average grade level is less than one. 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.

Appendix B: Treatment Details

B.1 Experimental procedures and treatment

This section provides a detailed description of the experimental procedures and treatment.

Recruitment of schools. In the summer of 2013, the research team met with the district's school principals as a group; described the "Evaluation Partnership Program," as the treatment was known; and solicited volunteers for the experiment the following school year. Of the district's 21 elementary and middle schools, 14 volunteered.

Making the list of one-to-one pairings. Matching teachers. In October 2013, prior to random assignment, the research team created at list of one-to-one teacher pairings or partnerships for each of the 14 participating schools. The matching algorithm which created those pairings is described in detail in section B.2. The inputs to the matching algorithm, also described in detail in B.2, are teachers' prior classroom observation scores on 19 specific teaching skills.

Across all 14 schools, there were 141 were teachers included in the matching process. Our analysis in this paper focuses on these 141 teachers for whom we (expected to) have student test score outcomes because they were teaching grades 4-8 math or reading/language arts. The algorithm initially identified 30.2 percent of teachers as a low-performing "target" teachers. Of those target teachers, 87.1 percent were matched by the algorithm with a high-performing "partner" teacher. Thus, just over one-quarter of teachers (26.4 percent) were target teachers paired with a partner, and so, because the matching is one-to-one, the same fraction of teachers were partner teachers paired with a target.

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 matching on student enrollment size.

Informing treatment principals. In late October 2013, the seven treatment school principals each received an Excel file listing the recommended one-to-one teacher pairings. An example is provided in Appendix Figure B1. The simple report has two rows for each low-performing "target" teacher. The first row shows the target teacher's name; and then, for each of the 19 skills, marks with an "o" the skills where the target teacher has "weak" performance, in her prior evaluations. The second row lists the recommended high-performing "partner," chosen by the matching algorithm described in section B.2. Then for the "partner" an "x" marks the skills where the partner is "strong." Specific definitions of target, partner, weak skill, strong skill, etc. are provided in section B.2.[1]

Accompanying the Excel pair report were additional materials (shown as figures in this appendix and described below) which were designed to aid the principal in carrying out her role described below.

Control principals did not have one-to-one matching reports. The research team did create a list of one-to-one teacher pairs for each of the control schools, identical to the lists for the treatment schools. However, the research team never created Excel reports, like Appendix Figure B1, for the control schools. Moreover, no one outside the authors of this paper had access to the control school pair lists (until well after the experiment year).

It is important to note that all school principals, both treatment and control, already had access to the data used to create these Excel reports. Indeed, school principals play a key role in creating the data as classroom observers themselves.

[1] These reports were prepared by the research team, but were sent to principals by the state of Tennessee's Department of Education (TDOE), Office of Research and Strategy. The research team never had access to teacher's names, or other personally identifying information. Once the Excel files listing the recommended one-to-one teacher pairings were created by the research team, TDOE replaced the masked ID numbers with actual teacher names.

A control school principal with some data skills and a little time (or a helpful friend) could create the "o" and "x" lists for each teacher. However, the matching algorithm, described in section B.2, was not revealed in detail to principals. Additionally, when asked, none of the seven control principals described creating such reports or attempting any new matching program during the experiment year.

The principal's role and responsibilities, in treatment schools. In the treatment—the Evaluation Partnership Program—the school principal played a critical but relatively small role. The principal's primary responsibility was to introduce teachers to the program and introduce the pairs to each other. First, each treatment principal was asked to meet with participating teachers individually and introduce the program. Second, after meeting with individuals, the principal was asked to meet with each pair of teachers and introduce them to each other as partners in the program.

Appendix Figure B2 is a list of suggested "talking points" for those meetings provided to principals. A brief description of the program provided to principals and teachers:

> **Evaluation Partnerships.** The Evaluation Partnership program is designed to help teachers use the information and feedback they receive in the teacher evaluation process. Currently, many teachers receive information about how they are doing but they may not obtain the necessary guidance and support to translate those evaluation scores into lasting changes in instructional practice. In the Evaluation Partnership program, teachers who struggle in a particular area of instructional practice will be paired with a partner who has demonstrated success in that area. We believe that, if done well, these partnerships will enable teachers to work together throughout the year to strengthen their instructional practices. We believe this program can provide clear benefits not only to the lower-performing teachers, as they receive guidance and advice, but also to the higher-performing teachers who will think about how to translate their expertise to help their peers.

Appendix Figure B3 is one-page program guide created for participating teachers. The guide includes a list of "Suggested activities" and a "Recommended

partnership timeline" along with some "Tips." The school principal provided these guides to teachers. We note, however, that as you can see the terms "target" and "partner" used in this paper were not used in communications with participating principals and teachers; "target" and "partner" are convenient short hand jargon.

