MATHEMATICA
Policy Research

FINAL REPORT

Measuring Teachers' Effectiveness: A Report from Phase 3 of Pennsylvania's Pilot of the Framework for Teaching

April 23, 2015

Stephen Lipscomb
Jeffrey Terziev
Duncan Chaplin

Submitted to:

Team Pennsylvania Foundation
100 Pine Street
9th Floor
Harrisburg, PA 17101

Project Officer: Ms. Abby Smith
Contract Number: 90982

Submitted by:
Mathematica Policy Research

Project Director: Duncan Chaplin
Reference Number: 40420-300

This page has been left blank for double-sided copying.

ACKNOWLEDGMENTS

This report benefited from the support of the Team Pennsylvania Foundation, the Pennsylvania Department of Education, and the Bill & Melinda Gates Foundation. We wish to thank Abby Smith and Rich Maraschiello for their guidance during this project, and John Nau, David Ream, and Thresa Stafford for their help in obtaining the data needed to do this work. We also want to thank the educators (teachers, principals, and support staff) who participated in Phase 3 of the Teacher Evaluation Pilot and are thereby helping to improve the teacher evaluation systems used in Pennsylvania and elsewhere.

We would like to recognize several Mathematica Policy Research staff who made substantive contributions to this report. Brian Gill provided helpful comments and contributions. Brynn Hagen and Alma Vigil provided excellent programming support. The paper was edited by Maura Butler and produced by Autumn Parker.

Any opinions expressed herein are those of the authors and do not necessarily represent the views of the Team Pennsylvania Foundation, the Pennsylvania Department of Education, or the Bill & Melinda Gates Foundation.

This page has been left blank for double-sided copying.

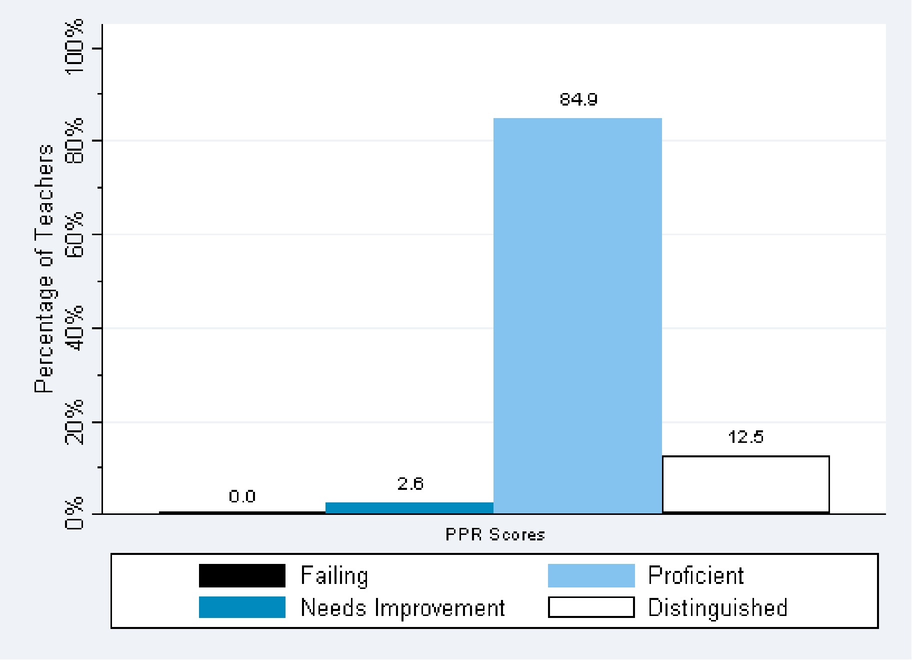
SUMMARY

Like many states throughout the nation, Pennsylvania is in the midst of major reforms to its teacher evaluation system. Under the new system, the state will base annual evaluations on several measures, including supervisor observations using the Framework for Teaching (FFT) and, for many teachers, their contributions to student achievement growth from a value-added model (VAM). In the past, there was concern that supervisor observations did not differentiate performance well or relate to true teacher performance. In this study, we investigate how well the new system has addressed these issues by analyzing the degree to which FFT scores differentiate performance, are internally consistent, and correlate with teachers' contributions to student achievement growth as measured by VAM scores.

This report is based on data from a pilot of the new system covering 6,676 teachers from 269 districts in the state of Pennsylvania, including Pittsburgh public schools. The data include the 22 components of the FFT, each of which is designed to capture a separate teaching practice. We used these data to estimate four domain scores and one overall Professional Practice Rating (PPR) score. We also merged these scores with data on teachers' estimated contributions to student achievement growth.

Based on these pilot data from the 2012–2013 school year, we estimate that, although less than 13 percent of teachers received the top rating (*distinguished*) for the overall PPR score, almost 85 percent were rated in the second highest category (*proficient*) (Figure S.1). Less than 0.1 percent were rated in the bottom category (*failing*). The remaining teachers (around 2.6 percent) were given *needs improvement* ratings.

Figure S.1. Distribution of Professional Practice Ratings



Source: Authors' calculations based on FFT pilot evaluation scores from the 2012–2013 school year provided by PDE.

Notes: See Appendix A, Table A.2 for information on proportions and sample sizes.

We found that FFT scores were internally consistent, meaning that the domains and the components within each domain appear to be measuring similar concepts. We also found that teachers with higher FFT scores tended to produce greater student achievement growth. The correlations of the FFT scores with VAM scores were all positive and generally statistically significant, ranging from 0.19 to 0.22 by domain.

We compared the results based on the 2012–2013 data with results based on 2011–2012 data from a previous pilot phase. For the most part, the findings were similar. More than 90 percent of teachers were rated in the top two performance categories in both phases, although the fraction of ratings in the top two categories decreased somewhat in Pittsburgh (which contributed more teachers to the pilot than any other district). The levels of internal consistency were in the acceptable to good ranges in both phases, with the overall PPR score having higher consistency than any of the domain scores in both phases. The correlations between parts of the FFT and VAM scores were almost always positive but also below 0.30 in both phases. The lowest correlations in 2011–2012 were slightly improved in 2012–2013.

In sum, although FFT scores are overwhelmingly concentrated in the top two performance categories, the positive correlations with VAM suggest that the FFT provides some meaningful differentiation and captures aspects of teacher skills related to student achievement growth.

CONTENTS

TABLES

FIGURES

I. INTRODUCTION

A. Rationale for the study

Pennsylvania is among many states that are developing and implementing new tools to evaluate teachers. Under recently enacted Pennsylvania law, the state must base half of a teacher's annual evaluation rating on a measure in which a supervisor—typically the school principal—judges the quality of the teacher's professional practices. For this purpose, the Pennsylvania Department of Education (PDE) is employing the Framework for Teaching (FFT), a commonly used classroom observation tool developed by Charlotte Danielson. Measures of student achievement form the basis for the remaining half of each teacher's annual evaluation rating.[1]

During the 2013–2014 school year, PDE implemented the FFT classroom observations statewide. The student achievement measures are being implemented statewide during the 2014–2015 school year. In preparation for statewide implementation of these evaluation measures, PDE conducted the Pennsylvania Teacher Evaluation Pilot in a subset of districts in three phases, starting in the 2010–2011 school year and continuing through the 2012–2013 school year. The pilot explores several aspects of the evaluation system, including the following:[2]

- The degree of variation across teachers in professional practice scores using the FFT
- The degree of internal consistency of the FFT
- The degree to which higher or lower FFT scores are indicative of teachers who make larger or smaller contributions to their students' growth in achievement

Mathematica Policy Research has been exploring these issues at the behest of PDE, using funding from the Team Pennsylvania Foundation (Team PA). Team PA, in turn, received the funds used for this research from the Bill & Melinda Gates Foundation. The findings from this research are being used to inform the early implementation of Pennsylvania's new teacher evaluation system. Mathematica has already examined data from the first two phases of the pilot. This study examines the third phase of the pilot.

B. The Framework for Teaching

The FFT is a classroom observation tool that school districts across the country use to evaluate teacher performance. The FFT specifies 22 teaching practices, known as components. Evaluating supervisors, typically school principals, rate teacher performance on each component using four performance categories: *distinguished* (3 points), *proficient* (2 points), *needs improvement* (1 point), or *failing* (0 points).[3] FFT components are grouped into four domains: (1)

[1] Measures of student achievement include value-added assessment system data, building-level achievement data, student learning objectives, and other measures.

[2] Another primary objective of the pilot was to collect information about principals' and teachers' experiences using the FFT during the pilot. Mathematica will not be collecting or analyzing this information.

[3] The component ratings are based on direct observation by a supervisor.

planning and preparation, (2) classroom environment, (3) instruction, and (4) professional responsibilities (see Appendix A, Table A.1 for a list of the components in each domain).

Supervisors also assign domain scores using the same rating scale used for the component scores (3, 2, 1, or 0 points), based on the preponderance of evidence from each domain. The phase 3 data collection did not include teachers' domain scores, only their component scores. We estimated teachers' domain scores by averaging their component scores from each domain, producing measures that are continuous between zero and three.[4] The domain scores are, in turn, averaged to estimate each teacher's Professional Practice Rating (PPR), the professional practice measure that constitutes half of teachers' overall evaluation scores.[5] For both the domain scores and the PPR, we followed PDE's rating tool for recording scores based on the new evaluation system and assigned scores to performance categories as follows: 0–0.5 (*failing*), 0.5–1.5 (*needs improvement*), 1.5–2.5 (*proficient*), 2.5–3.0 (*distinguished*). Table I.1 provides more information on the component, domain, and PPR scores.

Table I.1. Description of component, domain, and PPR scores

	Components (as implemented and as used in study)	Domains (as implemented)	Domains (as used in study)	PPR (as implemented and as used in study)
Number	22	4	4	1
Score values	0, 1, 2, or 3	0, 1, 2, or 3	0–3 (continuous)	0–3 (continuous)
Method of obtaining scores	Supervisors' perceptions based on classroom observations	Supervisors' perceptions of the preponderance of evidence from the component scores	Unweighted average of component scores	Weighted average of domain scores

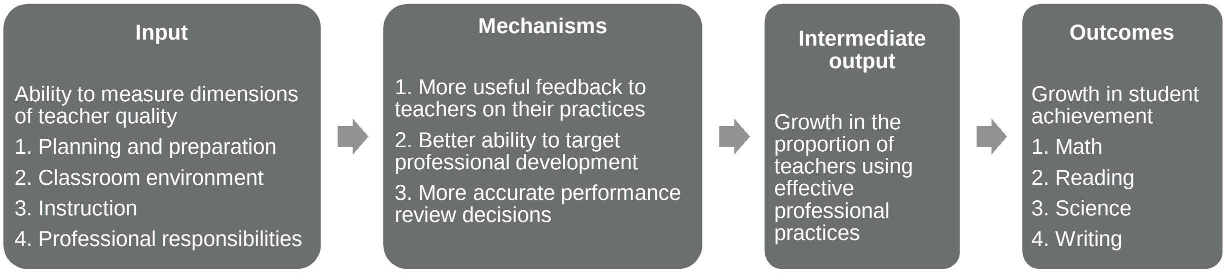
Notes: PPR = Professional Practice Rating. In calculating the PPR, domains 1 and 4 each receive a 20 percent weight. Domains 2 and 3 each receive a 30 percent weight. The components are the 22 teaching practices that constitute the FFT.

PDE is using the FFT to improve the evaluation of teacher effectiveness, with the ultimate goal of improving student outcomes (Figure I.1). Using the FFT for teacher evaluations may (1) prompt teachers to align their practices to the FFT, (2) improve the ability of schools and districts to target professional development opportunities, and (3) improve schools' and districts' ability to judge performance. This, in turn, may increase the extent to which teachers use effective professional practices in their classrooms to improve student achievement.

[4] In contrast, PDE's use of whole numbers will reduce the precision of the domain scores, potentially undermining both the ability of the FFT to differentiate teacher performance and the extent of correlations with value added to some extent.

[5] The PPR is a weighted average of domain scores. Domains 1 and 4 each receive a 20 percent weight. Domains 2 and 3 each receive a 30 percent weight.

Figure I.1. Conceptual framework for how measuring teachers' practices can improve student outcomes


Input
Ability to measure dimensions of teacher quality
1. Planning and preparation
2. Classroom environment
3. Instruction
4. Professional responsibilities
Mechanisms
1. More useful feedback to teachers on their practices
2. Better ability to target professional development
3. More accurate performance review decisions
Intermediate output
Growth in the proportion of teachers using effective professional practices
Outcomes
Growth in student achievement
1. Math
2. Reading
3. Science
4. Writing

C. The Pennsylvania Teacher Evaluation Pilot

The Pennsylvania Teacher Evaluation Pilot implemented the FFT with groups of teachers before introducing the tool statewide. Pilot evaluations served only to provide information; they were not used for formal evaluative purposes. A broad stakeholder steering committee selected the FFT in fall 2010, and four school districts agreed to participate in the first trial implementation during the spring of 2011 (phase 1). The pilot was expanded during the 2011–2012 school year to include 2,621 teachers from 105 districts (phase 2). The pilot was further expanded during the 2012–2013 school year to include 6,676 teachers from 269 districts (phase 3). The 2012–2013 pilot also placed greater emphasis on principal training, including offering the opportunity for reliability certification using the FFT developer's "gold standard" reviews.

1. Findings from phases 1 and 2

Mathematica's studies of the first two pilot phases found that the FFT produced limited differentiation in ratings of teacher performance, but the small amount of differentiation found was positively correlated with teachers' contributions to their students' growth in achievement. At least 90 percent of teachers received *proficient* or *distinguished* ratings on most components, and very few teachers received *failing* ratings (Walsh and Lipscomb 2013; Lipscomb et al. 2012). Consistent with other research on the FFT (Milanowski 2011; Kane and Staiger 2012), teachers receiving higher FFT scores in phase 2 were more likely to make larger contributions to student growth as measured by a value-added model (VAM).[6] VAMs make predictions about students' achievement scores based on their own test score history and background characteristics, where the prediction is based on how particular students would perform if served by the average teacher in the state. The average difference between actual and predicted achievement (positive or negative) across a teacher's students is considered a measure of the teacher's "value added," or effectiveness, relative to the average teacher.

