GETTING DOWN TO FACTS II

Technical Report

Can Teacher Evaluation Programs Improve Teaching?

Virginia Lovison
Harvard University

Eric S. Taylor
Harvard University

September 2018

About: The *Getting Down to Facts* project seeks to create a common evidence base for understanding the current state of California school systems and lay the foundation for substantive conversations about what education policies should be sustained and what might be improved to ensure increased opportunity and success for all students in California in the decades ahead. *Getting Down to Facts II* follows approximately a decade after the first Getting Down to Facts effort in 2007. This technical report is one of 36 in the set of *Getting Down to Facts II* studies that cover four main areas related to state education policy: student success, governance, personnel, and funding.

Stanford University

PACE Policy Analysis for California Education

Can Teacher Evaluation Programs Improve Teaching?

Virginia Lovison
Harvard University

Eric S. Taylor
Harvard University

Acknowledgements

We thank Katharine Strunk, Jason Grissom, Susanna Loeb, Matt Kraft, and other Getting Down to Facts II authors for helpful comments and early suggestions. We are equally grateful to several district administrators, from the five districts we highlight, for their insights and comments on our descriptions of their evaluation programs. Erika Byun contributed excellent research assistance.

Introduction

In this *Getting Down to Facts* report we focus on teacher evaluation programs, and further focus on the features of evaluation programs which may promote or hinder teachers' effectiveness in their work.

Why Focus on Teacher Evaluation and Teacher Effectiveness?

The past decade has brought dramatic growth in teacher evaluation in American public schools; growth in the money, time, and effort devoted to evaluation, but also growth in the sophistication and innovativeness of evaluation measures and other program features. There were many forces driving that growth. One force was incentives from the federal government, beginning notably with the Race to the Top competition and continuing with the requirements of NCLB waivers. A second force was new research evidence documenting large differences in teaching performance between individual teachers.

Advocates for teacher evaluation often point to the potential for evaluation to help individual teachers become more effective in the work of teaching. This was a third force or motivation behind the recent decade's growth, but it is not a new motivation for evaluation. The goal of improving teachers' effectiveness is fundamental, for example, in the peer assistance and review (PAR) programs which began in Toledo, Ohio in the 1980s, were further developed in places like the Poway and Mt. Diablo Unified School Districts (USDs), and had spread widely in California by the turn of the century.

This motivation—improving individuals' effectiveness in the work of teaching—is our present focus. As this paper progresses we will elaborate on what *more effective* means and *how* evaluation programs may promote or hinder such improvements. In short, evaluation may improve job performance, for example, by incentivizing teachers to give more attention or effort to specific teaching practices, or by providing objective feedback about teaching practices where an individual needs to focus efforts to improve, or by providing a new setting to practice and deepen teaching skills (Milanowski and Henemen 2001, Taylor and Tyler 2012).

Teacher Evaluation in California Today

Teacher evaluation in California today is a district responsibility, partly *de facto* and partly *de jure,* but California's school districts do act on that responsibility. Over both recent years and many decades, California districts have produced a range of substantively different approaches to teacher evaluation, demonstrating both the potential for district-level action generally and specific design options. Later in this report we highlight several district examples, but in the next paragraphs we discuss where things stand at the state level.

There is no state-mandated, or even state model, teacher evaluation program in California.[1] The previous sentence is not surprising to most California educators, but would be surprising to educators in other states where statewide evaluation programs have become the norm in recent years. A typical statewide program specifies many features, for example, the use of several specific types of performance measures and weights for calculating an overall score; but the typical state also allows local adaptation (for a review see Steinberg and Donaldson 2016). The proximate reason for new statewide evaluation policies was federal government requests during the Obama administration, especially through the NCLB waiver process, to which most states acceded. The state of California chose, ultimately, to not seek a waiver, at least partly, to avoid the federal government's requirements about teacher evaluation (LA Times 2013).[2]

While there is no statewide teacher evaluation *program*, there are state laws that govern teacher evaluation. In practice, however, those laws are not a binding constraint on districts' decisions about how to evaluate teachers. The existing state statutes are known as the Stull Act, first passed by the state legislature in 1976 (California Education Code §44660-44665). The Stull Act requires each district to have an evaluation program, but leaves most decisions to individual districts. As an example, one of the more notable and prescriptive provisions of the Stull Act (appears to) require that districts evaluate teachers based on, among other things, "the progress of pupils towards…the state adopted…standards as measured by state adopted [tests]" (§44662(b)(1)), but gives no more details.[3] Moreover, in practice districts have been allowed to ignore the provision quoted above and other provisions of the law. Legislative efforts to change the law in recent years have been unsuccessful.