The preceding paragraphs describe a principal's designed role in the treatment. In practice, one treatment school simply did not participate; the principal did not take any steps to start the partnerships. Another school did participate, but the program and partnerships were introduced in a group faculty meeting.

Each treatment principal was given a list of recommended one-to-one pairs to work from. But principals were told that they could make adjustments if they saw a need, for example, to avoid a pair which the principal knew from experience would not get along. To help principals choose an alternative partner, in case they decided a change was needed, for each target teacher we provided the report shown in Appendix Figure B4. It is structured quite similar to the report on recommended one-to-one pairs shown in Appendix Figure B1. This additional report, however, lists (up to) five possible alternative partners for each target teacher. Importantly, this list of five is not constrained by a one-to-one rule; a potential alternative partner may be listed for multiple target teachers. Our goal with this additional report was to provide principals easy access to information on skill-by-skill "weak"-to-"strong" comparisons for potential alternative partners, and thus encourage partnerships in the style of the program.

To be clear, all treatment effect estimates in the paper use only the teacher pairings as created by the matching algorithm described in section B.2. We do not use any pairings as adjusted by principals. Thus, the paper's results are, in that sense, interpreted as intent to treat pairings. First stage results for teacher participation are shown in Appendix Table A2.

The treatment principals' final role was to (hopefully) be generally supportive of the partnership program during the year. In some cases this included

concrete forms of support. For example, we know anecdotally that some principals provided time or arranged schedules so that teacher pairs could observe each other teaching; in other cases teachers made such arrangements on their own. However, we do not have systematic data on forms of principal support.

Teachers' and partnerships' role and responsibilities, in treatment schools. The core of the treatment design is the partnerships between teachers. As mentioned above, teacher pairs were introduced to each other by their school principal, and provided with the program guide in Appendix Figure B3. As included in those guides, pairs were encouraged to meet on a regular basis, with the first meeting and partner classroom observation occurring in the first month. The list of "Suggested Activities" includes reviewing the results of evaluations, observing each other in the classroom, asking for (and giving) constructive feedback and advice, developing strategies, and following up on each other's effort to improve, among other suggestions. The guide also includes a "Recommended…Timeline" and "Tips" for working with a partner.

The prior paragraph summarizes what teacher pairs were asked to do for the program. What can we say about what pairs actually did? The data we do have, though self-reported and incomplete, suggests observing each other teaching was a primary activity, along with discussing evaluation and providing feedback. First, we asked teachers about their activities in an end-of-year survey; the survey was anonymous and teachers self-reported their participation in the program. The survey response rate in treatment schools was 45 percent. Teachers who self-reported participating in the program also reported participating in peer observations more often: 80 percent said they observed a collogue, compared to 55 percent of non-participants in treatment schools; 67 percent said they were observed themselves, versus 51 percent. Self-reported participants were also a little more likely, 83 percent versus 77 percent, that they discussed their own evaluation results with other teachers. Second, we provided an online log for pairs to keep track of

their activities; the logs were not mandatory and some teachers reported not using the logs because their school's internet connection was bad. About half of pairs logged any activity. Though the logs are likely incomplete, of the activities logged 40 percent were observing each other's class and debriefing, while only 2 percent were discussing lesson plans. Most activities, 56 percent, were logged as meetings.

B.2 Details of the matching algorithm and our reasoning behind it

This section provides a detailed description of the algorithm we designed and used to identify, for each school, a set of one-to-one matches of a "target" teacher and "partner" teacher.

A critical feature is that the matches are one-to-one, i.e., each target teacher is in only one pair, and each partner is also in only one pair. To allocate the (potentially) scarce resource of good partners to targets, under this one-to-one constraint, we define a measure of "match quality" for each potential pairing of teachers; and then choose the set of pairings, for a given school, which maximizes the sum of the match quality scores. The match quality score, in brief, is the number of teaching skills, measured in prior classroom observation, where a target teacher's weakness in the skill is matched with a potential partner's strength.

The raw input data are classroom observation scores, provided by the Tennessee Department of Education. During a given observation, o, the teacher, j, is scored on (op to) 19 different teaching skills. The raw input data have one row per teacher per observation, jo, and columns with scores for each of the 19 skills. The skill scores are ordinal integer scores 1-5. The scores are labeled: (1) "Significantly below expectations," (2) "Below expectations," (3) "At expectations," (4) "Above expectations," and (5) "Significantly above expectations." We use data from observations, o, which occurred during the school year prior to the experiment, $t - 1$ = 2012-13.