The findings from the first two pilot phases can be interpreted only within a narrow context due to several features of the design of these phases. In particular, only four districts participated in the first pilot, severely limiting the external validity of the findings. Although the second pilot was larger, nearly two-thirds of the participating teachers taught in a single district (Pittsburgh Public Schools), again limiting the generalizability of findings. In addition, PDE issued different

[6] See Lipscomb et al. (2010) for a review of the value-added literature on teacher evaluation that Mathematica prepared as part of the first pilot study.

instructions to principals in phase 2 about the number of FFT components to use in their teachers' evaluations, compared with PDE's guidance for actual evaluations. Specifically, teachers in phase 2 were evaluated on a consistent set of three components that measured their mastery in planning coherent instruction, engaging students in learning, and using assessments to inform instruction. Principals were then instructed to choose at least five other components so that teachers were assessed on at least two components from each of the FFT's four domains. These instructions led to substantial differences among teachers for which components were used and, therefore, limited the comparability of teachers' domain scores. For actual teacher evaluations based on the FFT, PDE recommends that principals use all the components for which they feel evidence to support a rating exists.

2. Overview of phase 3

Phase 3 differed from the previous two phases in several important ways. First, phase 3 was more than 2.5 times larger than phase 2, in terms of the number of teachers, schools, and districts participating (Table I.2). Second, the proportion of participants from Pittsburgh was substantially smaller (16 versus 64 percent). Third, it included a greater emphasis on principal training, including opportunities for principals to compare their ratings on practice evaluations with official ratings from the FFT developer. Fourth, PDE's instructions to principals in phase 3 about the number of components to use mirrored the guidance for actual evaluations. Principals in phase 3 used, on average, 20 of the 22 components. In contrast, many principals in phase 2, especially those outside of Pittsburgh, used the minimum allowable number of components. The fact that phase 3 principals used nearly all components means that domain scores are more comparable across teachers, because they pertain to a mostly consistent set of practices.

Table I.2. The number of teachers, schools, and districts in the phase 2 and phase 3 pilots

		Sample sizes		
	Phase	Pittsburgh	Not Pittsburgh	Total
Teachers	2	1,673	948	2,621
	3	1,038	5,638	6,676
Schools	2	64	248	312
	3	58	849	907
Districts	2	1	104	105
	3	1	268	269

Source: Authors' calculations based on FFT pilot evaluation scores from the 2011–12 and 2012–13 school years provided by PDE.

Note: The number of Pittsburgh teachers declined between phases 2 and 3, because the data that Pittsburgh provided in phase 3 did not include teachers in the district's Supported Growth Project (SGP). These teachers have previously demonstrated proficiency in their teaching practices and do not participate in the formal observation process. They instead agree to be rated on a single focal component and carry forward their FFT scores from the previous year.

The phase 3 teacher sample resembled teachers across Pennsylvania according to some observable characteristics but not others (Table I.3). For example, the proportion of teachers in phase 3 who are female was representative of the teacher workforce statewide. However, phase 3 teachers were more likely than other teachers in Pennsylvania to be white, less likely to be Asian, more likely to have five or fewer years of experience, and less likely to have a master's degree. In addition, they had lower annual salaries on average.

Although teachers in phase 3 had different observable characteristics than other teachers in the state, the phase 3 sample more closely reflected teachers statewide than the phase 2 sample did. In particular, the race/ethnicity distribution in phase 3 and the proportion of teachers with a master's degree were closer to statewide teacher averages, compared with the phase 2 teacher sample (Walsh and Lipscomb 2013). The relatively smaller proportion of teachers from Pittsburgh in the phase 3 sample appears to partly be responsible for these changes.

Table I.3. Characteristics of Pennsylvania teachers in phase 3 and not in phase 3

Characteristic	Pennsylvania (not phase 3)	Phase 3		
		All	Pittsburgh	Not Pittsburgh
Female (percentage)	73.9	73.7	73.3	73.7
Race/ethnicity				
White (percentage)	94.2	95.0*	85.9 #	96.7
African American (percentage)	4.1	3.7	13.1 #	2.1
Hispanic (percentage)	0.9	0.7	0.3	0.7
Asian (percentage)	0.6	0.4*	0.6	0.3
Other race/ethnicity (percentage)	0.2	0.2	0.1	0.2
Total experience				
Five years or fewer (percentage)	20.1	26.9*	20.6 #	28.1
More than five years (percentage)	79.9	73.1*	79.4 #	71.9
Educational attainment				
Master's degree or higher (percentage)	54.8	45.4*	33.4 #	47.6
Bachelor's degree (percentage)	43.9	54.2*	66.2 #	52.0
Less than bachelor's degree (percentage)	1.3	0.4*	0.4	0.4
Annual salary ($)	$63,674	$59,339*	$73,327 #	$56,821
Number of Teachers	136,028	6,445	978	5,467

Source: Mathematica calculations based on data from Pennsylvania's longitudinal student database.

Notes: Test statistics allow for unequal variances across samples. We were unable to obtain data on background characteristics for 231 phase 3 teachers. These teachers are excluded from this table.

* Difference between Pennsylvania teachers participating in phase 3 and those not participating in phase 3 is statistically significant at the 5 percent level. Symbols are reported only in the column for the overall phase 3 sample.

# Difference between phase 3 Pittsburgh and non-Pittsburgh teachers is statistically significant at the 5 percent level. Symbols are reported only in the column for the phase 3 Pittsburgh sample.

D. Research questions

This study uses data from the third phase of the pilot to address three research questions:

1. **To what extent do FFT scores vary across teachers in phase 3, and how does this variation compare with phase 2?** The degree of score variation is an indication of how well the FFT differentiates between high-performing and low-performing teachers. When professional practice scores are at the high end of the scale—as they were in phase 2—the FFT may be less useful for distinguishing teaching effectiveness. We examine the distribution of FFT scores in phase 3, which may look different than in phase 2 because of the smaller proportion of teachers from Pittsburgh and the more rigorous training that principals received.
2. **How internally consistent are teachers' FFT ratings, and how does this consistency compare with phase 2?** Internal consistency measures the degree to which different parts of the FFT reach similar conclusions about a teacher's effectiveness. We attempt to confirm findings from phase 2 suggesting that the FFT and its domains have good or acceptable internal consistency, using the broader phase 3 teacher sample and in the context of more rigorous evaluator training.
3. **How strongly correlated are teachers' FFT scores and their estimated contributions to their students' growth in achievement, and how does this correlation compare with phase 2?** The strength of this correlation is a test of the validity of the conceptual framework underlying the use of the FFT. Findings from phase 2 suggested that teachers with higher FFT scores tend to be those who make larger contributions to their students' growth in achievement. This finding was particularly true for instructional practices. We re-examine and attempt to confirm these relationships in the broader phase 3 pilot. Specifically, we calculate teachers' contributions to student achievement growth among all 4th through 8th grade teachers in the state, and for teachers included in phase 3, correlate their VAM estimates with their FFT scores.

II. NEARLY ALL TEACHERS IN PHASE 3 RECEIVED PROFICIENT OR DISTINGUISHED FRAMEWORK FOR TEACHING SCORES, AS IN PHASE 2

A goal of any teacher evaluation system is to distinguish between higher-performing and lower-performing teachers. To achieve this objective, an evaluation system must have the capacity to give teachers different evaluation scores. If an evaluation system assigns similar scores to teachers who, in fact, vary in their effectiveness, the system will have limited usefulness for differentiating teachers' performance levels. In phase 2, at least 90 percent of teachers received either a *distinguished* or *proficient* rating on most components (Walsh and Lipscomb 2013). Although we do not know what the ideal distribution of FFT ratings should be, these findings suggested that the FFT, as implemented in phase 2, differentiated teacher performance only to a limited degree. One possible contributor might have been that some principals did not apply the FFT as it was intended to be used, and were not using the two lowest performance categories. This concern led PDE to provide more rigorous training to principals in phase 3 in how to use the FFT. The phase 3 training included opportunities for principals to compare their ratings on practice evaluations against official ratings by the FFT developer. Because of the potential impact of the additional training, we examine in this study the variation in FFT scores obtained during phase 3, and explore the change in the distribution of these scores between phases 2 and 3. These analyses shed light on the likely ability of PDE's evaluation system to distinguish between more and less effective teachers.

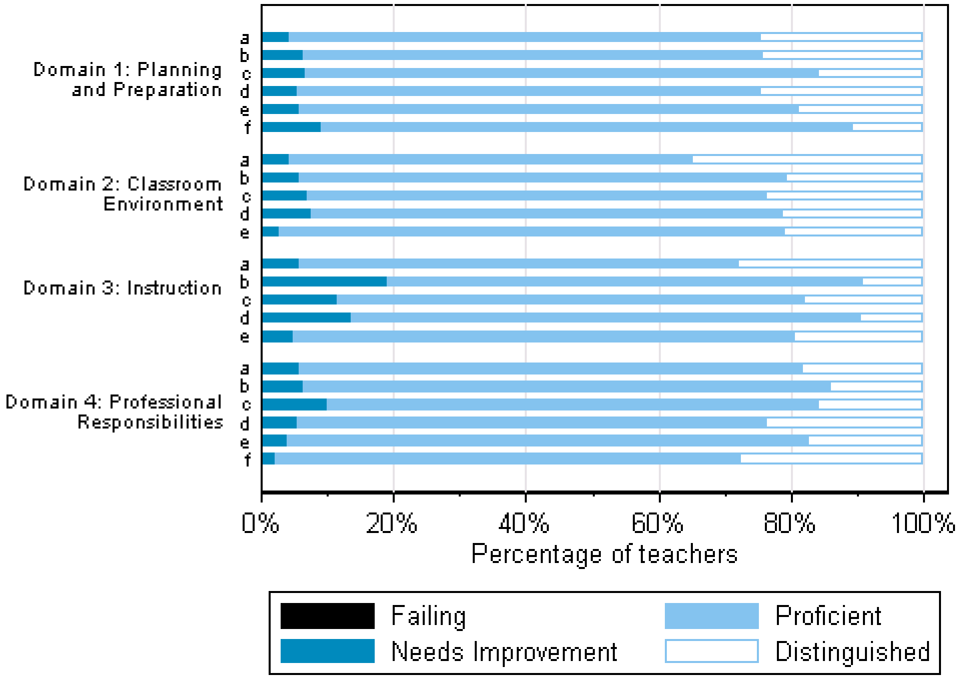
A. Summary of Framework for Teaching scores obtained during phase 3

1. More than 90 percent of teachers received proficient or distinguished Framework for Teaching ratings on most components and domains during phase 3

On 19 of 22 components, 10 percent or fewer teachers received a *failing* or *needs improvement* rating, meaning that 90 percent or more were rated as *proficient* or *distinguished* (Figure II.1; see Appendix A, Table A.1 for more detail). *Proficient* was the most common rating. Between 60.7 and 79.7 percent of teachers received a *proficient* rating depending on the component, and overall, 72.7 percent of all ratings were *proficient* across all components. *Distinguished* was typically the second most common rating. Between 9.5 and 35.0 percent of teachers received a *distinguished* rating depending on the component, and overall, 20.3 percent of all ratings were *distinguished* across components. In total, 93 percent of all FFT component ratings were either *proficient* or *distinguished*.

On three components, somewhat larger proportions of teachers in phase 3 received *failing* or *needs improvement* ratings. All three of these components were in the instruction domain: 3b (using questioning and discussion techniques), 3c (engaging students in learning), and 3d (using assessment to inform instruction). For each of these components, 0.2 percent of teachers received a *failing* rating, consistent with the proportion receiving a *failing* rating for other components. However, between 11.3 and 19.0 percent of teachers received *needs improvement* ratings, larger than for other components. Principals in phase 3 appear to have felt that teachers could improve the most on instructional practices.

Figure II.1. Phase 3 FFT component scores by component—all districts



Source: Authors' calculations based on FFT pilot evaluation scores from the 2012–2013 school year provided by PDE.

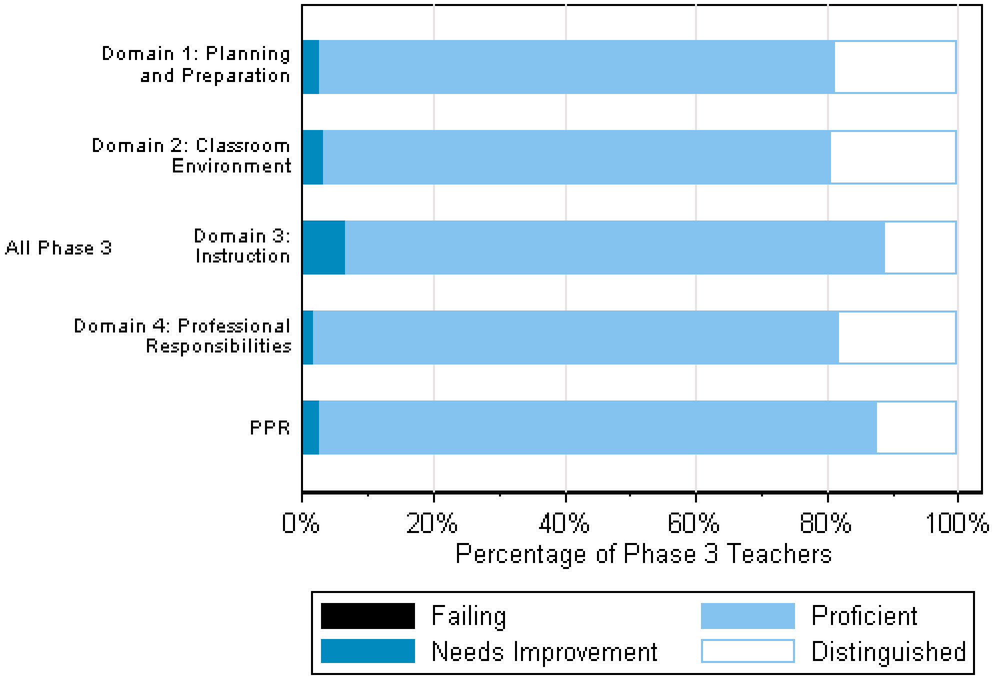
Notes: Bars for *failing* are not visible in the figure, because only very small proportions of teachers received *failing* ratings (between 0.1 and 0.2 percent depending on the component).