In short, teacher evaluation is, and will likely remain, the responsibility of each California school district. Thus we have written this report primarily with district leaders, managers, and policymakers in mind. This report was not written to argue for or against a change in state policy.

This Report

In this report we discuss several key features of evaluation programs which may promote or hinder teachers' effectiveness in their work. We do not attempt to prescribe a single evaluation program design, made up of specific features, for all of California. Instead our purpose is to provide an introduction to key issues and evidence for California's policymakers and school leaders who are concerned about teacher evaluation in their districts and schools.

The examples and evidence we summarize do identify some promising evaluation design features—promising in the sense that they have, at least in one or two cases, helped improve teacher effectiveness in other states and districts. But, in general, research evidence on whether

---

[1] The California Commission on Teacher Credentialing does provide the California Standards for the Teaching Profession (CSTP) and the accompanying Continuum of Teaching Practice rubric as discussed below.

[2] Several California districts, working together as the "CORE districts," did receive a NCLB waiver. The CORE waiver application included changes to the districts' teacher evaluation programs.

[3] There is disagreement among stakeholders on how to interpret the language of this provision and the accompanying statutory language, and, as a result, disagreement about just how compulsory the provision is (*Doe v. Antioch* 2016, LA Times 2016).

and how evaluation promotes teaching effectiveness is still relatively scarce compared to other aspects of managing schools. Leaders and policymakers should proceed with thoughtful caution. We have also pointed out some known cautions in the discussion below.

Our report is organized around four themes of contemporary teacher evaluation programs: First, evaluation which is based, at least in part, on multiple classroom observations structured by and scored with a detailed rubric. Second, making clear, easy, direct connections between an individual's evaluation results and resources to help that individual in her efforts to improve. Third, evaluation using multiple measures of effectiveness in teaching. One potential measure being subjective evaluations from school principals or other close supervisors. Fourth, programs which do or do not attach consequences to evaluation results, most notably tenure decisions.

For each of these four features, we provide examples of different approaches in practice in California school districts. We highlight five California districts and summarize key features of their evaluation programs; the five are Poway, Long Beach, Los Angeles, San Jose, and San Juan Unified School Districts. These five districts were not selected because they represent typical California districts or typical evaluation programs. We selected these five to show a diversity of evaluation programs in use by California districts today. We also selected these five because they show different approaches to the four features we highlight. For example, some use multiple measures while others do not.

Also for each of the four features, we summarize scholarly research which provides evidence on which approaches are more or less likely to promote improvements in teachers' effectiveness at their work. In selecting the research to include we have set a high bar: we focus primarily on experiments and quasi-experiments which are most likely to sort out causal relationships, not simply report correlations.

Before taking up the four topics, we first report results from a recent survey of California teachers and principals. These results provide some insight into teachers' and principals' current beliefs and attitudes about teacher evaluation in California. And then after discussing the four topics, we include some discussion of the costs of evaluation—both budgetary costs and costs in the form of educators' time and effort which would otherwise be applied to different productive tasks.

California Teachers' and Principals' Current Opinions

Do California's teachers and principals believe their own school's (district's) current evaluation program can improve teaching? When asked to describe their own experience with evaluation, teachers were evenly split. Half of California teachers said evaluation in their school is primarily, mostly, or entirely to *grade teachers for accountability*. The other half felt the opposite; that evaluation in their school is primarily, mostly, or entirely to *help teachers improve their teaching.* These survey results are shown in Figure 1. Principals' responses to the same question were quite different. Just under 20 percent of principals said evaluation in their schools is primarily, mostly, or entirely to *grade teachers*, compared to 50 percent of teachers. And most of

those 20 percent (three-quarters of the 20 percent) said it was primarily about grading teachers, but also somewhat to help teachers improve.[4]