In a pre-processing step, we create a dataset with just one row per teacher, j, and 19 skill scores. Each skill score is the simple average across observations, o, for teacher j. In our data, the average number of observations, o, per teacher, j, for a given skill is 3.6 (st.dev. 1.5, IQR 3-4). Table 1 lists the 19 skills and provides descriptive statistics for the skill scores using this dataset. In these data we also include the school where teacher j works in the experiment year, t = 2013-14. This dataset is the input to the matching algorithm.

The details and steps of the matching algorithm are below. To simplify exposition, the steps as written here describe the process for one school. This process was repeated for each of the 14 schools in the study sample, both treatment and control schools. There is no loss of detail by focusing on a given school; the algorithm does not make use of any information or data outside the input data for a given school.

1. Make a list of "target" teachers

   Let R_{js} be the score for teacher j in skill s, with $s \in \{1,2,\dots,19\}$. And let $W_{js} = 1\{R_{js} < 3\}$, an indicator $= 1$ if teacher j has a "weakness" in skill s. Teacher j is a target teacher if the following two conditions hold:

   a. $\left[\sum_{s=1}^{19} W_{js}\right] \geq 1$, teacher j has one or more "weak" skill areas

   b. $\left[\frac{1}{19}\sum_{s=1}^{19} R_{js}\right] \leq 3$, teacher j's average score across all 19 skills is "At expectations" or lower

   Let J^T be the number of target teachers.

2. Make a list of potential "partner" teachers

   Let $V_{js} = 1\{R_{js} \geq 4\}$, an indicator $= 1$ if teacher j has a "strength" in skill s. Teacher j is a potential partner teacher if the following three conditions hold:

a. Teacher j was not identified as a target in step 1

b. $\left[\sum_{s=1}^{19} V_{js}\right] \geq 1$, teacher j has one or more "strong" skill areas

c. $\left[\frac{1}{19}\sum_{s=1}^{19} R_{js}\right] \geq 3.5$

Let J^P be the number of potential partner teachers.

3. Define the "optimal set of one-to-one pairings" of teachers

a. Define a "pairing" of teachers

A paring, (j, k), is the combination of one target teacher, j, with one partner teacher, k.

b. Define a "set of one-to-one pairings"

A single "set of one-to-one pairings", p, includes many pairings, $p = \{(1, k), (2, k'), \dots, (J^T, k'')\}$; but each j is in only (at most) one pairing, and each k is in only (at most) one pairing. A school will have P possible different sets of one-to-one pairings.

c. Define a "match score"

For a potential paring of teachers, (j, k), define the "match score" $m(j, k) = \sum_{s=1}^{19} W_{js} * V_{ks}$. The match score is the number of skills on which there is a "weakness" to "strength" match between target and potential partner, respectively.

d. Define "optimal"

For a given target teacher j the best possible partner is $k^* = \underset{k}{\text{argmax}}\big(m(j, k)\big)$, the partner which maximizes the "match score."

But k^* may be the same person for multiple target teachers in a school, and we are working to create a set of one-to-one matches. Thus, we need a school-level objective to optimize. We use the simple sum of match scores for a given set of one-to-one pairings: $M(p) = m(1,k) + m(2,k') + \cdots + m(J^t, k'')$. The "optimal set of one-to-one pairings" is $p^* = \underset{p \in P}{\text{argmax}}\big(M(p)\big)$.

4. Finding the optimal set of one-to-one pairings, p^*

   To find p^* we use the Kuhn-Munkres algorithm (Kuhn 1955, also known as the Hungarian algorithm). We refer the reader to Kuhn (1955) or other presentations for an explanation of the algorithm. The Stata code written for this project is available from the authors.

   The input to the Kuhn-Munkres algorithm is a matrix. In our case the matrix is $J^T \times J^P$. The rows represent target teachers, $j \in J^T$, identified in step 1. The columns represent potential partner teachers, $k \in J^P$, identified in step 2. The cells of the matrix contain the match score, $m(j,k)$, defined in 3.c.[2]

The list of items 1-4 above is the matching algorithm used in the experiment. We now highlight and discuss some key decisions implicit in the algorithm.

In step 1, we define a "weak" skill as $R_{js} < 3$. We chose this threshold for two reasons: (1) It matches the state's threshold for "At expectations." For teacher j to have $R_{js} < 3$ requires that in at least one observation, o, she scored

[2] The standard Kuhn-Munkres algorithm is designed to find $\underset{p \in P}{\text{argmin}}\big(M(p)\big)$, so the matrix cells actually contain $-m(j,k)$.

"(Significantly) Below expectations." When talking with participant teachers we did not use the language "weak skill," instead we would describe these as "skills where you scored below 3." Teachers understood that these were areas where they needed to improve, or at least where their formal performance evaluation indicated they needed to improve.