See Appendix A, Table A.1 for information on proportions and sample sizes.

Figure II.2 shows that most teachers received *proficient* or *distinguished* domain scores in phase 3, as well (see Appendix A, Table A.2 for a tabular format).[7] At most, 6.7 percent of teachers were rated as *failing* or *needs improvement* in any domain, meaning that at least 93.3 percent were either *proficient* or *distinguished*. Similar to component scores, *proficient* was again the most common rating and *distinguished* the second most common. The proportion of teachers receiving a *proficient* domain score ranged between 77.1 and 82.0 percent. The proportion of teachers receiving a *distinguished* domain score ranged between 11.3 and 19.6 percent. As with component scores, very few teachers received *failing* domain ratings (between 0.0 and 0.1 percent) or *needs improvement* ratings (between 1.7 and 6.6 percent). In short, the range of variation in actual FFT scores is less than the scale permits, because *needs improvement* ratings are used rarely and *failing* ratings almost never.

[7] As mentioned above, we estimated domain scores by averaging the component scores within each domain and treated scores below 0.5 as *failing*, at least 0.5 but less than 1.5 as *needs improvement*, at least 1.5 but less than 2.5 as *proficient*, and at least 2.5 as *distinguished*. We did not have access to teachers' actual domain scores, which are calculated based on the preponderance of evidence within each domain.

Figure II.2. Summary of phase 3 domain and PPR scores—all districts



Source: Authors' calculations based on FFT pilot evaluation scores from the 2012–2013 school year provided by PDE.

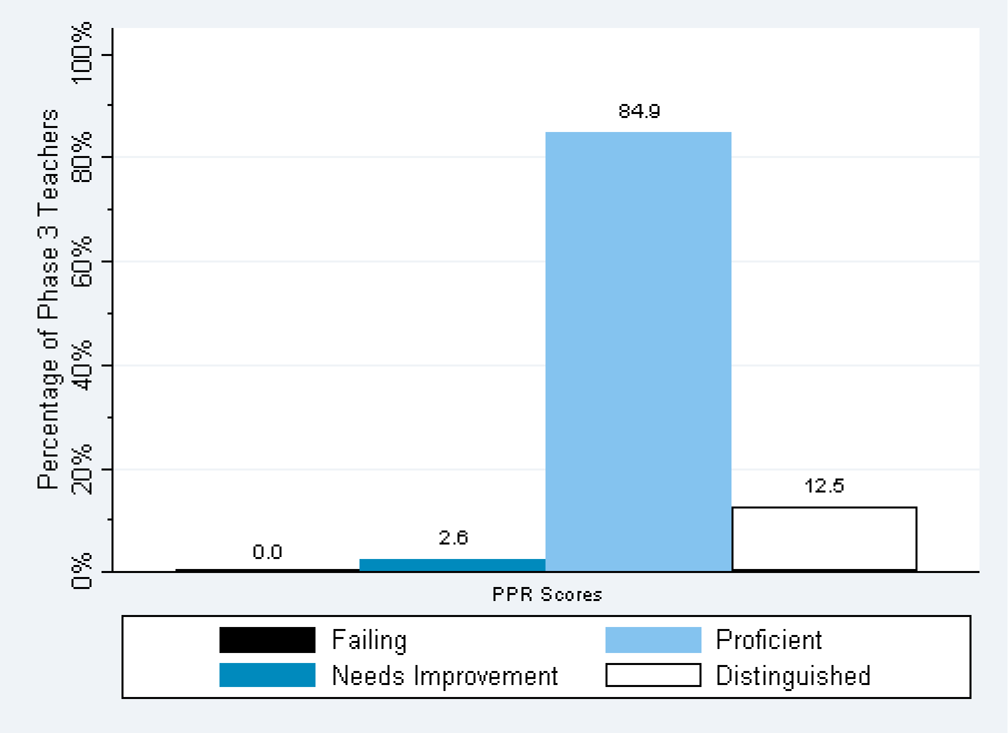
Notes: PPR = Professional Practice Rating. Bars for *failing* are not visible in the figure, because only very small proportions of teachers received *failing* domain or PPR scores (between 0.0 and 0.1 percent).

See Appendix A, Table A.2 for information on proportions and sample sizes.

2. High performance on FFT components and domains led most teachers in phase 3 to receive high Professional Practice Ratings

The PPR is the measure that is ultimately used in teachers' evaluation ratings. Figure II.3 shows the proportion of teachers in phase 3 with PPR scores in each performance category. The PPR scores exhibited the same concentration of scores in the *proficient* and *distinguished* ranges as the component and domain scores. In particular, 12.5 percent of teachers' PPR scores in phase 3 were *distinguished*, 84.9 percent were *proficient*, and 2.6 percent were *needs improvement*. No teacher received a *failing* PPR score in phase 3 (Appendix A, Table A.2). Table A.3 shows the continuous distribution of PPR scores between 0 and 3 and, again, highlights the concentration of PPR scores in the higher end of the score range.

Figure II.3. Distribution of Professional Practice Ratings in phase 3—all districts



Source: Authors' calculations based on FFT pilot evaluation scores from the 2012–2013 school year provided by PDE.

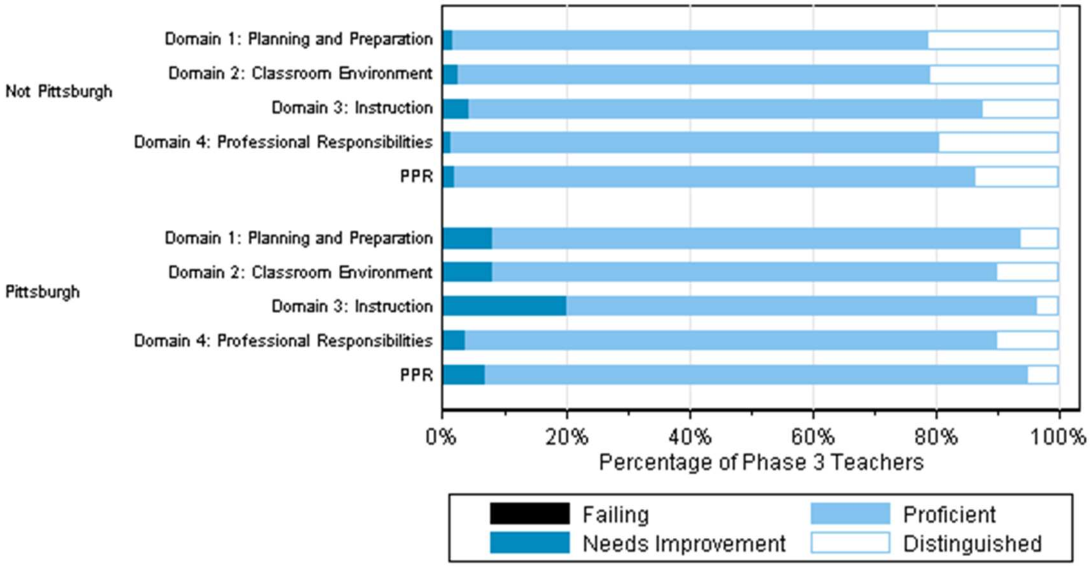
Notes: See Appendix A, Table A.2 for information on sample sizes.

3. Among teachers in phase 3, Pittsburgh teachers received lower professional practice scores, on average, than teachers in other districts

Although most phase 3 teachers across districts received high professional practice scores, Pittsburgh teachers tended to receive lower scores compared with teachers who taught in other districts (Figure II.4; see Appendix A, Table A.4 for tabular format). Depending on the domain, the proportion of phase 3 Pittsburgh teachers who received a *proficient* or *distinguished* domain score was between 2.3 and 15.7 percentage points lower than for phase 3 teachers outside of Pittsburgh. PPR scores were lower, on average, as well. In particular, the proportion of phase 3 Pittsburgh teachers with a PPR in the *proficient* or *distinguished* ranges was 5.3 percentage points lower than for phase 3 teachers outside of Pittsburgh (93.0 versus 98.3 percent).

Pittsburgh teachers' lower domain and PPR scores could indicate lower performance or be a result of their principals having higher evaluation standards. In phase 2, Walsh and Lipscomb (2013) found that Pittsburgh teachers also had slightly lower average VAM scores, suggesting that the evaluation standards of principals in Pittsburgh and in other districts may not have been substantially different. We re-examined this finding in phase 3 and again found that Pittsburgh teachers tended to have slightly lower average VAM scores (see Appendix A, Table A.5).

Figure II.4. Summary of phase 3 domain and PPR scores, by district



Source: Authors' calculations based on FFT pilot evaluation scores from the 2012–2013 school year provided by PDE.

Notes: PPR = Professional Practice Rating. Bars for *failing* are not visible in the figure, because only very small proportions of teachers received *failing* domain or PPR scores (between 0.0 and 0.1 percent).

See Appendix A, Table A.4 for information on proportions and sample sizes.

B. Comparison of Framework for Teaching scores and Professional Practice Ratings obtained during phases 2 and 3

As previously noted, the findings from phase 2 indicated that a large majority of teachers received the two highest ratings on most of the 22 FFT components, possibly indicating that PDE's evaluation system may not be providing substantial variation in scores. Principals may be reluctant to use the two lowest ratings even when appropriate, a possibility that led PDE to provide principals with more rigorous training about how to use the FFT as part of phase 3.

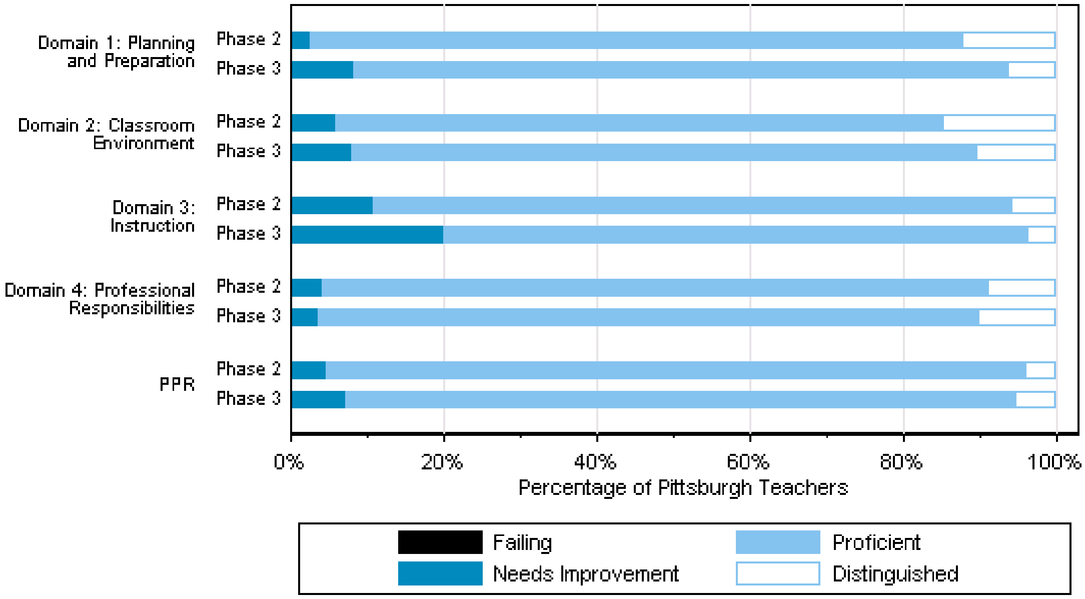
We examined changes in the distribution of FFT and PPR scores between phases 2 and 3 to assess whether principals became more willing to use all four performance categories. Due to the significant difference in the proportion of Pittsburgh teachers between the two phases and the fact that Pittsburgh teachers received lower evaluation scores than non-Pittsburgh teachers in both phases 2 and 3, we examined changes in the distribution of domain and PPR scores separately for Pittsburgh teachers and for teachers outside of Pittsburgh.[8]

[8] Examining the change in the distribution of teachers' evaluation scores between phases 2 and 3 for the entire sample in each phase gives a misleading impression that the proportion of teachers receiving *proficient* or *distinguished* scores was higher in phase 3 than phase 2. This misinterpretation happens because Pittsburgh teachers, who overall have lower scores than other teachers in the pilot, represented a substantially smaller share of the phase 3 sample compared with their share of the phase 2 sample. To adjust for this factor, we estimate results separately for Pittsburgh and non-Pittsburgh teachers.

1. Relative to phase 2, the percentage of teachers given *needs improvement* scores increased in Pittsburgh but not in other districts

For Pittsburgh teachers, we found that the proportion of teachers receiving *needs improvement* ratings rose across most domains and for the PPR, most notably nearly doubling in domain 3 (Figure II.5; see Appendix A, Table A.6 for a tabular format). Correspondingly, the proportion of teachers receiving evaluation scores within either the *proficient* or *distinguished* score ranges was between 2.2 and 9.2 percentage points lower in phase 3 than in phase 2 for domains 1, 2, and 3, was similar for domain 4, and was 2.5 percentage points lower for the PPR. The proportion of teachers receiving *failing* ratings was negligible in both years, ranging from 0.0 to 0.1 percent.

Figure II.5. Comparison of phases 2 and 3 domain scores and Professional Practice Ratings—Pittsburgh teachers only



Source: Authors' calculations based on FFT pilot evaluation scores from the 2011–2012 and 2012–2013 school years provided by PDE.

Notes: PPR = Professional Practice Rating.