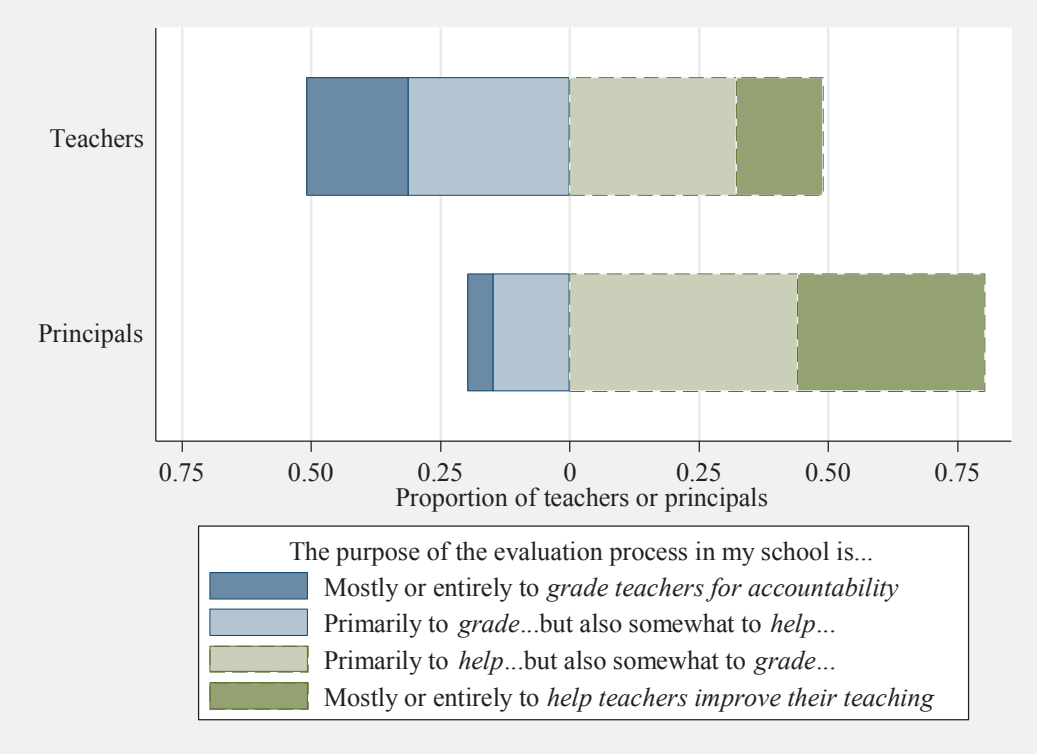

Figure 1. Teacher and principal assessment of the purpose of current teacher evaluation programs

Note: Authors' calculations using RAND ATP/ASLP October 2017 Survey for GDTFII. The full text of the question stem is: "Which of these statements comes closest to describing your own experience? The purpose of the teacher evaluation process in my school is…" The four answer choices are shown above. The full text of the second choices is: "Primarily to grade teachers for accountability, but also somewhat to help teachers improve their teaching." The full text of the third choices follows the same pattern.

[4] We estimate that of the total variation in teachers' opinions about teacher evaluation, perhaps 20-30 percent is between districts. This estimate holds for the overall assessment summarized in Figure 1, and the more detailed questions summarized in Figure 2; it also holds for principals' opinions. However, these results come with an important limitation due to the size of the GDTFII survey sample. Half of the teacher sample (55 percent) and nearly three-quarters of the principal sample (72 percent) are observations where we have just one or two teacher (principal) observations per district. To calculate our estimates, we select a subsample of districts based on the total number of teacher (principal) observations in the district. The range of 20-30 percent arises because our estimate changes as we pick different subsamples (e.g., exclude singleton districts, exclude all districts with only 1-2 observation, 1-3, etc.).

We do not have comprehensive data on the characteristics of district evaluation programs; if we had such data we would investigate whether teachers' and principals' opinions are correlated with those characteristics. The small samples in the GDTFII survey also limit our ability to characterize district level differences. For example, LAUSD has the largest sample, of course, but even that sample is fewer than 30 teacher observations.

The survey responses described in this section were collected specifically for *Getting Down to Facts II* in the last quarter of 2017. The respondents include 459 teachers and 318 principals, which correspond to response rates of 57 percent and 31 percent respectively. The usual caveats with surveys are applicable in this case as well. The particular sample of teachers (principals) who responded to the survey may have unusually (un)favorable opinions about evaluation. The survey process itself, including the question wording, may have elicited unusually (un)favorable opinions. Nevertheless, it seems unlikely these caveats would, for example, overturn the conclusion from the previous paragraph that a strong majority of principals see evaluation as about helping teachers improve.[5]

Teachers' and principals' beliefs and attitudes about evaluation are more than simply context for this paper's discussion. Those beliefs and attitudes can be a barrier to, or an input to, using evaluation to improve teaching effectiveness. For example, only half of California teachers (52 percent) agree with the statement: *"The evaluation process provides me with a clear roadmap of what professional development opportunities to pursue in order to address my areas for improvement."* More teachers, but still not all teachers, (72 percent) agree with the simpler statement: *"The evaluation process in my school helps me identify areas where I can improve."* To be sure, a teacher's opinion of these statements may be different from other, perhaps more objective, ways to assess an evaluation program's key characteristics. But in practice evaluation is much less likely to benefit the one-quarter to one-half of teachers who disagree with these statements. Survey results for these two items and several others are summarized in Figure 2.