(2) Empirically, scoring $R_{js} < 3$ is relatively rare, as we knew from looking at the data before defining the algorithm. As shown in Table 1, the proportion of teachers scoring < 3 on a given skill (the sample estimate of $E[W_{js}|s]$) ranges from 0.05 to 0.23 with a median of 0.13. In other words, for a given skill, roughly 1 in 8 teachers is "weak" by our definition. The mean number of weak skills is about 2½, and the mean number conditional on having any weak skills is just under 6.

Our initial goal was to identify approximately one-quarter to one-third of teachers as "target" teachers in step 1. Among our 14 study schools, the mean proportion of target teachers in the school, $\frac{J^T}{J}$, is 0.302. In other words, the marginal target teacher is at about the 30th percentile of performance in her school's distribution.

In step 2, we define a "strong" skill as $R_{js} \geq 4$. Again, our choice of threshold was partly based on the rubric's labels. For teacher j to have $R_{js} \geq 4$ requires that she score "(Significantly) Above expectations" in all of her observations, o. Given the skew in skill scores, this threshold is easier to achieve empirically. As shown in Table 1, the proportion of teachers scoring ≥ 4 on a given skill (the sample estimate of $E[V_{js}|s]$) ranges from 0.22 to 0.72 with a median of 0.48. The mean number of strong skills for teachers is 9 or 10. Thus, our conditions for identifying potential "partner" teachers, in step 2, are fairly inclusive. Among our 14 study schools, the mean proportion target teachers in the school, $\frac{J^P}{J}$, is 0.621.

With steps 1-4 and the details in the prior few paragraphs, we can make the following characterizations of the kinds of pairings which could be created by the matching algorithm. First, there may (will) be pairings of two teachers who are both in the bottom half of the school's teacher performance distribution. The marginal target teacher is at about the 30th percentile, and the marginal potential partner is at about the 40th percentile. These kind of pairings between two seemingly "average" teachers were not unintended. Our goal was not to simply pair "great" teachers with "bad" teachers, where great and bad are defined in some broad all-skills sense. We could have done so with overall observation scores or test-score-based "value-added" scores. There are existing programs for that kind of pairing of teachers on overall performance. Rather, our goal from the beginning was to (1) match on weakness and strength in specific skill areas, and (2) include most (many) of a school's teachers in pairings. Under this goal, we hypothesized, we could use more of a school's own teachers as partners or mentors, because a marginal teacher might be a great mentor in some specific skills even if she was still developing herself in other skills. Similarly, focusing on specific skills would draw in more teachers who could improve in some specific skills even if they were doing well in others. Note, however, that the algorithm does not preclude the overall-great teachers being matched with overall-bad.

Next, we are often asked about whether the algorithm can or did create "bidirectional" matches. That is, matches between teacher A and teacher B where both (i) some of A's weak skills were matched by B's strong skills, and (ii) some of B's weak skills were matched by A's strong skills. First, we simply did not consider bidirectional matches when designing the matching algorithm. When we designed the algorithm we thought of these two types as mutually exclusive. This is implied by steps 1 and 2. By conditions 2.a. and 2.c., a teacher identified as a "target" teacher in step 1 cannot be a potential "partner," even if she has some "strong" skills as defined in step 2. As a result, the chances were limited for creating

bidirectional matches in the experiment. Among all the possible pairings the algorithm considered—i.e., possible pairings defined by a row and column of the matrix in step 4—less than one percent (0.87) were bidirectional.[3] Among the optimal set of one-to-one pairings there were zero bidirectional matches. The lack of bidirectional matches is partly of a function of the specifics of the algorithm: the weak and strong skill thresholds, the match score, the maximization objective used in Kuhn-Munkres, etc.

Moving away from the specifics of the algorithm used in the experiment, one could design a weak-to-strong skill matching process in different ways. Under the one-to-one pairing constraint, any algorithm needs some objective function over some potential pair characteristic(s). But that objective function could be designed to give more weight to bidirectional pairs, or more (less) weight to certain skill areas, or focus on some overall measure of performance, etc. One reasonable question is the extent to which the results in this paper, with its specific matching approach, would generalize to other matching approaches.

References

Kuhn, Harold W. (1955). The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly*, *2*, 83–97.

[3] In the set of all possible combinations of any two teachers in the same school, a $J \times J$ matrix, there are even fewer bidirectional matches, 0.076 percent.