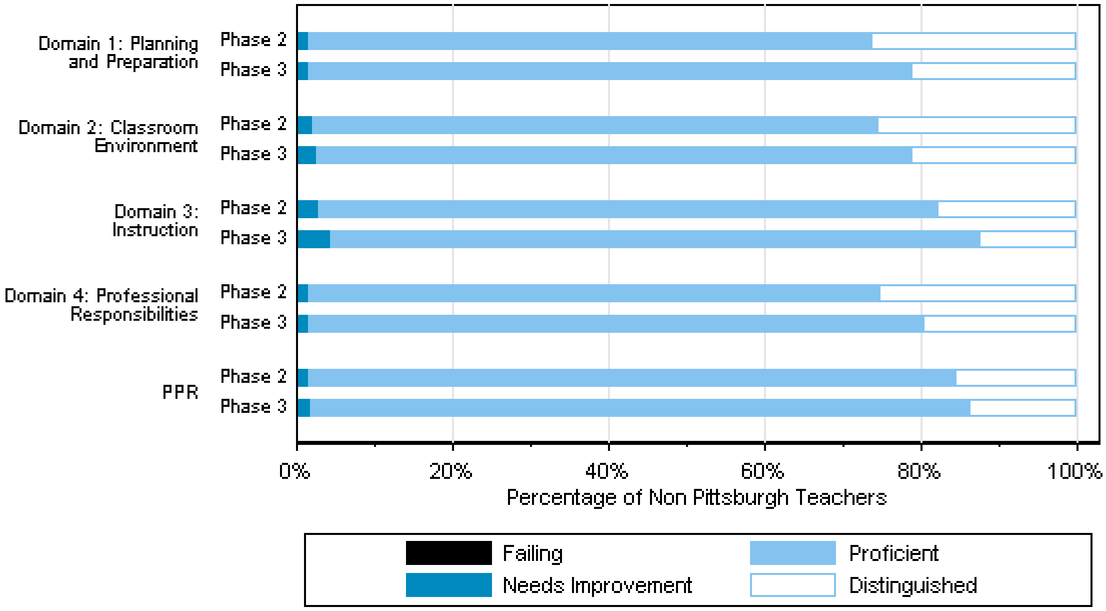
Bars for *failing* are not visible in the figure, because only very small proportions of teachers received *failing* domain or PPR scores (between 0.0 and 0.1 percent).

See Appendix A, Table A.6 for information on proportions and sample sizes.

The findings across phases for teachers in districts other than Pittsburgh were somewhat different (Figure II.6; see Appendix A, Table A.7 for a tabular format). In particular, the proportion of teachers from outside Pittsburgh receiving *proficient* or *distinguished* domain scores was similar between phases 2 and 3 for domains 1, 2, and 4, and was 1.5 percentage points lower in phase 3 for domain 3. However, for all four domains, the proportion of these teachers who were given a *distinguished* domain rating was between 4.4 and 5.6 percentage points lower in phase 3, and the proportion with *proficient* ratings was between 3.8 and 5.8 percentage points higher, indicating that principals outside Pittsburgh may be awarding *distinguished* ratings less frequently and *proficient* ratings more frequently during phase 3.[9] We also found that, unlike the trend in Pittsburgh teachers' PPR, the proportion of teachers from outside Pittsburgh receiving *proficient* or *distinguished* PPR was very similar in phases 2 and 3. The proportion of teachers outside of Pittsburgh receiving *needs improvement* or *failing* domain ratings or PPR scores also remained similar between the two phases.

In sum, ratings in Pittsburgh shifted slightly away from *proficient* and *distinguished* toward *needs improvement*, while ratings outside of Pittsburgh shifted away from *distinguished* toward *proficient*, with minimal change in the low percentages receiving *needs improvement* ratings. Very few teachers received *failing* ratings regardless of the district in which they taught.

Figure II.6. Comparison of phases 2 and 3 domain scores and Professional Practice Ratings—not Pittsburgh teachers only



Source: Authors' calculations based on FFT pilot evaluation scores from the 2011–2012 and 2012–2013 school years provided by PDE.

Notes: PPR = Professional Practice Rating.

Bars for *failing* are not visible in the figure, because only very small proportions of teachers received *failing* domain or PPR scores (between 0.0 and 0.1 percent).

See Appendix A, Table A.7 for information on proportions and sample sizes.

2. Findings were similar if the samples are restricted to teachers participating in both phase 2 and phase 3

As previously noted, we analyzed teachers in Pittsburgh separately from those outside of Pittsburgh to account for large changes across phases in the proportion of the teacher sample

[9] The differences in the proportions of teachers receiving *distinguished* and *proficient* ratings were calculated by subtracting the percentage of phase 2 teachers receiving that rating from the percentage of phase 3 teachers also receiving that rating (from Table A.7).

coming from Pittsburgh. However, the teacher samples changed between phases within each of these two groups of teachers, as well, changes that could explain differences in FFT scores. In particular, teachers in Pittsburgh's Student Growth Project (SGP), who have previously demonstrated proficiency on the FFT, were included in the data from phase 2 but not in phase 3. The exclusion of this group of high-performing teachers could explain the apparent increase in the proportion of ratings in the *needs improvement* category in phase 3.

To account for any compositional effects, we repeated our analyses only for teachers who participated in both phases. By focusing on teachers who were rated in both phases, any differences in average evaluation scores between phases represent either changes in their performance or a systemic change in the way they were evaluated. The findings for this subset of teachers, shown in Appendix A, Tables A.8 and A.9, are quite similar to those presented in Figures II.5 and II.6 for both Pittsburgh teachers and teachers outside of Pittsburgh. Collectively, the findings suggest that changes in the composition of both groups may not have substantially impacted changes in evaluation scores between the two phases for either group.

III. THE FFT HAD GOOD OR ACCEPTABLE INTERNAL CONSISTENCY IN PHASE 3, AS IN PHASE 2

Information about the distribution of teachers' classroom observation scores is useful only if the framework used to determine the scores is reliable. There are several ways to measure this reliability. However, because PDE's teacher evaluation system does not involve multiple observers rating each teacher, and each teacher is rated only once, we could examine only one dimension of reliability—the internal consistency of the evaluation scores. Internal consistency assesses the similarity of teachers' FFT scores on measures designed to capture similar aspects of their performance. If a teacher's observation scores vary substantially across measures that pertain to the same underlying concept, then the observation system may not be reliably measuring that teacher's performance.

1. The full Framework for Teaching had good internal consistency and its domains had at least acceptable internal consistency in phase 3

Using phase 3 data, we computed for each domain and for the PPR Cronbach's alpha (Cronbach 1951), a commonly used measure of internal consistency. We applied David de Vaus' recommendation (from his widely cited textbook on surveys in social research) that Cronbach's alpha values above 0.8 are considered good and alpha values above 0.7 are considered acceptable (de Vaus 2002). We also estimated the contributions of each component and domain to the internal consistency of domain scores and PPR to determine how sensitive the results are to omitting a particular score.

We found the internal consistency of the FFT was acceptable or good during phase 3. Table III.1 shows that Cronbach's alpha for domains 2, 3, and 4 fell within the acceptable range, and that Cronbach's alpha for domain 1 met the criterion for a good rating. Table III.1 also shows that Cronbach's alpha for the PPR, 0.87, is higher and within the good range. Tables A.10 and A.11 present the "leave-out" scores calculated to assess the contribution of each component and domain of the FFT. The leave-out alphas for each component and domain do not indicate that any single component is inconsistent with the other components within a domain, or that any domain is inconsistent with the other domains.

Table III.1. Phase 3 Cronbach's alpha values for Framework for Teaching domains and Professional Practice Rating scores

Framework for Teaching domain	Number of items in scale	Cronbach's alpha	Sample size
Domain 1: planning and preparation	6	0.80	6,422
Domain 2: classroom environment	5	0.76	6,401
Domain 3: instruction	5	0.77	6,373
Domain 4: professional responsibilities	6	0.77	5,975
PPR (Professional Practice Rating)	4	0.87	6,675

Source: Authors' calculations based on FFT pilot evaluation scores from the 2012–13 school year provided by PDE.

Note: Sample sizes may vary because only teachers with ratings for all components within a domain or all domains of the PPR are included in the calculation of Cronbach's alpha.

2. Internal consistency rose slightly between phases 2 and 3

We examined changes in the FFT's internal consistency between phases 2 and 3 to investigate whether the FFT continues to estimate teachers' component, domain, and PPR scores reliably. Table III.2 indicates that the internal consistency of the rubric increased slightly in phase 3, consistent with evaluators having a better understanding of the FFT and how its components interrelate.

Table III.2. Cronbach's alpha values for Framework for Teaching domain scores and Professional Practice Ratings

		Phase 2		Phase 3	
FFT domain	Number of items in scale	Cronbach's alpha	Sample size	Cronbach's alpha	Sample size
Domain 1: planning and preparation	6	0.78	1,659	0.80	6,422
Domain 2: classroom environment	5	0.75	1,639	0.76	6,401
Domain 3: instruction	5	0.72	1,646	0.77	6,373
Domain 4: professional responsibilities	6	0.75	1,440	0.77	5,975
PPR (Professional Practice Rating)	4	0.84	2,487	0.87	6,675

Source: Authors' calculations based on FFT pilot evaluation scores from the 2011–12 and 2012–13 school years provided by PDE.

Note: Sample sizes may vary because only teachers with ratings for all components within a domain or all domains of the PPR are included in the calculation of Cronbach's alpha.

IV. TEACHERS' FRAMEWORK FOR TEACHING SCORES WERE POSITIVELY CORRELATED WITH THEIR CONTRIBUTIONS TO GROWTH IN STUDENT ACHIEVEMENT, AS IN PHASE 2

The ultimate purpose of improving the evaluation of teacher effectiveness in Pennsylvania is to help students achieve (see Figure I.1). For the Framework for Teaching (FFT) to be a useful tool for improving student learning, teachers who exhibit the practices assessed by the FFT must be the same teachers who are more effective at raising student achievement. Measuring the strength of the relationship between teachers' FFT scores and their contributions to student achievement growth is a way to validate the core rationale underlying Pennsylvania's drive to improve the evaluation of teacher effectiveness.

Walsh and Lipscomb (2013) found that Pennsylvania teachers with higher FFT scores in the phase 2 pilot tended to make larger contributions to student achievement growth. They measured contributions to student achievement growth using a value-added model (VAM) that predicted students' Pennsylvania System of School Assessment (PSSA) outcomes (all subjects) in grades 4 through 8 based on students' own prior achievement scores and background characteristics. VAMs give an effectiveness score to each teacher based on the extent to which students' actual assessment outcomes exceed (or fall short of) their predicted outcomes, where the prediction represents how well the students would have done if served by the average teacher. This effectiveness score, called a VAM score or a value-added estimate, is a measure of teachers' contributions to their students' achievement growth.

The estimated correlations between teachers' FFT and VAM scores during the phase 2 pilot were positive and statistically significant across most components and at the domain and PPR levels. However, the magnitudes of the correlations were small, consistent with prior research on the FFT and other professional practice measures for teachers (Kane and Staiger 2012). Small correlations between FFT and VAM scores can occur for several reasons, including the following:

- Teacher performance over an entire school year may differ from their performance on the days when principals are able to observe them.
- The FFT might describe teaching practices that are strongly tied to growth in academic or non-academic outcomes that are not indicated well by PSSA scores in grades 4 through 8.
- Principals might not be applying the FFT correctly in evaluating teachers, thereby obscuring the true closeness of the relationship between FFT and VAM scores.

At the request of the Pennsylvania Department of Education (PDE), we re-examined the relationships between FFT and VAM scores to see how accurately findings from phase 2 represent the correlations between FFT and VAM scores in the broader phase 3 sample. Re-examining these relationships using the phase 3 data is opportune given PDE's efforts to improve evaluator training, which may ameliorate concerns about whether principals' are applying the FFT as intended.

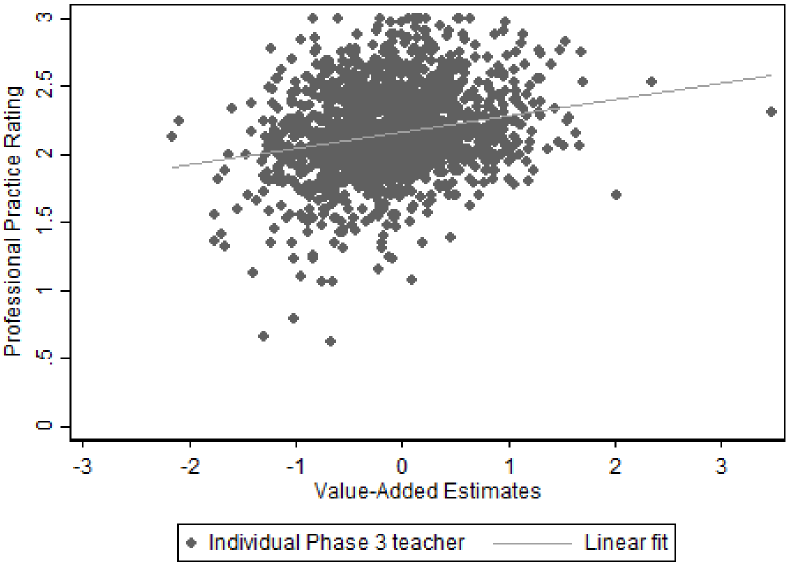
Below, we correlate phase 3 teachers' FFT scores at the component, domain, and PPR levels with their VAM scores. We then re-examine the strength of these correlations for subgroups of teachers by grade range and subject, and compare these findings with those from phase 2. We used the same methods as in phase 2 to estimate teacher VAM scores (see Appendix B for more detail). The VAMs include school years 2010–2011 through 2012–2013 and compare the effectiveness of all teachers in grades 4 through 8 across the entire state during that period, using PSSA scores as the achievement measure, mirroring PDE's plans for actual teacher evaluations.[10] For teachers of multiple subjects and/or grades, we combined their estimates to a single overall VAM score.

A. Teachers with higher component, domain, and PPR scores were more likely to have higher value-added scores in phase 3, similar to phase 2

Phase 3 teachers with higher FFT scores tended to make larger contributions to student achievement growth as measured by PSSA outcomes in grades 4 through 8, than those with lower FFT scores. This finding was true across components and domains, and for the PPR. Figure IV.1 plots PPR and VAM scores for all 1,730 phase 3 teachers who taught students in at least one tested subject in grades 4 through 8 (25 percent of the phase 3 sample) to illustrate the positive relationship between the two measures of effectiveness. The degree of correlation is 0.24 on a scale between -1 and 1, where positive values indicate that higher FFT scores are associated with higher VAM scores. The degree of correlation would be higher were teachers bunched more closely to the positively sloped line in the figure; the correlation would be closer to zero if teachers were more scattered across the chart.