A different approach is to ask teachers about outcomes instead of inputs. *"The teacher evaluation process used in my school has led to improvements in my teaching."* More than two-thirds of teachers (69 percent) agreed with this outcomes statement. Two-thirds is an encouraging result. But the remaining one-third (or more) is still a substantial opportunity to improve teaching in California.

We should be cautious, however, in making strong inferences based on these results. Self-assessments are useful, but imperfect, ways to measure job performance, especially improvements in performance. One common problem in surveys is that respondents overstate success or satisfaction; once we have invested effort in something we want it to have been a good investment. We also note that similar issues may arise in the principals' self-assessment of their giving the "right feedback" discussed in the next paragraph.

[5] The survey was fielded by the RAND Corporation using its American Teacher Panel and American School Leader Panel (https://www.rand.org/education/projects/atp-aslp.html). The questions were primarily written by GDTFII researchers. Survey dates were October 27, 2017 through January 5, 2018. The results presented here use RAND's sampling weights which use observable characteristics to adjust for oversampling (undersampling), relative to the population of California teachers and principals, in the construction of the sampling frame; and to adjust for differential unit nonresponse. Even with the weights applied, there may remain selection bias due to unobservable characteristics.

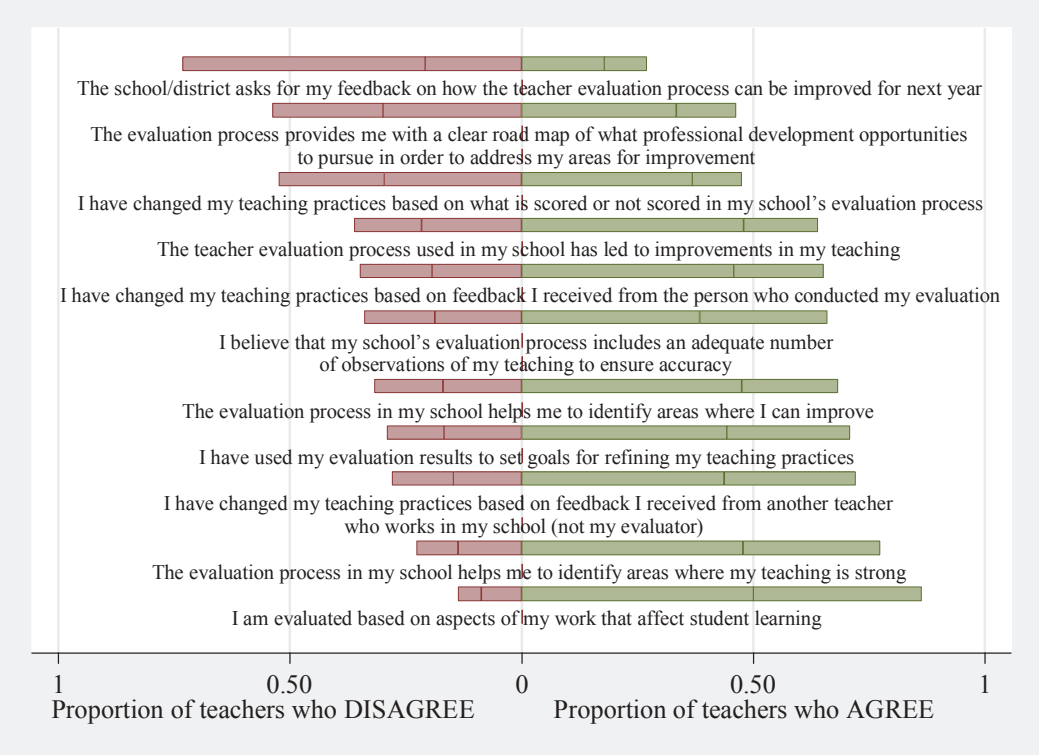

Figure 2a. Teachers' assessment of their district's (school's) evaluation program

Note: Authors' calculations using RAND ATP/ASLP October 2017 Survey for GDTFII. The full text of the question stem is: "To what extent to you agree or disagree with the following statements about teacher evaluation?" Options were: strongly disagree, somewhat disagree, somewhat agree, and strongly agree. The bars above are divided by strongly and somewhat with the proportion somewhat close to the center line.