Teacher Name	Potential Match Teacher Name							
George Washington								
	Abigail Adams							
Thomas Jefferson								
	Eliza Hamilton							
Note: An 'x' indicates that the teacher had an average score of 4 or higher on that element, an 'o' indicates an average score of less than 3.								
Instructional Plans (IP)		o	x					
Student Work (SW)			x					
Assessment (AS)		o	x			x		
Expectations (EX)			x			x		
Managing Student Behavior (MSB)			x		o			
Environment (ENV)			x		o			
Respectful Culture (RC)			x			x		
Standards and Objectives (SO)		o	x		o	x		
Motivating Students (MS)			x		o	x		
Presenting Instructional Content (PIC)		o	x			x		
Lesson Structure and Pacing (LS)			x			x		
Activities and Materials (ACT)			x			x		
Questioning (QU)		o	x					
Academic Feedback (FEED)		o	x					
Grouping Students (GRP)		o	x					
Teacher Content Knowledge (TCK)			x			x		
Teacher Knowledge of Students (TKS)			x			x		
Thinking (TH)						x		
Problem Solving (PS)		o			o	x		

Appendix Figure B1—Example of report for treatment school principals showing recommended one-to-one teacher pairings

Talking to Lower-Performing Teachers

- **Provide context**
  - Explain the Evaluation Partnership project and its goals
  - Be clear that they are not bad teachers, just struggling in particular area(s) of practice
- **Make the conversation about them (not you, their partner, or the project)**
- **Frame as opportunity, not obligation**
  - Opportunity to incorporate feedback, improve practice, improve evaluation scores, and develop a relationship with a colleague
  - Opportunity to assess their work and aspirations, and develop a plan that will help them achieve aspirations
- **Listen**
  - Address their personal concerns and encourage questions

Talking to Higher-Performing Partners

- **Provide context**
  - Explain the Evaluation Partnership project and its goals
  - Explain why they were matched, and why you think they would be a good mentor
- **Make the conversation about them (not you, their partner, or the project)**
- **Frame as opportunity, not obligation**
  - Partnership will be beneficial both for the school and for themselves
  - They can learn and benefit from experience: they can hone their own craft and develop transferable skills
  - Partnering is an opportunity to pay it forward and help a colleague
- **Signal their expertise and that they need not feel expert**
  - Acknowledge any concerns but reiterate that they have something to offer
  - More resources on partnership strategies are available on the online portal
- **Listen**
  - Address their personal concerns and encourage questions

Kick-Off Meeting

- **Introduce the teachers to each other (if necessary)**
- **Explain why they are paired, and how they can both benefit**
- **Provide context**
  - Give both parties enclosed letters and materials (if not already distributed)
  - Be clear about goals: partnership should align with your school's mission and values
- **Lay out expectations**
  - Openness about strengths and weakness
  - Consistent contact throughout school year
- **Be enthusiastic about the partnership**
  - A successful program needs a "champion" who is encouraging and believes in the program
- **Point them to existing resources**
  - Refer them to the online portal and encourage them to track progress

Appendix Figure B2—Principal "talking points" for introducing teachers and pairs to the program

Evaluation Partnership Project: A Guide

Partnership Tips

1. ***Build Trust.*** Set shared relationship norms, including when and how to communicate.
2. ***Be vulnerable and open.*** Don't try to impress your partner. The best way to learn is to be open about your strengths and weaknesses.
3. ***Help your partner help you.*** Seek advice, listen, ask questions, & be willing to try things out.
4. ***Seek direction, not results.*** Don't expect immediate results; work steadily toward your goals.
5. ***Set multiple concrete goals, and track progress.*** Goals should be varied and have clear timelines for completion with markers of "success".
6. ***Give and ask for feedback.*** Track what works.

Suggested Activities

- Define goals for the school year
- Review evaluation feedback together
- Set up regular meeting times
- Observe each other's classroom practice
- Ask for (and give) constructive feedback
- Review lesson plans & classroom strategies
- Develop strategies and share advice
- Follow up on efforts to improve
- Log goals, meetings, activities, & progress on-line

Recommended Partnership Timeline

1. First Meeting
When: As soon as possible
Purposes: Set norms, review & discuss evaluation results, and develop goals.

2. Observation(s)
When: Regularly (first observation this month)
Purposes: Identify effective practices & areas for improvement. Compare goals to feedback.

3. Further Meetings and Other activities
When: On a regular basis
Purposes: Give feedback. Review plans. Develop & refine strategies. Track goals. See above for ideas.

4. Final Meeting
When: Last weeks of school year
Purposes: Assess progress on goals, and how to build on skills next year.