[10] An alternative approach would be to estimate a VAM that is based entirely on the 2012–13 school year. Walsh and Lipscomb (2013) provide some evidence that scores from this single- and same-year VAM are slightly more related to FFT scores. However, these findings may overestimate the true relationship between FFT and VAM scores, because the increase in correlation may partly reflect the influence of an unusually high- or low-achieving group of students that affects VAM scores and evaluators' impressions of teachers' professional practices in the same way.

Figure IV.1. The Professional Practice Ratings and value-added estimates of phase 3 teachers in grades 4 through 8



Note: The correlation coefficient between PPR scores and value-added estimates is 0.24.

Source: Mathematica calculations based on phase 3 classroom observation data in the 2012–2013 school year and value-added estimates using data from school years 2010–2011 through 2012–2013.

The correlations between teachers' FFT and VAM scores in phase 3 were in the range of those found in phase 2 (Walsh and Lipscomb 2013) and are consistent with findings from the Measures of Effective Teaching project (Kane and Staiger 2012).[11] At the PPR and domain levels, the correlations ranged from 0.19 to 0.24 and were statistically significant (column 1 of Table IV.1). That is, we can say with confidence that the correlations are positive. At the domain level, the largest correlation with value added was in domain 1—planning and preparation (0.22). The second highest was in domain 3—instruction (0.21). In phase 2, domain 3 had the highest correlation (0.28) of any domain-level score with value added (column 2 of Table IV.1).

The correlations between VAM scores and individual FFT components for the full sample were all positive and statistically significant in phase 3, ranging from 0.11 (demonstrating flexibility and responsiveness) to 0.20 (demonstrating knowledge of content and pedagogy; managing classroom procedures). In phase 2, some estimated correlations were not statistically significant, perhaps in part due to smaller samples. The magnitude of the correlations in phase 3 were similar to the correlations found in phase 2 for most components. However, the correlations

[11] The findings pertain to correlations between FFT scores and *underlying value added*—the value added measure we would obtain if we could eliminate estimation error. Imprecision in value-added estimates tends to lower correlations with professional practice scores. We sought to eliminate this estimation error to focus on the portion of value-added scores that is a signal rather than noise. To achieve this objective, we followed the well-known approach described in Jacob and Lefgren (2008) of adjusting the correlations by the inverse of the square root of the reliability of the value-added estimates, calculated using the estimated standard errors of the value-added estimates. Presumably, the correlations would be even larger if we could also adjust for the error in the FFT scores.

for the components in domain 3 (instruction), while on par with those in other domains, were not the largest, as they had been in phase 2.

Table IV.1. Correlations between 4th through 8th grade teachers' Framework for Teaching scores and their value-added model scores, by phase

	Full sample		Pittsburgh		Not Pittsburgh	
FFT measure	Phase 3	Phase 2	Phase 3	Phase 2	Phase 3	Phase 2
Professional practice rating (PPR)	0.24*	0.24*	0.27*	0.22*	0.22*	0.22*
Domain 1: planning and preparation	0.22*	0.23*	0.29*	0.20*	0.20*	0.21*
Domain 2: classroom environment	0.20*	0.19*	0.21*	0.18*	0.19*	0.16*
Domain 3: instruction	0.21*	0.28*	0.24*	0.27*	0.20*	0.24*
Domain 4: professional responsibilities	0.19*	0.17*	0.21*	0.16*	0.19*	0.11
1a: demonstrating knowledge of content and pedagogy	0.20*	0.12*	0.29*	0.06	0.17*	0.17*
1b: demonstrating knowledge of students	0.12*	0.19*	0.19*	0.18*	0.10*	0.18*
1c: setting instructional outcomes	0.19*	0.14*	0.22*	0.10	0.18*	0.19*
1d: demonstrating knowledge of resources	0.15*	0.17*	0.14*	0.15*	0.14*	0.15
1e: planning coherent instruction	0.13*	0.18*	0.16*	0.19*	0.11*	0.13*
1f: designing ongoing formative assessments	0.14*	0.16*	0.21*	0.16*	0.12*	0.10
2a: creating a learning environment of respect and rapport	0.12*	0.14*	0.12	0.16*	0.12*	0.09
2b: establishing a culture for learning	0.16*	0.20*	0.27*	0.18*	0.12*	0.21*
2c: managing classroom procedures	0.20*	0.18*	0.11	0.24*	0.21*	0.10
2d: managing student behavior	0.12*	0.16*	0.09	0.17*	0.12*	0.12
2e: organizing physical space	0.12*	0.04	0.18*	0.01	0.10*	0.04
3a: communicating with students	0.15*	0.25*	0.12	0.28*	0.14*	0.20*
3b: using questioning and discussion techniques	0.16*	0.24*	0.23*	0.24*	0.14*	0.21*
3c: engaging students in learning	0.18*	0.22*	0.22*	0.21*	0.16*	0.19*
3d: using assessment to inform instruction	0.15*	0.17*	0.11	0.19*	0.15*	0.08
3e: demonstrating flexibility and responsiveness	0.11*	0.18*	0.18*	0.17*	0.09*	0.17*
4a: reflecting on teaching and student learning	0.14*	0.13*	0.21*	0.14*	0.13*	0.08
4b: system for managing students' data	0.13*	0.09*	0.20*	0.06	0.11*	0.09
4c: communicating with families	0.13*	0.12*	0.14*	0.15*	0.12*	0.00
4d: participating in a professional community	0.12*	0.13*	0.12	0.07	0.11*	0.15*
4e: growing and developing professionally	0.14*	0.12*	0.17*	0.09	0.13*	0.11
4f: showing professionalism	0.12*	0.10	0.10	0.03	0.12*	0.12
Sample sizes (for PPR and domain scores)	1730	666	265	395	1465	271

Sources: Mathematica calculations based on phase 3 classroom observation data paired with value-added estimates from school years 2010–11 through 2012–13. Findings for phase 2 are reproduced from Walsh and Lipscomb (2013).

Notes: Correlations are adjusted for estimation error in value-added estimates (Jacob and Lefgren 2008). The sample sizes for components are lower than the sample size reported for the PPR and domain scores in the bottom row when teachers are not rated on particular components.

* Statistically significant at the 5 percent level.

The correlations in phase 3 were larger for Pittsburgh teachers than for non-Pittsburgh teachers (also true in phase 2, albeit less consistently across components). This finding was true for the PPR, the domain-level scores, and most components. Several factors could be contributing to this pattern, although two in particular may be likely. First, Pittsburgh Public Schools has been using a version of the FFT in teacher evaluations for several years, meaning that principals in that district may have more experience applying it than other principals evaluating teachers in the pilot. Second, principals in Pittsburgh have access to their teachers' value-added scores from the prior year, and could be using them to inform their judgments about teacher performance in subsequent years.

B. The positive correlations were systemic across teachers in different grades and subjects

We explored the relationships between teachers' FFT and VAM scores for groups of phase 3 teachers based on the grade levels and subjects in which they teach students, to provide additional context for the overall findings. Specifically, we focused on four groups. The first group includes teachers in grades 4 through 6 who are responsible for teaching more than one subject, referred to as generalist elementary teachers. The remaining three groups include departmentalized math, English-language arts (ELA), and science teachers, respectively, in grades 6 through 8. We did not include in this analysis 7th and 8th grade teachers who taught students in multiple subjects.

The findings, shown in Table IV.2 for the PPR and domain-level scores, indicate that higher FFT scores are positively correlated with higher VAM scores across these four teacher groups. The findings for generalist elementary teachers in phase 3 are most consistent with the overall findings in Table IV.1. The correlations with value added for this group of teachers are somewhat larger than in phase 2. In addition, all of the correlations at the PPR and domain levels are statistically significant, which was not true in phase 2, perhaps due in part to smaller samples. For departmentalized math and ELA teachers, the magnitudes of correlations between FFT and VAM scores were somewhat lower than in phase 2, except for domain 4. This finding was particularly prevalent among math teachers in domain 3 (instruction), where the correlation is not statistically significant, perhaps partly explaining the smaller magnitude of the correlation with domain 3 scores in the overall sample. Nevertheless, most of the correlations across domains are positive and statistically significant for teachers in both subjects. As in phase 2, the correlations for departmentalized science teachers were larger than in other subjects or grades, possibly meaning that the practices included on the FFT are stronger predictors of contributions to student achievement growth in science than in other subjects.[12] Regardless of the explanation, the consistently positive relationships between FFT and VAM scores for teachers across these grades and subjects suggest that higher FFT scores might be predictive of larger contributions to student learning in nontested grades and subjects, as well.

[12] The findings for science are based on 8th grade teachers only, because the science PSSA is given only to middle school students in 8th grade. The larger correlations in science compared with math and ELA do not appear to be related to grade level, however. In particular, the findings for math and ELA are consistent when the teacher sample is restricted to just those teaching students in 8th grade. However, fewer estimates are statistically significant, because the sample sizes are smaller.

Table IV.2. Correlations between teachers' value-added scores and their Framework for Teaching domain scores and Professional Practice Ratings, by grade span and subject—all districts

	Grades 4–6 generalist elementary teachers		Grades 6–8 departmentalized teachers					
			Math		ELA		Science	
	Phase 3	Phase 2	Phase 3	Phase 2	Phase 3	Phase 2	Phase 3	Phase 2
Professional practice rating	0.24*	0.17	0.12*	0.22*	0.17*	0.16	0.30*	0.46*
Domain 1: planning and preparation	0.23*	0.08	0.13*	0.28*	0.16*	0.24*	0.20	0.21
Domain 2: classroom environment	0.23*	0.15	0.07	0.13	0.12*	0.21*	0.29*	0.32*
Domain 3: instruction	0.21*	0.17*	0.08	0.29*	0.16*	0.20*	0.26*	0.56*
Domain 4: professional responsibilities	0.18*	0.18*	0.18*	−0.03	0.13*	0.03	0.32*	0.36*
Sample sizes	662	134	308	121	412	172	85	48

Sources: Mathematica calculations based on phase 3 classroom observation data paired with value-added estimates from school years 2010–11 through 2012–13. Findings for phase 2 are reproduced from Walsh and Lipscomb (2013).

Notes: Correlations are adjusted for estimation error in value-added estimates (Jacob and Lefgren 2008). Total sample sizes in each table row are smaller than in Table IV.1, because teachers in grades 7 and 8 are excluded if they teach multiple subjects.

* Statistically significant at the 5 percent level.

V. CONCLUSION

One key goal of a teacher evaluation system is to distinguish between higher- and lower-performing teachers. To achieve this objective, an evaluation system must have the capacity to give teachers different evaluation scores. An evaluation system that assigns similar scores to teachers whose effectiveness differs substantially will limit the types of decisions that can be made using that system. When studying data from the 2011–2012 school year (phase 2 of the pilot of this teacher evaluation system), Walsh and Lipscomb (2013) found that on most components, at least 90 percent of teachers received either a *distinguished* or *proficient* rating. Our analyses of the phase 3 data from the 2012–2013 school year shows a similar pattern: more than 90 percent were rated in the top two categories (*proficient* or *distinguished*) for most components as well as for the overall PPR score and for each domain. Less than 0.3 percent received scores in the bottom category (*failing*). The rest (between 2 and 20 percent) were given the second lowest rating: *needs improvement*. The fraction scoring in the top two categories decreased somewhat in Pittsburgh, especially in domain 3 (instruction), but not in the other pilot districts.

Although we do not know what the ideal distribution of FFT ratings should be, our findings suggest that the FFT, as implemented in both phases, differentiates teacher performance only to a limited degree. One possible contributor might have been that some principals did not apply the FFT as it was intended to be used and were reluctant to use the two lowest ratings. This concern led PDE to provide more rigorous training to principals in phase 3 on how to use the FFT. The phase 3 training included opportunities for principals to compare their ratings on practice evaluations against official ratings by the FFT developer. The lack of noticeable shifts in the ratings between phases 2 and 3 (outside of Pittsburgh) may suggest the need for continued training and/or monitoring of the system.

Although the data could be taken as cause for concern, we also found some encouraging patterns. In many evaluation systems, including the one that existed in Pennsylvania before the recent reforms, almost all teachers were given the highest possible score. The fact that less than 20 percent of teachers received the highest possible score for any domain and only 12.5 percent for the overall PPR score in our study indicates a substantial change since the pre-reform period. At the other end of the scale, it may be concerning that less than 0.3 percent of teachers received the lowest possible score, however it should be noted that teachers can be fired for receiving a *failing* score for their final evaluation.[13] Similarly, the fraction receiving *needs improvement* scores might also be seen as cause for concern, because less than 3 percent of teachers received an overall PPR score in that range—but again, it should be noted that a teacher who within a decade receives two final evaluation scores in the *needs improvement* range can be let go.

We also found that the FFT rubric measures were internally consistent, with the overall PPR score having higher consistency than any of the domain scores in both phases. The internal

[13] http://www.pacode.com/secure/data/022/chapter19/s19.1.html. Decisions about teacher tenure are made based on the final evaluation score, which combines the supervisor ratings, student growth measures, and other data. These final evaluation scores use a four-point scale with categories of failing, needs improvement, proficient, and distinguished.

consistency of the overall domain-level scores ranged from 0.76 to 0.80 in Phase 3 as compared with 0.72 to 0.78 in Phase 2. Thus, in both phases they were in the acceptable range and improved slightly between the two phases. These levels of internal consistency suggest that the different components of the FFT are measuring similar or highly correlated attributes of teacher quality (within and across domains).

In phase 3, as in phase 2, teachers with higher FFT scores tended to make modestly larger contributions to student achievement growth compared to teachers with lower FFT scores. The correlations of the FFT domain scores with VAM for the phase 3 sample overall were consistently statistically significant, ranging from 0.19 to 0.22. For Pittsburgh teachers, the correlations of the FFT domain scores with VAM were higher in phase 3 than in phase 2 in all domains except for domain 3 (instruction), which declined slightly. The correlation of FFT domain 3 scores with VAM also declined slightly for teachers outside of Pittsburgh. The correlation of PPR scores with VAM rose in Pittsburgh (from 0.22 to 0.27) and remained the same outside of Pittsburgh (0.22). The magnitudes of these correlations with VAM may have been somewhat higher had we been able to adjust for measurement error in the FFT scores and not just in VAM. The fact that the correlations are well below 0.5 may suggest that the FFT measures are somewhat noisy and/or that FFT scores may be capturing aspects of teacher quality that are not captured by VAM.