While there is meaningful variation among California teachers' opinions about evaluation, the differences are not strongly correlated with basic characteristics of teachers or their districts. We examined several potential correlates: (i) district size, student demographics, SBAC test scores, and teacher workforce characteristics; (ii) features of the district's teacher evaluation program, as reflected in the local collective bargaining agreement; and (iii) teacher respondent characteristics and other opinions collected in the GDTFII survey. Across all these potential predictors of teachers' evaluation opinions, the correlation was rarely stronger than 0.10 (in absolute value) for any survey item. For example, teachers with less experience were more positive about evaluation, especially pre-tenure teachers, but experience explains at best 1-2 percent of the variation in teachers' opinions (correlations of a most 0.12). Besides the characteristics described in the next two paragraphs, the result for experience is typical of other correlates.

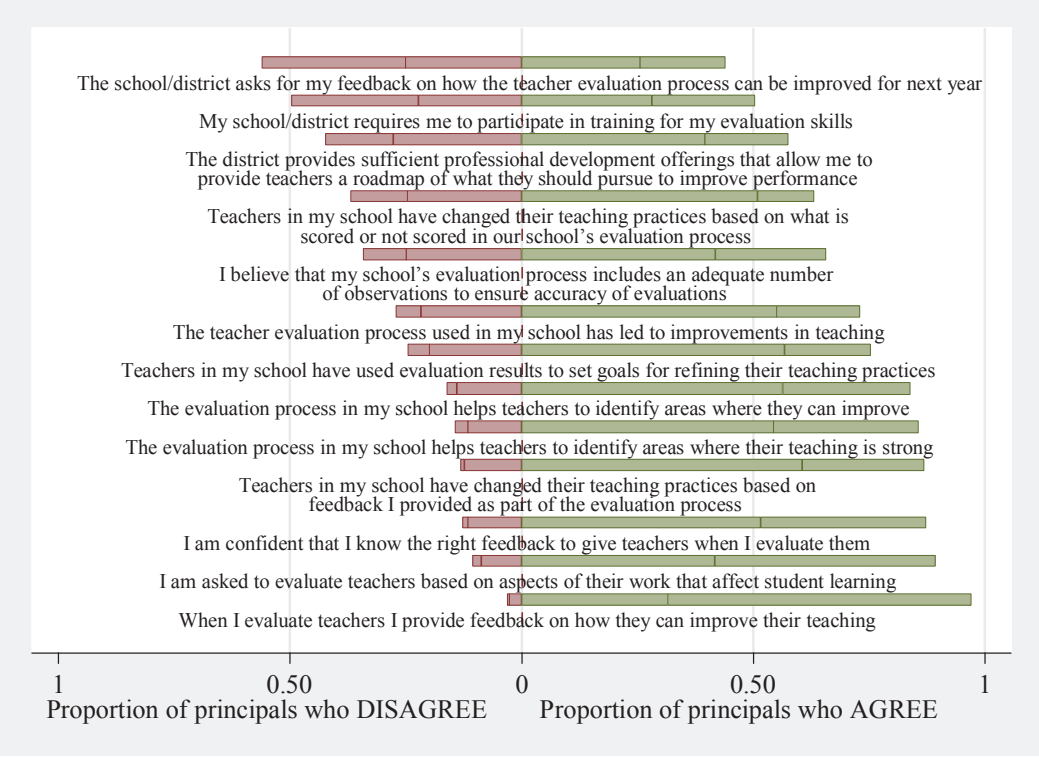

Figure 2b. Principals' assessment of their district's (school's) evaluation program

Note: Authors' calculations using RAND ATP/ASLP October 2017 Survey for GDTFII. The full text of the question stem is: "To what extent to you agree or disagree with the following statements about teacher evaluation?" Options were: strongly disagree, somewhat disagree, somewhat agree, and strongly agree. The bars above are divided by strongly and somewhat with the proportion somewhat close to the center line.

The strongest predictor of teacher's opinions about evaluation was her opinion of her own school's principal. In the survey teachers were asked a series of items like "The principal at my school communicates a clear vision for the school," "…sets high standards for student learning," "…is supportive and encouraging," and "I am pleased with the way my principal runs this school." While the strongest correlation, ratings of the principal explain only about 15 percent of the variation in opinions about the school's (district's) evaluation program (a correlation of about 0.40).