Appendix Figure B3—Program guide provided to teachers

Teacher Name		Potential Match Teacher Name	Instructional Plans (IP)	Student Work (SW)	Assessment (AS)	Expectations (EX)		Managing Student Behavior (MSB)	Environment (ENV)	Respectful Culture (RC)		Standards and Objectives (SO)	Motivating Students (MS)	Presenting Instructional Content (PIC)	Lesson Structure and Pacing (LS)	Activities and Materials (ACT)	Questioning (QU)	Academic Feedback (FEED)	Grouping Students (GRP)	Teacher Content Knowledge (TCK)	Teacher Knowledge of Students (TKS)	Thinking (TH)	Problem Solving (PS)	
George Washington			o		o							o						o	o				o	
		Abigail Adams	x	x	x	x		x	x	x		x	x	x	x	x	x	x	x	x	x			
		Dolly Madison	x	x	x	x		x	x	x		x	x	x	x	x	x		x	x	x			
		Mercy Otis Warren	x		x	x		x	x	x			x	x	x	x		x		x	x			
		Martha Washington	x		x	x		x	x	x			x	x	x	x		x		x	x			
		Eliza Hamilton			x	x				x		x	x	x	x	x					x	x		
Thomas Jefferson								o	o			o	o										o	
		Abigail Adams	x	x	x	x		x		x		x	x	x	x	x	x	x	x	x	x			
		Eliza Hamilton			x	x				x		x	x	x	x	x					x	x		
		Dolly Madison	x	x	x	x		x	x	x		x	x	x	x	x	x		x	x	x			
		Mercy Otis Warren	x		x	x		x	x	x			x	x	x	x		x		x	x			
		Martha Washington	x		x	x		x	x	x			x	x	x	x				x	x			

Note: An 'x' indicates that the teacher had an average score of 4 or higher on that element, an 'o' indicates an average score of less than 3.

Appendix Figure B4—Example of report for treatment school principals showing additional potential partners, if needed

Appendix C: Attrition

This appendix provides further detail on the topic of attrition. In it we (i) describe the scope for attrition to threaten identification and interpretation of the estimated treatment effects, (ii) estimate differences in attrition across treatment conditions, and (iii) estimate bounds on effects following Lee (2009) and Horowitz and Manski (2000).

The school-level treatment effect estimates—specifically Table 3 panel B—are not threatened by attrition, at least not in the traditional sense. No schools attrited. This is notable because the experimental design was to randomly assign schools to treatment (the partnership program) or control, and thus the first-order treatment effect estimate is the school level estimate.

While no schools attrited, some teachers did attrit. Teacher attrition is relevant to the interpretation of the school-level effect estimates, even if those estimates are not threatened by attrition in the traditional sense. Moreover, teacher attrition is certainly relevant to identification and interpretation of teacher-level effect estimates, like the separate effects for target, partner, and other teachers in Table 3 panel C and Table 4.

C.1 Teacher attrition during the experiment year

We begin with (the potential for) teacher attrition during the experiment year, t. Throughout the paper, with only one exception, all student test score outcomes are tests taken at the end of the experiment year, 2013-14. (That one exception is Table 3 column 3, where the outcome is tests taken at the end of the year following the experiment, 2014-15. We discuss teacher attrition for these estimates later.) Thus, with that one exception, the attrition concern is limited to teachers attriting during the experiment year.

What constitutes attrition in this case? As detailed in Appendix B, at the time of random assignment we identified 141 teachers who met two criteria: (i) we expected them to teach math or reading/language arts (or both) in grades 4-8 during

the experiment year, and (ii) we had prior classroom observation score data for them which was necessary for inclusion in the teacher matching algorithm. At the end of the experiment year, 5 of the 141 did not have student test score data—that is, they had not taught math or reading/language arts in grades 4-8 (at least not for any substantial portion of the year). Thus, a fairly-low attrition rate of 3.6 percent.

The attrition rate is low largely because the scope for attrition during the experiment year was limited. Schools were randomly assigned to treatment or control October 2, 2013, after the school year had already begun and teachers were teaching students.

In Appendix Table C1 column 1 we report treatment effects on attrition during the experiment year, t. There is no difference in overall attrition rates; the point estimate is zero and far from statistically significant (panel A). Additionally, no target teachers attrited, neither treatment targets nor control targets. Recall that roles were assigned by an algorithm, and thus we observe assigned role in both treatment and control schools. Thus, the overall treatment effects and target teacher effects are unlikely to be threatened by teacher attrition. There are some potential differences for partner teachers and no role teachers, though the differences are not statistically significant.

To examine the sensitivity of our estimates to teacher attrition, we calculate Manski-style bounds (Horowitz and Manski 2000). The intuition, briefly, is to first impute missing outcomes for attriters with (i) the highest possible value for attriting control units, and (ii) the lowest possible value for attriting treatment units. This provides the lower bound. Then reverse the imputation to find the upper bound. The upper bound can be written in population terms:

$$(1-p^1)E[Y|T=1]+p^1\min(Y)-[(1-p^0)E[Y|T=0]+p^0\max(Y)]=\delta^{UB} \quad \text{(C.1)}$$

where $p^t = E[1\{attrit\}|T=t]$, $t \in \{1,0\}$, is the probability of attriting given treatment status. Strictly speaking, these bounds are undefined when the outcome

is unbounded, but we provide estimates using plausible values for $\min(Y)$ and $\max(Y)$. Moreover, as min-max range gets wider the approach is less and less informative, because of the strong assumption that attriters are at the extremes of the distribution.