Although some schools in the state of Pennsylvania have been applying the FFT for a number of years, others are just starting to learn how the process works. All findings to date are also based on data from no-stakes measures. Both theory and prior evidence on testing regimes suggest that attaching stakes to the measures (that is, including FFT scores in formal evaluations, as PDE is doing) will inflate the scores—and could undermine their correlation to value added, as well. Hence, it will be important to make sure that principals are applying the framework faithfully going forward. State and district staff may want to consider taking steps to confirm that the ratings are correct (for example, through inter-rater reliability checks and by continuing to triangulate the results with other sources of information, such as student surveys and VAM results). Similarly, although the system allows for four possible ratings, most teachers receive only two of those four ratings. Hence, it will be important to continue to monitor the distribution of scores to ensure that they reflect policy priorities and goals.

In total, these results suggest that, in comparison to the pre-reform situation, the state of Pennsylvania has made important strides toward improving its teacher evaluation system by bringing in more sources of information (that is, VAM), by increasing the amount of differentiation in supervisor ratings, and by implementing a measure of professional practice that is more clearly related to student achievement growth than the previous measure. We also find evidence that the FFT and VAM are complementary measures of teacher quality. However it remains likely that the FFT could be improved to further enhance its ability to improve education outcomes in the state of Pennsylvania. For example, adding additional classroom observations and employing multiple raters, raters that work across schools, and independent raters (rather than colleagues of the staff they are rating) are steps that all have the potential to further improve the validity and reliability of FFT scores.

Additional research might also help to inform future improvements in the FFT and the overall evaluation system. Some of this research could be done with current data. For example, understanding how teacher characteristics are associated with both FFT and VAM might inform how PDE develops and recruits staff. Similarly, understanding whether FFT scores vary across years among teachers, depending on the characteristics of their students, might suggest a need to adjust scores based on those student characteristics. Obtaining answers to other questions might also be useful but would require additional data. For example, the scores that principals give teachers may change after principals receive certain types of training and/or receive information on teachers' VAM scores. We would be glad to explore these topics and others with Team PA and PDE.

This page has been left blank for double-sided copying.

VI. REFERENCES

Buonaccorsi, J.P. *Measurement Error: Models, Methods, and Applications*. Boca Raton, FL: Chapman & Hall/CRC, 2010.

Cronbach, Lee J. “Coefficient Alpha and the Internal Structure of Tests.” *Psychometrika*, vol. 16, no. 3, 1951, pp. 297–334.

de Vaus, David A. *Surveys in Social Research*, 5th edition. Crows Nest, Australia: Allen & Unwin, 2002.

Hock, H., and E. Isenberg. “Methods for Accounting for Co-Teaching in Value-Added Models.” Working paper. Washington, DC: Mathematica Policy Research, 2012.

Jacob, B.A., and L. Lefgren, “Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education.” *Journal of Labor Economics*, vol. 26, no. 1, 2008, pp. 101–136.

Kane, T.J., and D.O. Staiger. “Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.” Seattle, WA: Bill & Melinda Gates Foundation, Measures of Effective Teaching Project, 2012.

Lipscomb, S., H. Chiang, and B. Gill. “Value-Added Estimates for Phase 1 of the Pennsylvania Teacher and Principal Evaluation Pilot.” Cambridge, MA: Mathematica Policy Research, 2012.

Lipscomb, S., B. Teh, B. Gill, H. Chiang, and A. Owens “Teacher and Principal Value Added: Research Findings and Implementation Practices.” Cambridge, MA: Mathematica Policy Research, 2010.

Milanowski, A. “Validity Research on Teacher Evaluation Systems Based on the Framework for Teaching.” March 18, 2011. Paper presented at the American Educational Research Association Annual Meeting, New Orleans, LA, April 10, 2011. Available at http://www.eric.ed.gov/PDFS/ED520519.pdf. Accessed December 18, 2014.

Walsh, E., and S. Lipscomb. “Classroom Observations from Phase 2 of the Pennsylvania Teacher Evaluation Pilot: Assessing Internal Consistency, Score Variation, and Relationships with Value Added.” Cambridge, MA: Mathematica Policy Research, 2013.

This page has been left blank for double-sided copying.

APPENDIX A

DATA SOURCES AND DESCRIPTIVE INFORMATION

This page has been left blank for double-sided copying.

1. Data sources

We rely on two types of data to address the study's research questions. The first is phase 3 Framework for Teaching (FFT) classroom observation data, which are used in analyses to address all three research questions. The second is statewide, student-level longitudinal data, which are used in analyses to address only the third research question.

A. Framework for Teaching classroom observation scores

The phase 3 FFT data include classroom observation scores on 6,676 participating teachers from the 2012–2013 school year. Of these teachers, 1,038 teach in Pittsburgh Public Schools (PPS) and 5,638 teach in 268 of the approximately 500 school districts in Pennsylvania. The data include principals' ratings of teachers on the 22 FFT components but not domain-level or Professional Practice Rating (PPR) scores.[14] We average teachers' component scores within each domain to calculate domain-level scores, using data from any components that were rated within the domains. We calculate PPR scores as a weighted average of teachers' domain scores, weighting domains 2 and 3 at 30 percent each and domains 1 and 4 at 20 percent each. We exclude from the PPR calculation teachers missing at least one domain score (meaning that they are missing scores for all components in a domain).

Findings in Sections II and III are based on FFT scores from the entire phase 3 sample, but the findings in Section IV are based only on phase 3 teachers who teach math, reading, science, and/or writing to students in grades 4 through 8. This subset includes 1,730 teachers, or 26 percent of the phase 3 sample.

B. Statewide, student-level longitudinal data

We estimate teachers' contributions to their students' achievement growth using student-level longitudinal data from two agencies within the Pennsylvania Department of Education (PDE). Test score data come from the Bureau of Assessment and Accountability and include all Pennsylvania System of School Assessment (PSSA) scores for students in grades 3 through 8 in math, reading, science, and writing during school years 2009–2010 through 2012–2013. Data on student-level characteristics, course records, and teacher links are derived from the Pennsylvania Information Management System (PIMS). The PIMS data include school years 2010–2011 through 2012–2013. We used PPS data to link Pittsburgh students with their teachers, because PIMS records for students in that district were incomplete.

We use these data in teacher value-added models (VAMs) that estimate the size of teachers' contributions to student achievement growth. The VAMs cover school years 2010–2011 through 2012–2013. The test score data extend back one additional year, to 2009–2010, so that students' prior-year scores can be included in the VAMs. See Appendix B for more detail.

[14] Principals were not asked to provide the domain or PPR scores they assigned, only component scores, as part of the Pennsylvania Teacher Evaluation Pilot.

2. Distribution of phase 3 Framework for Teaching scores for teachers in Pittsburgh and in other districts

Tables A.1 and A.2 show the phase 3 distribution of FFT scores across performance categories for all teachers in the sample. Table A.3 presents the distribution of all phase 3 teachers' PPR scores across eight score ranges, each spanning 0.5 FFT points. Table A.4 compares domain and PPR scores for phase 3 teachers in Pittsburgh and in other districts. Table A.5 compares the average FFT and VAM scores separately for Pittsburgh and non-Pittsburgh teachers. Tables A.6 and A.7 compare the distribution of teachers' domain scores and PPR scores between phases 2 and 3 for all PPS teachers and for all teachers in other districts who were rated in at least one of the two phases.

Table A.1. Summary of phase 3 classroom observation data—all districts

		Percentage of teachers earning:				
Rubric component	Number of teachers with scores	*Failing*	*Needs improvement*	*Proficient*	*Distinguished*	Average FFT score
1a: demonstrating knowledge of content and pedagogy	6,544	0.1	4.2	70.8	25.0	2.2
1b: demonstrating knowledge of students	6,585	0.2	6.1	69.1	24.6	2.2
1c setting instructional outcomes	6,590	0.2	6.6	77.2	16.1	2.1
1d: demonstrating knowledge of resources	6,500	0.1	5.3	69.6	25.0	2.2
1e: planning coherent instruction	6,595	0.2	5.8	75.0	19.1	2.1
1f: designing ongoing formative assessments	6,523	0.2	9.0	79.7	11.1	2.0
2a: creating a learning environment of respect and rapport	6,558	0.1	4.1	60.7	35.0	2.3
2b: establishing a culture for Learning	6,606	0.2	5.8	73.2	20.8	2.1
2c: managing classroom procedures	6,550	0.2	6.8	69.0	24.0	2.2
2d: managing student behavior	6,551	0.2	7.5	70.9	21.4	2.1
2e: organizing physical space	6,458	0.0	2.8	75.9	21.3	2.2
3a: communicating with students	6,553	0.2	5.6	66.1	28.2	2.2
3b: using questioning and discussion techniques	6,566	0.2	19.0	71.2	9.5	1.9
3c: engaging students in learning	6,629	0.2	11.3	70.4	18.1	2.1
3d: using assessment to inform instruction	6,578	0.2	13.5	76.7	9.7	2.0
3e: demonstrating flexibility and responsiveness	6,447	0.2	4.7	75.3	19.8	2.1
4a: reflecting on teaching and student learning	6,435	0.2	5.6	75.8	18.5	2.1
4b: system for managing student data	6,421	0.2	6.3	79.1	14.4	2.1
4c: communicating with families	6,255	0.2	9.8	74.0	16.0	2.1
4d: participating in a professional community	6,354	0.1	5.4	70.5	24.0	2.2
4e: growing and developing professionally	6,345	0.1	4.1	78.3	17.6	2.1
4f: showing professionalism (PPS Data)	6,372	0.2	1.9	70.0	27.9	2.3
All components	6,676	0.2	6.9	72.7	20.3	2.1

Source: Authors' calculations based on Framework for Teaching pilot evaluation scores from the 2012–2013 school year provided by the Pennsylvania Department of Education.

Table A.2. Summary of phase 3 classroom observation data—all districts

Domain	Number of teachers with scores	Percentage of teachers earning: *Failing*	*Needs improvement*	*Proficient*	*Distinguished*	Average FFT score	Standard deviation
Domain 1	6,675	0.0	2.6	78.3	19.1	2.1	0.4
Domain 2	6,676	0.0	3.3	77.1	19.6	2.2	0.4
Domain 3	6,676	0.1	6.6	82.0	11.3	2.1	0.4
Domain 4	6,676	0.1	1.7	79.9	18.4	2.1	0.3
PPR	6,675	0.0	2.6	84.9	12.5	2.1	0.3

Source: Authors' calculations based on Framework for Teaching pilot evaluation scores from the 2012–2013 school year provided by the Pennsylvania Department of Education.

Note: PPR = Professional Practice Rating

Table A.3. Distribution of Professional Practice Ratings—all districts

	Full PPR
Mean score	2.1
Standard deviation of scores	0.3
Percentage of scores that are below 0.5	0.0
Percentage of scores that are at least 0.5, below 1.0	0.3
Percentage of scores that are at least 1.0, below 1.5	2.2
Percentage of scores that are at least 1.5, below 2.0	23.8
Percentage of scores that are exactly 2.0	13.5
Percentage of scores that are above 2.0, below 2.5	47.6
Percentage of scores that are at least 2.5, below 3.0	11.7
Percentage of scores that are exactly 3.0	0.9

Source: Authors' calculations based on Framework for Teaching pilot evaluation scores from the 2012–2013 school year provided by the Pennsylvania Department of Education.

Note: PPR = Professional Practice Rating

Table A.4. Summary of phase 3 domain and PPR scores for teachers in Pittsburgh and not in Pittsburgh—all districts

Domain	Number of teachers with scores	District	Percentage of teachers earning:				Difference in percentage *proficient* or *distinguished* (PPS minus non-PPS)
			Failing	*Needs improvement*	*Proficient*	*Distinguished*	
Domain 1	1,038	PPS	0.1	8.1	85.4	6.5	-6.5
	5,637	non-PPS	0.0	1.6	77.0	21.4	
Domain 2	1,038	PPS	0.0	8.0	81.5	10.5	-5.6
	5,638	non-PPS	0.0	2.4	76.3	21.3	
Domain 3	1,038	PPS	0.1	19.9	76.1	3.9	-15.7
	5,638	non-PPS	0.1	4.2	83.0	12.7	
Domain 4	1,038	PPS	0.0	3.7	86.0	10.3	-2.3
	5,638	non-PPS	0.1	1.3	78.8	19.8	
PPR	1,038	PPS	0.0	7.0	87.7	5.3	-5.3
	5,637	non-PPS	0.0	1.8	84.4	13.9	

Source: Authors' calculations based on Framework for Teaching pilot evaluation scores from the 2012–2013 school year provided by the Pennsylvania Department of Education.

Notes: The difference in the percentage of teachers earning *proficient* or *distinguished* ratings is calculated by adding the percentage receiving *proficient* to the percentage receiving *distinguished* for Pittsburgh and non-Pittsburgh teachers separately and then subtracting the combined non-Pittsburgh percentage from the combined Pittsburgh percentage. A negative value indicates that the percentage of Pittsburgh teachers receiving *proficient* or *distinguished* ratings was smaller than the proportion of non-Pittsburgh teachers receiving either of the same two ratings.

PPS = Pittsburgh Public Schools teachers; non-PPS = teachers from districts other than Pittsburgh Public Schools; PPR = Professional Practice Rating.

Table A.5. Comparison of average FFT and value-added scores for phase 3 teachers in Pittsburgh and in other districts

Measure	Pittsburgh	Non-Pittsburgh
Professional Practice Rating	2.0**	2.2
	(0.3)	(0.3)
Value-added score	-0.2**	-0.1
	(0.5)	(0.6)
Sample size	265	1465

Source: Authors' calculations based on value-added scores and phase 3 Framework for Teaching pilot evaluation scores provided by the Pennsylvania Department of Education.

Notes: Standard deviation reported below each average. The sample includes phase 3 teachers with VAM estimates.