Teachers also had more positive opinions of evaluation when their district's evaluation program provides explicit supports to low scoring teachers, including formal assistance plans and additional classroom observations.[6] This is an example of explicitly linking resources for improvement to evaluation results, a topic which we will return to below. But again these features

[6] In this analysis of correlates, our measures of evaluation program features come from data collected from district collective bargaining agreements in 2015 (Strunk and Reardon 2010, Strunk et al. 2018). We thank Katharine Strunk for sharing these data.

explain only 2-4 percent of the variation in teacher's opinions of evaluation (maximum correlations of 0.20). Other features of the district's evaluation program were much weaker predictors of opinions, including the frequency of evaluation, number of classroom observations, number of rating categories, and whether the evaluation includes student test scores or other measure of achievement.

In general, principals' opinions were similar to teachers, though perhaps somewhat more optimistic about evaluation as an opportunity for teacher improvement. For example, 84 percent of principals, compared to 72 percent of teachers, agreed that the evaluation process helps teachers identify areas where they can improve.[7] Similarly, on the outcomes measure, 73 percent of principals agreed that the evaluation process has led to improvements in teaching.

California principals are quite confident about the feedback they give teachers after evaluation. Nearly all principals (97 percent) reported that they do in fact "…provide feedback on how [teachers] can improve…" as part of evaluation. Simply giving any feedback, good or bad, may be an easy bar to cross, yet nearly nine out of every ten principals (87 percent) agreed with the statement: *"I am confident that I know the right feedback to give teachers when I evaluate them."* This confidence seems at least somewhat at odds with the other beliefs of principals and teachers that suggest room for improvement in the feedback process.[8]

Multiple, Rubric-Based Classroom Observations

Modern rubric-based classroom observation programs have become a common feature of teacher evaluation, both in California and around the country. This section focuses on rubric-based observations—the first of four themes of contemporary teacher evaluation programs we will discuss in detail. We summarize (quasi-)experimental evidence which demonstrates that rubric-based observation programs can improve teaching effectiveness. The evidence is encouraging, but limited in that it comes from just two different district teacher evaluation programs.

A Brief Primer on Rubric-Based Observations

Classroom observations structured with detailed-rubrics are now quite common in public schools. In California, as examples, rubric-based observations are part of teacher evaluation programs in Fresno, Oakland, Poway, San Jan, and Soledad USDs. Additionally, the California

---

[7] This is a comparison of California teachers generally with California principals generally. We do not have a large sample of cases where a teacher and principal from the same school are surveyed.

[8] We also examined correlates of teachers' opinions about evaluation, among the potential correlates measures in the GDTFII survey data. The strongest correlation by far was a teacher's opinions about evaluation and her opinion of her school's principal (or administration more generally). The latter is measured with a series of items like "The principal at my school communicates a clear vision for the school," "…sets high standards for student learning," "…is supportive and encouraging," and "I am pleased with the way my principal runs this school." While the strongest correlation, ratings of the principal explain only about 15 percent of the variation in opinions about the school's (district's) evaluation program. Other characteristics and opinions with smaller correlations include: teacher experience, teacher reported characteristics of the students in the school, and overall job satisfaction.

Department of Education and Commission on Teacher Credentialing have produced a rubric paralleling the California Standards for the Teaching Profession, called the Continuum of Teaching Practice. Many other examples from California and beyond are readily available online.

What differentiates rubric-based observations from classroom observations typical of the past? The first differentiator, as "rubric-based" implies, is the use of a detailed rubric. The typical rubric covers several (sometimes dozens) of teaching practices, for example, "managing student behavior" and "questioning techniques." Figure 3 shows an example of one practice from the rubric used by the Los Angeles Unified School District, and one practice from the Poway Unified School District rubric. For each practice, the rubric describes what one would need to observe a teacher doing in the classroom to judge the teaching as effective or ineffective. Most rubrics describe four or more separate levels of effectiveness, sometimes attaching labels like "highly effective," "effective," "developing," and "ineffective"; or "accomplished," "effective," "approaching effective," and ineffective." These levels are translated into scores for evaluations.

A detailed rubric can be more than simply a guide for scoring. First, a rubric can create clear, shared expectations between teachers and administrators. Second, by describing what effective teaching "looks like" in practice, a well-designed rubric can also guide teachers' individual or collective efforts to improve.

A rubric is not the only important characteristic of a classroom observation program. Modern rubric-based observation programs often involve multiple observations of a given teacher over the course of a school year. The observers may be either the school principal and other administrators, or a specialized evaluator who is or recently was a teacher. Different rubric-based programs have different requirements for pre- and post-observation conversations between the teacher and the observer.

Can Rubric-Based Observations Improve Teacher Effectiveness?