Estimates of Manski-style bounds are provided in Appendix Table C1 columns 3 and 4. To estimate the bounds we first set $E[Y|T=0]=0$, thus $E[Y|T=1]=\delta$. Our estimates of each δ are repeated in Appendix Table C1 column 2 for convenience. Our estimate of p^0 is the "Control attrition rate" reported in column 1, and $p^1 = p^0 +$ the attrition difference estimate also in column 1. We set $\max(Y) =$ 99th percentile of the teacher "value-added" distribution, and $\min(Y) =$ 1st percentile. "Value-added" is short hand for the teacher's contribution to student test scores. The between-teacher standard deviation in value-added is typically estimated to be 0.15σ (see Hanushek and Rivkin 2010, and Jackson, Rockoff, and Staiger 2014). Thus we set $\max(Y) = 0.15 * 2.33 = 0.35$, and $\max(Y) = -0.35$.[1]

For the average treatment effect, the bounds are 0.038 to 0.088. These bounds are not dramatically different from our estimate of 0.065, despite the strong assumption of the Manski-style approach that attriters are those at the extremes of the distribution. In the case of partner (no-role) teachers the lower bounds approach (include) zero, but the main estimates for these two groups also suggest there was no significant effect.

C.2 Teacher attrition for the year following the experiment

We now turn to (the potential for) teacher attrition during the year following the experiment, $t+1$. This attrition concern is relevant to the estimates in is Table 3 column 3, where the outcome is tests taken at the end of the year following the

[1] This assumes value-added is normally distributed, which is also consistent with existing literature.

experiment, 2014-15. In short, the purpose of these $t+1$ estimates is to measure whether changes in teacher performance persisted after the experiment year when teachers had been assigned a new group of students.

Attrition rates were much higher in $t+1$. Of the original 141 teachers, we have student test score data for only 96 at the end of $t+1$, an attrition rate of 31.9 percent. These 96 are teachers who taught math or reading/language arts in grades 4-8 in the school year following the experiment, 2014-15.[2] The higher attrition rates are partly due to much greater scope for normal job turnover. The 40 teachers who attrited between t and $t+1$ will include teachers who were reassigned to other subjects or grade levels, teachers taking leave, retirements, teachers who moved outside Tennessee, etc. Such changes may have been affected by treatment, yes, but there is much more scope for change than there was during the experiment year.

In Appendix Table C2 column 1 we report treatment effects on attrition between t and $t+1$. In the treatment group 31.6 percent of teachers attrited compared to 27 percent in the control group; the difference is not statistically significant (panel A). Attrition rates were highest for target teachers: 32.3 percent of control target teachers attrited, and 39 percent of treatment target teachers; though again the difference is not statistically significant (panel B).

The patterns of attrition suggest that exposure to the treatment program may have induced more target teachers to turnover. We estimate that treatment target teachers were performing better even in the year following the experiment. But that positive estimate may be due in part to differential attrition; for example, it may be that among target teachers the especially low-performing teachers were more likely to turnover when treated.

[2] Changing schools between t and $t+1$ is necessarily not attrition. We can follow teachers anywhere they teach in Tenessee's public schools.

To examine the sensitivity of our estimates to teacher attrition, we calculate both Manski-style bounds (Horowitz and Manski 2000) and Lee-style bounds (Lee 2009). The Manski-style bounds are estimated as described in section C.1 above, and reported in Appendix Table C2 columns 3 and 4. Not surprisingly, the Manski-style bounds are much wider, and thus not very informative, because of the higher attrition rates in both treatment and control.

Lee (2009) defines an alternative bounding approach, typically resulting in much tighter bounds than the Manski approach. The intuition, briefly, is to "trim" (drop from the data) control observations until the new attrition rate in the control equals the treatment attrition rate (or the reverse if initially the attrition rate is higher in the control). For the upper bound, trim control observations with the highest observed outcome values. For the lower bound, the reverse. These bounds are tighter in part because they do not require imputing extreme values to all attriters; indeed, there is no imputation of missing values. Estimates of Lee-style bounds are provided in Appendix Table C2 columns 5 and 6.

Our setting requires a modification to the standard Lee bounds procedure. We are addressing attrition at the teacher level, but our outcome data and regression specifications are at the student level. Thus, determining which teachers to "trim" is not is not a simple function of a scalar observed dependent variable.