**Significantly different from non-Pittsburgh teachers at the 0.05 level, two-tailed test.

Table A.6. Summary of phase 3 domain and PPR scores in phases 2 and 3 for teachers in at least one phase—Pittsburgh only

		Percentage of teachers earning:				Difference in percent receiving *proficient* or *distinguished* (phase 3 minus phase 2)
Domain	Phase	*Failing*	*Needs Improvement*	*Proficient*	*Distinguished*	
Domain 1	Phase 2	0.0	2.7	85.1	12.2	-5.4
	Phase 3	0.1	8.1	85.4	6.5	
Domain 2	Phase 2	0.1	5.7	79.3	14.9	-2.2
	Phase 3	0.0	8.0	81.5	10.5	
Domain 3	Phase 2	0.1	10.8	83.2	6.0	-9.2
	Phase 3	0.1	19.9	76.1	3.9	
Domain 4	Phase 2	0.0	4.0	87.0	9.1	0.2
	Phase 3	0.0	3.7	86.0	10.3	
PPR	Phase 2	0.0	4.5	91.3	4.2	-2.5
	Phase 3	0.0	7.0	87.7	5.3	

Source: Authors' calculations based on Framework for Teaching pilot evaluation scores from the 2011–2012 and 2012–13 school years provided by the Pennsylvania Department of Education.

Notes: The difference in the percentage of teachers earning *proficient* or *distinguished* ratings is calculated by adding the percentage receiving *proficient* to the percentage receiving *distinguished* for phase 2 and 3 teachers separately and then subtracting the combined phase 2 percentage from the combined phase 3 percentage. A negative value indicates that the percentage of phase 3 teachers receiving *proficient* or *distinguished* ratings was smaller than the percentage of phase 2 teachers receiving either of the same two ratings.

Table A.7. Summary of phase 3 domain and PPR scores in phases 2 and 3 for teachers in at least one phase—not Pittsburgh only

		Percentage of teachers earning:				Difference in percent receiving *proficient* or *distinguished* (phase 3 minus phase 2)
Domain	Phase	*Failing*	*Needs Improvement*	*Proficient*	*Distinguished*	
Domain 1	Phase 2	0.0	1.6	71.9	26.5	-0.0
	Phase 3	0.0	1.6	77.0	21.4	
Domain 2	Phase 2	0.0	2.0	72.2	25.7	-0.3
	Phase 3	0.0	2.4	76.3	21.3	
Domain 3	Phase 2	0.0	2.8	79.2	18.0	-1.5
	Phase 3	0.1	4.2	83.0	12.7	
Domain 4	Phase 2	0.0	1.6	73.0	25.4	0.2
	Phase 3	0.1	1.3	78.8	19.8	
PPR	Phase 2	0.0	1.4	82.8	15.7	-0.2
	Phase 3	0.0	1.8	84.4	13.9	

Source: Authors’ calculations based on Framework for Teaching pilot evaluation scores from the 2011–2012 and 2012–2013 school years provided by the Pennsylvania Department of Education.

Notes: The difference in the percentage of teachers earning *proficient* or *distinguished* ratings is calculated by adding the percentage receiving *proficient* to the percentage receiving *distinguished* for phase 2 and 3 teachers separately and then subtracting the combined phase 2 percentage from the combined phase 3 percentage. A negative value indicates that the percentage of phase 3 teachers receiving *proficient* or *distinguished* ratings was smaller than the percentage of phase 2 teachers receiving either of the same two ratings.

3. Distribution of phase 3 Framework for Teaching scores for teachers in Pittsburgh and in other districts who were rated in both phases

To account for differences in the composition of the Pittsburgh and non-Pittsburgh teacher samples in phases 2 and 3, we examined changes in the distribution of domain and PPR scores for Pittsburgh and non-Pittsburgh teachers who were rated in both phases. Overall, 965 teachers participated in both phases of the pilot, including 824 from Pittsburgh and 141 from other districts. We then compared those changes with changes in the distributions of scores for the full phase 2 and 3 Pittsburgh and non-Pittsburgh samples (teachers rated in at least one phase). A difference in the change in scores between those rated in both phases and those rated in at least one phase would indicate that compositional differences in the teacher samples may partly explain the changes between phases for the overall samples.

We did find a slightly larger decrease in the proportion receiving *distinguished* domain 3 ratings and a slightly smaller decrease in the proportion receiving *proficient* domain 3 ratings among Pittsburgh teachers rated in both phases (Table A.8). However, in general, we found that changes in the distributions of scores for both Pittsburgh and non-Pittsburgh teachers were similar for teachers rated in both phases and those rated in at least one phase (Tables A.8 and A.9). These findings suggest that changes in the composition of Pittsburgh and non-Pittsburgh teachers, for the most part, may not have impacted the distribution of evaluation scores across the two phases.

Table A.8. Summary of rubric domain scores—comparison of Pittsburgh teachers rated in at least one or both phases

		Change between phases 2 and 3 in percentage of teachers earning (phase 3 minus phase 2):			
Domain	Phases rated in	*Failing*	*Needs improvement*	*Proficient*	*Distinguished*
Domain 1	At least one	0.1	5.4	0.3	-5.8
	Both	0.1	4.7	1.4	-6.3
Domain 2	At least one	-0.1	2.3	2.2	-4.4
	Both	0.0	3.0	3.0	-6.0
Domain 3	At least one	0.0	9.2	-7.1	-2.1
	Both	0.0	9.1	-4.7	-4.4
Domain 4	At least one	0.0	-0.3	-0.9	1.2
	Both	0.0	-0.8	-1.3	2.1
PPR	At least one	0.0	2.5	-3.6	1.1
	Both	0.0	2.0	-3.4	1.4

Source: Authors' calculations based on Framework for Teaching pilot evaluation scores from the 2011–2012 and 2012–2013 school years provided by the Pennsylvania Department of Education.

Notes: The change between phases 2 and 3 in the percentage of teachers earning each performance rating is calculated by subtracting the percentage of phase 2 teachers receiving that rating from the corresponding phase 3 percentage. A negative value indicates that the percentage of phase 3 teachers receiving that performance rating was smaller in phase 3 than in phase 2.

PPR = Professional Practice Rating.

Table A.9. Summary of rubric domain scores–comparison of non-Pittsburgh teachers rated in at least one or both phases

		Change between phases 2 and 3 in percentage of teachers earning (phase 3 minus phase 2):			
Domain	**Phases rated in**	***Failing***	***Needs improvement***	***Proficient***	***Distinguished***
Domain 1	At least one	0.0	0.0	5.0	-5.0
	Both	0.0	0.7	4.7	-5.4
Domain 2	At least one	0.0	0.4	4.0	-4.4
	Both	0.0	-0.7	11.1	-10.3
Domain 3	At least one	0.1	1.4	3.8	-5.3
	Both	0.0	-2.1	5.8	-3.7
Domain 4	At least one	0.1	-0.3	5.7	-5.5
	Both	0.0	0.0	7.5	-7.5
PPR	At least one	0.0	0.4	1.5	-1.9
	Both	0.0	0.7	4.0	-4.7

Source: Authors' calculations based on Framework for Teaching pilot evaluation scores from the 2011–2012 and 2012–2013 school years provided by the Pennsylvania Department of Education.

Notes: The change between phases 2 and 3 in the percentage of teachers earning each performance rating is calculated by subtracting the percentage of phase 2 teachers receiving that rating from the corresponding phase 3 percentage. A negative value indicates that the percentage of phase 3 teachers receiving that performance rating was smaller in phase 3 than in phase 2.

PPR = Professional Practice Rating.

4. Framework for Teaching internal consistency leave-out scores

Tables A.10 and A.11 present leave-out scores for each FFT component and domain. Leave-out scores assess the internal consistency of a scale with a particular item excluded. A leave-out score that is substantially different from the Cronbach's alpha when no components or domains are excluded indicates that a particular component or domain is generally inconsistent with the other components in its domain or the other domains in the FFT.

Table A.10. Cronbach's alpha values for Framework for Teaching domains when particular components are excluded

	Phase 2		Phase 3	
Portion of the Framework for Teaching excluded when calculating alpha	**Cronbach's alpha**	**Sample size**	**Cronbach's alpha**	**Sample size**
Domain 1: planning and preparation, excluding:				
No components	0.78	1,659	0.80	6,422
1a: demonstrating knowledge of content and pedagogy	0.74	1,663	0.77	6,425
1b: demonstrating knowledge of students	0.76	1,671	0.79	6,431
1c: setting instructional outcomes	0.74	1,667	0.76	6,428
1d: demonstrating knowledge of resources	0.76	1,678	0.78	6,444
1e: planning coherent instruction	0.73	1,675	0.76	6,423
1f: designing ongoing formative assessments	0.75	1,681	0.78	6,459
Domain 2: classroom environment, excluding:				
No components	0.75	1,639	0.76	6,401
2a: creating a learning environment of respect and rapport	0.70	1,658	0.71	6,405
2b: establishing a culture for Learning	0.70	1,651	0.72	6,406
2c: managing classroom procedures	0.70	1,661	0.71	6,413
2d: managing student behavior	0.69	1,651	0.70	6,418
2e: organizing physical space	0.76	1,688	0.75	6,450
Domain 3: instruction, excluding:				
No components	0.72	1,646	0.77	6,373
3a: communicating with students	0.68	1,688	0.73	6,377
3b: using questioning and discussion techniques	0.66	1,652	0.73	6,397
3c: engaging students in learning	0.64	1,654	0.71	6,375
3d: using assessment to inform instruction	0.68	1,652	0.73	6,390
3e: demonstrating flexibility and responsiveness	0.69	1,707	0.73	6,455
Domain 4: professional responsibilities, excluding:				
No components	0.75	1,440	0.77	5,975
4a: reflecting on teaching and student learning	0.71	1,448	0.74	6,049
4b: system for managing student data	0.74	1,445	0.74	6,011
4c: communicating with families	0.75	1,463	0.76	6,052
4d: participating in a professional community	0.70	1,459	0.73	5,985
4e: growing and developing professionally	0.70	1,457	0.73	6,007
4f: showing professionalism (PPS data)	0.71	1,586	0.72	5,982

Source: Authors' calculations based on Framework for Teaching pilot evaluation scores from the 2011–2012 and 2012–2013 school years provided by the Pennsylvania Department of Education.

Note: The sample sizes for each component differ because only teachers with a rating for all other components in the domain are included in calculating the leave-out scores.

Table A.11. Cronbach's alpha values for the full Framework for Teaching Professional Practice Rating when particular domains are excluded

	Phase 2		Phase 3	
Portion of the Framework for Teaching excluded when calculating alpha	Cronbach's alpha	Sample size	Cronbach's alpha	Sample size
No components	0.84	2,487	0.87	6,675
Domain 1: planning and preparation	0.79	2,489	0.83	6,676
Domain 2: classroom environment	0.80	2,491	0.84	6,675
Domain 3: instruction	0.77	2,489	0.82	6,675
Domain 4: professional responsibilities	0.82	2,499	0.85	6,675

Source: Authors' calculations based on Framework for Teaching pilot evaluation scores from the 2011–2012 and 2012–2013 school years provided by the Pennsylvania Department of Education.

Note: The sample sizes for each domain differ because only teachers with a rating for all other domains are included in calculating the leave-out scores.

APPENDIX B

ESTIMATING TEACHER EFFECTIVENESS BASED ON VALUE ADDED

This appendix provides an overview of the process we use for obtaining teacher effectiveness estimates through a value-added model (VAM). This process involves first estimating VAMs for each subject and grade between 4th and 8th grade. The next step is to create for each teacher an overall value-added measure that combines the teacher's VAM estimates across the grades and subjects that the teacher served. See the description in Appendix B of Walsh and Lipscomb (2013) for more information on this process.

1. Estimate the teacher value-added models

The VAMs estimated in this report provide measures of teachers' contributions to student learning in 4th through 8th grade math and reading, 5th and 8th grade writing, and 4th and 8th grade science. We use Pennsylvania System of School Assessment (PSSA) scores in these grades and subjects as outcomes, and students' own prior PSSA scores as baselines. The VAMs base teachers' effectiveness estimates on as many as three years of teaching.

A. The value-added model

The following regression equation, estimated separately for each grade-subject combination, describes the teacher VAMs:

$$A_{itcy} = \beta' P_{i(y-1)} + \gamma' X_{iy} + \theta' C_{itcy} + \delta' T_{ity} + \varphi' Y_y + e_{itcy} \quad (1)$$

In the model, A_{itcy} is an assessment score for student *i*, taught by teacher *t* in class *c*, in year *y* between 2010–2011 and 2012–2013. For example, A_{itcy} could be a 5th grade PSSA math assessment. The sample would comprise student-teacher-class-year combinations across the state over a set period of years in which the student took a particular assessment and was taught by a particular teacher in the subject of the assessment. The vector $\mathbf{P}_{i(y\text{-}1)}$ includes school-year-specific variables for student *i*'s prior-year PSSA scores. We include prior-year math and reading scores in all VAMs, and prior-year science and writing scores in VAMs where those scores would be available in the prior year. Including prior-year scores in two or more subjects captures a broader range of prior inputs than if only a same-subject prior-year score were used. For most students, prior-year scores come from the previous grade. However, prior scores for grade repeaters come from the same grade as the outcome variable. Therefore, the vector $\mathbf{P}_{i(y\text{-}1)}$ also includes a set of variables containing grade repeaters' same-grade PSSA scores from the previous year.

The vector $\mathbf{X}_{iy}$ is a set of variables for observed individual student characteristics. The vector $\mathbf{C}_{itcy}$ is a set of variables for the characteristics of student *i*'s classroom peers. The vector $\mathbf{Y}_y$ includes year indicators for the school years in the VAM. The coefficients in $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\theta}$ are the estimated relationships between students' assessment scores and each respective student characteristic, controlling for the other factors in the model. The variable e_{itcy} is the error term.[15]

The vector $\mathbf{T}_{ity}$ includes a teacher dummy variable for each teacher in the VAM that is equal to one for students taught by the teacher, and zero otherwise. Students taught by multiple teachers are included in the model on multiple rows, once for each teacher, and each student-teacher-course-year observation has exactly one non-zero element in $\mathbf{T}_{ity}$. We use a weighted

[15] We use a standard cluster-robust variance estimator to obtain standard errors that adjust for clustering of observations by student and account for heteroskedasticity.

least squares regression to accurately attribute the exposure of students to teachers during the school year. This approach gives less weight to students in calculating a teacher’s value added when students are also taught by another teacher in the same subject, grade, and year. A student contributes up to a total of 100 percent of his or her dosage to one or more teachers. A student’s dosage is split between teachers when the student takes multiple courses in the same subject. This approach is known as the “full-roster” method of estimating VAM (Hock and Isenberg 2012).