Can rubric-based classroom observation programs improve teachers' effectiveness in the work of teaching? There is persuasive evidence from programs in Cincinnati and Chicago that, yes, rubric-based observations programs *can* improve teaching. The results are encouraging. We should, however, exercise caution in predicting positive results wherever rubric-based observations are used; we discuss some considerations after reviewing the two cases.

The first case comes from the Cincinnati Public Schools and its long-standing multiple-observation rubric-based Teacher Evaluation System (TES). The encouraging research evidence is that, in short, Cincinnati's teachers became more effective teachers as a result of their participation in TES. This result comes from a quasi-experimental analysis of historical data by Taylor and Tyler (2012a, 2012b) where teaching effectiveness is measured by teachers' contributions to student math achievement test scores.

Standard 3: Delivery of Instruction

Component 3b: Using Questioning and Discussion Techniques

Effective teachers design questions that provide cognitive challenge and engineer discussions among students to ensure all students participate. The highly effective teacher designs instruction that provides opportunities for students to develop their own cognitively challenging questions and to engage in various types of student-to-student discussions.

Element	Ineffective	Developing	Effective	Highly Effective
3b1. Quality and Purpose of Questions *Questions are designed to challenge students and elicit high-level thinking* **CO**	Teacher's questions do not invite a thoughtful response or are not relevant. Questions do not reveal student understanding about the content/concept or text under discussion, or are not comprehensible to most students.	Teacher's questions are a combination of both high and low quality, or delivered in rapid succession. Only some questions invite a thoughtful response that reveals student understanding about the content/concept or text under discussion. Teacher differentiates questions to make them comprehensible for some students.	Teacher's questions require rigorous student thinking. Most questions invite and reveal student understanding about the content/concept or text under discussion. Teacher differentiates questions to make learning comprehensible for student subgroups.	Teacher's questions require rigorous student thinking and invite students to demonstrate understanding through reasoning. Students themselves formulate questions to advance their understanding about the content/concept or text under discussion. Teacher differentiates questions to make learning comprehensible for all students in the class.
3b2. Discussion Techniques and Student Participation *Techniques are used to ensure that all students share their thinking around challenging questions* **CO**	Teacher makes no attempt to differentiate discussion. Interactions between the teacher and the students are characterized by the teacher generating the majority of questions and most answers.	Teacher makes some attempt to use differentiated strategies to engage all students in discussion with uneven results. Only some students participate in the discussion and/or the discussion is not intellectually challenging.	Teacher uses intentional, differentiated strategies to engage all students in discussion, attempting gradual release from teacher-directed to student-initiated conversation. Students participate in intellectually challenging discussions.	Teacher uses intentional, differentiated strategies to engage all students in intellectually challenging student-to-student discussions. Teacher creates conditions for students to assume considerable responsibility for the success of the discussions.

Figure 3a. Selection from LAUSD's teacher observation rubric

Note: Los Angeles Unified School District. May 2013 version. "CO" indicates scores based on classroom observation, as opposed to professional conversation or artifacts.

DOMAIN II: Accomplished Teachers Instruct ALL Students

(Assessed by self-reflection, observation, documentation, and conference)

ELEMENT 4: Using Assessment to Guide Instruction and Advance Student Learning

A. Uses a variety of formative assessment strategies during instruction to monitor learning outcomes for all students
B. Provides timely and consistent feedback to students to improve their learning
C. Ensures that all students know and understand the criteria for assessment of the learning outcomes
D. Facilitates opportunities for all students to assess, monitor, and take responsibility for their own learning