Our Lee-style bounds are estimated as follows: Assume for this explanation that the attrition rate is higher in the treatment, as it is overall in year $t+1$. First, estimate the number of control teachers to trim, $N_{trim} = (p^1 - p^0) * N^0$, where N^0 is the number of control teachers. Sample estimates of the p^t terms are taken from Appendix Table C2 column 1.[3] Second, note that there are $\binom{N^0}{N_{trim}} = M$ different ways to trim N_{trim} teachers from the data. Let $\hat{\delta}_m$ be the treatment effect estimated

[3] Our estimate of $(p^1 - p^0) * N^0$ may not be an integer. In Appendix Table C2, we use ceil$[(p^1 - p^0) * N^0]$, but the estimates are robust to using the floor instead.

for a given possible trim m. The Lee-style bounds are $\min(\hat{\delta}_m)$ and $\max(\hat{\delta}_m)$. We find $\min(\hat{\delta}_m)$ and $\max(\hat{\delta}_m)$ by simply estimating all M possible trims. This is computationally feasible because M is on the order of thousands in our setting. Since we are interested in estimating treatment effects by teacher role, we find N_{trim}, $\min(\hat{\delta}_m)$, and $\max(\hat{\delta}_m)$ separately for each role (see Lee 2009 Section 3.2).

References

Horowitz, Joel L. & 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.

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

Appendix Table C1—Teacher attrition during the experiment year

		Treatment effect estimate		
		Main (Table 3	Manski-style bounds	
	Attrited	col 2)	Lower	Upper
	(1)	(2)	(3)	(4)
(A) Average treatment effects				
All teachers	-0.000	0.056	0.029	0.079
	[0.968]	[0.080]		
	(1.000)	(0.250)		
Control attrition rate	0.036			
(B) Treatment effects by teacher role				
Low-performing target teachers	a	0.123	a	a
		[0.000]		
		(0.031)		
Control attrition rate				
High-performing partner teachers	-0.109	0.029	0.003	0.056
	[0.052]	[0.252]		
	(0.203)	(0.547)		
Control attrition rate	0.092			
No assigned role	0.122	0.029	-0.012	0.063
	[0.108]	[0.468]		
	(0.125)	(0.625)		
Control attrition rate	-0.007			

Note: Column 1: Each column within panels reports estimates from a separate linear probability model regression, with 141 teacher observations. The dependent variable is an indicator = 1 if the teacher attrited during the experiment year t. Specifically, attrited = 1 if the teacher did not teach math or reading/language arts in grades 4-8 during the experiment year. The sample is 141 teachers. Column 2 simply repeats Table 3 column 2 for convenience. Please see the note on Table 3 for details. Columns 3 and 4 report Manski-style (Horowitz and Manski 2000) bounds. The calculation is described in the text. 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.

(a) No target teachers attrited, neither treatment targets nor control targets.

Appendix Table C2—Teacher attrition the year following the experiment

		Treatment effect estimate				
		Main (Table 3	Manski-style bounds		Lee-style bounds	
	Attrited	col 3)	Lower	Upper	Lower	Upper
	(1)	(2)	(3)	(4)	(5)	(6)
(A) Average treatment effects						
All teachers	0.046	0.106	-0.132	0.277	0.051	0.141
	[0.568]	[0.220]			[0.556]	[0.212]
	(0.547)	(0.375)			(0.563)	(0.375)
Control attrition rate	0.270					
(B) Treatment effects by teacher role						
Low-performing target teachers	0.067	0.252	-0.095	0.403	0.207	0.340
	[0.724]	[0.068]			[0.052]	[0.040]
	(0.703)	(0.422)			(0.438)	(0.422)
Control attrition rate	0.323					
High-performing partner teachers	-0.001	0.056	-0.145	0.227	-0.025	0.072
	[0.976]	[0.684]			[0.776]	[0.624]
	(0.969)	(0.641)			(0.859)	(0.641)
Control attrition rate	0.267					
No assigned role	0.098	0.013	-0.193	0.210	0.008	0.047
	[0.400]	[0.908]			[0.864]	[0.536]
	(0.500)	(0.891)			(0.906)	(0.563)
Control attrition rate	0.240					

Note: Column 1: Each column within panels reports estimates from a separate linear probability model regression, with 136 teacher observations. The dependent variable is an indicator = 1 if the teacher attrited in the year follow the experiment. Specifically, attrited = 1 if the teacher did not teach math or reading/language arts in grades 4-8 in Tennessee public schools during year $t + 1$. The sample is 136 teachers. Column 2 simply repeats Table 3 column 3 for convenience. Please see the note on Table 3 for details. Columns 3 and 4 report Manski-style (Horowitz and Manski 2000) bounds. The calculation is described in the text. Columns 5 and 6 report Lee-style (Lee 2009) bounds. Our setting requires a modified Lee bounds approach; please see the text for details. 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.