The vector $\boldsymbol{\delta}$ is a set of coefficients to be estimated, one for each teacher in the VAM. Each coefficient in $\boldsymbol{\delta}$ identifies a teacher’s contribution to student learning—the extent to which the actual achievement of students tends to be above or below what is predicted for an average teacher. The average value-added score is set equal to zero but does not mean that student learning is zero for the teacher with the average value-added score. Rather, a positive value-added estimate represents above-average teacher performance, and a negative estimate represents below-average performance. The reference point for determining the average teacher contribution depends on the sample of teachers in the model. Because the model includes students and teachers across the state, the value-added estimates are calculated relative to the contribution of the average teacher in Pennsylvania in the grade, subject, and school years covered by the VAM. Teachers’ final value-added scores are based on a weighted average of these coefficient estimates (see section B below).

B. Correcting for measurement error in the pre-tests

The VAMs rely on students’ own prior achievement scores as indicators of their academic abilities before entering a teacher’s classroom. Standardized tests are imperfect measures of students’ true abilities. The measurement error introduced by using prior assessment scores as ability measures causes standard regression techniques to produce biased estimates of teacher effectiveness. We correct for measurement error by incorporating directly into the regression models the test/retest reliability of the PSSA tests. This approach, called an errors-in-variables (EIV) regression, eliminates bias due to the known amount of measurement error in students’ prior-year tests (Buonaccorsi 2010). In terms of Equation (1), EIV provides a better estimate of $\boldsymbol{\beta}$ than would be obtained by ordinary regression.

C. Controlling for students’ prior-year achievement and other background measures

We control for students’ test score histories by including in the VAMs their assessment scores in all subjects from the previous year. We include separate prior-year variables for PSSAs in each subject-grade-year combination to allow the relationships between each prior-year test and achievement to vary across grade-year combinations. Students who repeat a grade are included in the VAM. For such students, we include additional prior-year PSSA variables (because the grade level of the prior-year test will be different from that of non-grade-repeaters).

Because students do not take the science and writing PSSAs in consecutive grades, we cannot include prior-year scores in science and writing in these VAMs. We use prior-year math and reading scores instead. The lack of a same-subject, prior-year test does not prevent the VAM from determining whether students’ scores (for example, 4th grade PSSA science) are higher or lower than predicted. Although science and writing VAMs are, in this way, feasible to estimate, we expect the resulting estimates will be relatively less precise than estimates from VAMs that can include a same-subject, prior-year test.

Table B.1 lists the outcome and prior-year assessments in the VAMs for students who do not repeat a grade. We require that any student included in a VAM have at least one prior-year test score. We impute a small fraction of scores (less than 1 percent) for students who are missing one or more of the prior-year test scores but not the same-subject score.[16] The imputations are based on the relationships with other prior-year scores and observed characteristics of students who have nonmissing scores. For grade repeaters, the prior-year baselines come from the same grade as the outcome assessment and enter the VAM as different variables from the prior-year baselines for nonrepeaters.

Table B.1. PSSAs used as outcomes and baselines in the teacher value-added models

Outcome		Prior-year baseline	
Subject	Grade	Subject	Grade
Math	4	Math, reading	3
Reading	4	Math, reading	3
Science	4	Math, reading	3
Math	5	Math, reading, science	4
Reading	5	Math, reading, science	4
Writing	5	Math, reading, science	4
Math	6	Math, reading, writing	5
Reading	6	Math, reading, writing	5
Math	7	Math, reading	6
Reading	7	Math, reading	6
Math	8	Math, reading	7
Reading	8	Math, reading	7
Science	8	Math, reading	7
Writing	8	Math, reading	7

Note: Baseline scores for grade repeaters are their prior-year scores in the same grade as the outcome variable.

To help isolate the effect of teachers on student achievement, the VAMs also include control variables for observable student and classroom background characteristics. Table B.2 lists these

[16] For this purpose, we treat math as the same-subject prior-year test for science assessments, and reading as the same-subject prior-year test for writing assessments. Students missing the same-subject pre-test are dropped.

variables, which enter Equation (1) through the vectors X_{iy} and C_{itcy}. The factors that are included are thought to be correlated with student performance and outside the control of teachers.

Table B.2. Student and classroom characteristics included in the value-added models

Control variable	Definition
Free meals	Free meals eligibility {0,1}
Reduced-price meals	Reduced-price meals eligibility {0,1}
English-language learner (ELL)	ELL in outcome year {0,1}
Specific learning disability (SLD)	Designation of SLD under IDEA {0,1}
Speech or language impairment (SLI)	Designation of SLI under IDEA {0,1}
Emotional disturbance (ED)	Designation of ED under IDEA {0,1}
Intellectual disability (ID)	Designation of ID under IDEA {0,1}
Autism (AUT)	Designation of AUT under IDEA {0,1}
Physical/sensory impairment (PSi)	Designation of hearing impairment, visual impairment, deaf-blindness, or orthopedic impairment under IDEA {0,1}
Other impairment	Designation of other health impairment, multiple disabilities, developmental delay, or traumatic brain injury under IDEA {0,1}
Mobility	Attended multiple schools during school year {0,1}
Grade repeater	Repetition of the current grade {0,1}
Behind grade	More than 1.5 years older than expected for grade {0,1}
Age	Student age in years as of September 1
PSSA-modified (outcome)	Outcome is a PSSA-M score (PSSA outcomes only) {0,1}
PSSA-modified (prior-year math)	Prior-year math score is a PSSA-M score {0,1}
PSSA-modified (prior-year reading)	Prior-year reading score is a PSSA-M score {0,1}
Gender	Female {0,1}
Race/ethnicity	Indicators for African American, Hispanic, Asian Pacific Islander, or other race/ethnicity {0,1}
Classroom-level characteristics	Classroom average prior math and reading test scores (separate variables). Also included are classroom average standard deviations of prior math and reading test scores, and classroom size.

Notes: Peers are defined as a student's classmates in a particular classroom. IDEA = Individuals with Disabilities Education Act; PSSA-M = PSSA-Modified.

2. Obtain overall value-added scores by combining each teacher's individual scores

To obtain an overall value-added measure for each teacher, we combine teachers' value-added estimates for their grades and subjects. The composite measure represents the average contribution of teachers to their students' achievement across grades and subjects. To calculate the composite value-added measure, we first remove any estimates for teachers that are based on fewer than 10 student full-time equivalents. We then standardize teachers' estimates to have the same variance across grades and subjects. Finally, we average their grade- and subject-specific estimates to obtain a composite measure. The average is weighted based on the number of the teacher's students who contribute to the VAM for the grade-subject combination. We also calculate the precision of teachers' composite measures based on the precision of their grade- and subject-specific estimates and the covariance between their estimates across subjects.[17,18]

3. Number of students and teachers included in the Value-Added Models

In Table B.3, we show student sample sizes for each PSSA subject and grade-level assessment. The first column of data lists the number of students with assessment scores. The VAMs include about 90 percent of these student scores. In particular, the VAMs include those students with nonmissing (or imputed) prior scores and student background characteristics, who can be linked to a teacher in a course aligned with the subject area of the assessment.

Table B.3. Number of students with assessment scores and that are included in the teacher value-added models, by subject and grade level

Outcome	Number of students with an assessment score	Number of students included in the teacher VAMs
Math PSSA, grade 4	368,528	327,802
Math PSSA, grade 5	373,389	331,016
Math PSSA, grade 6	377,796	347,351
Math PSSA, grade 7	380,296	366,225
Math PSSA, grade 8	380,632	359,964
Reading PSSA, grade 4	368,438	328,246
Reading PSSA, grade 5	373,315	333,833
Reading PSSA, grade 6	377,683	350,648
Reading PSSA, grade 7	380,111	363,881
Reading PSSA, grade 8	380,351	358,810
Writing PSSA, grade 5	369,731	330,879
Writing PSSA, grade 8	377,046	356,197
Science PSSA, grade 4	367,528	311,100
Science PSSA, grade 8	377,948	363,859

Source: Mathematica calculations based on Pennsylvania student data.

Note: Sample sizes refer to unique student observations. Students are counted only once if they appear in a sample in multiple years.

PSSA = Pennsylvania System of School Assessment; VAM = value-added model.

[17] We calculate the standard error of the combined estimate as the square root of the weighted sum of variances and covariances, divided by the total student equivalents taught by the teacher across all VAMs. The weights in the sum are the squared student equivalents for the specific VAM. We approximate each covariance as the correlation between value-added scores in the two subjects (within a grade), multiplied by the standard errors of a teacher's estimates in the subjects. We account for covariances only between subjects, and not between grades. This choice reflects the likelihood that teachers do not typically share many of the same students across the grades they teach, whereas many teachers are responsible for instructing the same students in multiple subjects.

[18] When calculating correlations between value-added estimates and FFT scores, we use *pre-shrinkage* value-added estimates—estimates that are not adjusted using an empirical Bayes shrinkage procedure—and adjust the correlations for imprecision in value added using the method in Jacob and Lefgren (2008). Although the shrinkage procedure is an appropriate way to reduce misclassification of teachers in many value-added contexts, it can also lead to bias in value-added estimates and generally is not used for value-added unless teacher-level results are being used to make policy decisions for individual teachers.

We report in Table B.4 the number of teachers with VAM estimates, by outcome. To be included in Table B.4, teachers must have taught students in 2012–2013, the year of the phase 3 teacher evaluation pilot. In addition, they must have taught at least 10 students across grades and subjects over the three-year period covered by the VAMs.

Table B.4. Number of teachers with value-added estimates, by outcome

Outcome	2010–2011 to 2012–2013
Math PSSA, grade 4	5,047
Math PSSA, grade 5	4,863
Math PSSA, grade 6	3,684
Math PSSA, grade 7	2,734
Math PSSA, grade 8	2,709
Reading PSSA, grade 4	5,146
Reading PSSA, grade 5	5,017
Reading PSSA, grade 6	4,344
Reading PSSA, grade 7	3,492
Reading PSSA, grade 8	3,294
Writing PSSA, grade 5	5,058
Writing PSSA, grade 8	3,348
Science PSSA, grade 4	4,734
Science PSSA, grade 8	1,817
Teachers with at least one VAM estimate	25,404
Phase 3 teachers with at least one VAM estimate	1,730

Source: Mathematica calculations based on Pennsylvania student data.

Note: Teachers are included in multiple VAMs if they have students in multiple grades or subjects. The number of teachers with estimates excludes teachers whose estimates were based on fewer than 10 student equivalents across all grades and subjects they teach and teachers who did not teach any students in the most recent year included in the VAM (2012–2013).

4. Descriptive statistics from the distributions of value-added model estimates

Table B.5 provides technical results from the teacher VAMs. The first column of data reports the standard deviation of value-added estimates. The standard deviation is a measure of the wideness of the value-added distribution. In any distribution, the most effective teachers (those at the rightmost tail of the distribution) make the largest contributions to student achievement growth. When the value-added distribution is flatter (that is, more spread out and with a larger standard deviation), the amount of growth in student achievement is more positive for the most effective teachers and more negative for the least effective teachers than when the value-added distribution is tightly concentrated. A standard deviation of 0.23, the value for math in grade 4, means that a teacher at the 84th percentile of effectiveness raises student achievement by 0.23 standard deviations of student test scores more than the teacher at the 50th percentile of effectiveness. This result is equivalent to raising students’ 4th grade math scores from the 50th percentile to the 59th percentile.

Table B.5. Descriptive Characteristics of the VAM distributions

Outcome	84th minus 50th percentile of "underlying" value added (in z-score units)	Average standard error (in z-score units)	Percentage of estimates that are statistically significant
Math PSSA, grade 4	0.23	0.08	0.54
Math PSSA, grade 5	0.22	0.07	0.55
Math PSSA, grade 6	0.22	0.07	0.56
Math PSSA, grade 7	0.22	0.07	0.56
Math PSSA, grade 8	0.21	0.07	0.55
Reading PSSA, grade 4	0.22	0.08	0.50
Reading PSSA, grade 5	0.18	0.08	0.43
Reading PSSA, grade 6	0.17	0.08	0.40
Reading PSSA, grade 7	0.19	0.08	0.47
Reading PSSA, grade 8	0.18	0.08	0.39
Writing PSSA, grade 5	0.35	0.10	0.62
Writing PSSA, grade 8	0.33	0.10	0.59
Science PSSA, grade 4	0.26	0.08	0.58
Science PSSA, grade 8	0.22	0.07	0.61

Source: Mathematica calculations based on Pennsylvania student data.

Notes: The VAMs are based on statewide samples of teachers and students. Teachers' VAM estimates are based on students in their classrooms at any time during the specified analysis periods.

One z-score unit is equal to one standard deviation of student outcomes. One standard deviation of student outcomes is approximately equal to 230 PSSA points in math, 220 points in reading, 280 points in writing, and 190 points in science.

The 84th minus 50th percentile of underlying VAM estimates is the estimated difference in "underlying" value added for the teachers at these percentiles (that is, perfect measures of value added that do not have any estimation error). This value is calculated as the standard deviation of value-added estimates with an adjustment for the amount of estimation error using the method in Morris (1983).

All estimates for individual subject-grade combinations are pre-shrinkage.

The percentage of estimates that are statistically significant uses a 95 percent confidence interval. Statistically significant teacher VAM estimates can be distinguished above or below average performance.

PSSA = Pennsylvania System of School Assessment; VAM = value-added model.

The second and third columns of data show the average standard error of the VAM estimates and the proportion of estimates that are statistically significant, respectively. The average standard error is a measure of noise in the estimates. When VAM estimates have more noise, they need to be larger (or smaller) than the average estimate by a greater margin to be distinguished statistically from the average estimate. The percentage of estimates that are statistically significant shows the percentage of estimates that can be distinguished above or below average performance with 95 percent confidence, given the standard deviation of estimates and their average standard error.

www.mathematica-mpr.com

Improving public well-being by conducting high quality, objective research and data collection

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■ WASHINGTON, DC

MATHEMATICA
Policy Research