PERFORMANCE LEVEL DESCRIPTORS

Performance in these two columns results in the need for a guided improvement of practice goal in one or more areas for professional learning.		*Performance in these two columns results in a self-directed continuous improvement goal in one or more areas for professional learning.*	
Ineffective Practice	**Approaching Effective Practice**	**Effective Practice**	**Accomplished Practice**
A. Does not and/or ineffectively uses formative assessment strategies to monitor student learning	• Monitors learning through limited use of formative assessment strategies and may not elicit evidence of understanding from all students; formative assessments may not be purposeful	• Frequently monitors learning through a variety of purposeful formative assessment strategies; evidence of understanding of learning outcomes is elicited from all students	• Consistently monitors learning through a variety of purposeful formative assessment strategies; consistently elicits evidence of understanding of learning outcomes from all students to guide instructional and advance student learning
B. Does not provide feedback to students and/or provides feedback that is not specific, timely, or purposeful	• Occasionally provides feedback to students, but quality of feedback may be inconsistent and lack timeliness to guide their learning	• Frequently provides timely, accurate, and specific feedback that guides students in their learning	• Consistently provides a variety of ongoing feedback, from both teacher and peers, that is accurate, individualized, and advances learning
C. Does not and/or ineffectively communicates criteria for assessment of learning	• Occasionally communicates criteria for assessment; expectations for demonstrating mastery of the learning outcomes may be unclear	• Ensures that students are aware of the assessment criteria and that they understand the expectations for demonstrating mastery of the learning outcomes	• Consistently ensures that students clearly understand the expectations for demonstrating mastery of the learning outcomes; there is evidence that students have contributed to the assessment criteria
D. Opportunities for students to assess and monitor learning are not evident and/or students may not be held responsible for their own learning	• Occasionally provides opportunities for students to monitor and/or assess their own learning; some students may be held responsible	• Frequently facilitates opportunities for students to monitor and assess their own learning; all students are held responsible	• Consistently facilitates opportunities for students and peer groups to monitor, assess, and reflect on their own learning; students may initiate opportunities for self and peer assessment

Figure 3b. Selection from Poway USD's teacher observation rubric

Note: Poway Unified School District. July 2013 version.

Cincinnati's Teacher Evaluation System—designed collaboratively by the district and local union, and launched in 2000-01—has several key features. Each teacher's year-long evaluation involves multiple, typically 4-6, classroom observations structured and scored using a detailed rubric based on Charlotte Danielson's *Framework for Teaching* (1996). LAUSD's rubric is also based on Danielson's framework. Observations are conducted by a school administrator. Observers provide written feedback after at least one observation, and meet with the teacher at least once to discuss the results.[1]

Some readers will recognize Cincinnati's TES as an example of a peer assistance and review program. In that regard, Cincinnati's program shares features with the peer assistance and review programs implemented by some California districts around the same time. Among the five districts we highlight in this paper Poway and San Juan are exemplars of peer assistance and review.

Taylor and Tyler (2012a, 2012b) used historical data to track the effectiveness of individual teachers over many school years, and then to test whether effectiveness changed, on average, the year a teacher participated in TES. During the years the researchers studied, Cincinnati teachers were only evaluated periodically; experienced teachers were only evaluated every five years. The periodicity allowed the researchers to measure teaching effectiveness in the years before evaluation, the year of evaluation, and the years after evaluation. Comparing each teacher only to her own prior performance, the study shows an increase in average teacher effectiveness in the year of evaluation and that effectiveness continues at the new higher level in the years that follow evaluation.

That pattern of change in teachers' effectiveness over the years is important. One common hypothesis about evaluation is that employees simply work harder while they are being scored. If that had been the case in Cincinnati we would have seen an improvement in the year of evaluation, but then a return to the prior lower effectiveness level. The actual results from Cincinnati are consistent with a different hypothesis: that teachers learned or changed something during their evaluation year which improved their teaching effectiveness in a permanent way. The most straightforward explanations for these results lie in the key features of Cincinnati's approach to evaluation: multiple rubric-based classroom observations followed by feedback, often from peer evaluators.[2]

The second case comes from the Chicago Public Schools and a pilot of a new multiple-observation rubric-based evaluation program called the Excellence in Teaching Project (EITP). The results from a random-assignment field experiment in Chicago are similar to the results from Cincinnati: teachers' effectiveness improved as a result of their evaluation. Steinberg and Sartain

---

[1] We can only briefly highlight key features of TES in this paper, and some features have changed in recent years. For more information about TES see https://www.cps-k12.org/about-cps/employment/tes, as well Holtzapple (2003), Milanowski (2004), Milanowski, Kimball, and White (2004), Kane, Taylor, Tyler, and Wooten (2011), and Taylor and Tyler (2012a).

[2] These benefits may well be specific to rubric-based evaluation programs. As a contrasting example, Goldhaber and Anthony (2007) study the National Board of Professional Teaching Standards evaluation process. They do not find evidence that participating in the NBPTS process improves teacher effectiveness.

Poway Unified School District

Teacher Professional Learning and Effectiveness System

Key components. Initial self-assessment and goal setting. Multiple, rubric-based classroom observations. Post-observation conferences. Teacher self-reflection throughout the year, and an end-of-year reflection submitted to the administrator before the final evaluation rating is determined. Evaluation results linked to peer assistance and other resources for improvement.

Frequency. Non-tenured teachers are evaluated twice per year: once in January and once in April or May. Tenured teachers with a history of satisfactory evaluations are evaluated every other year. Tenured teachers with an unsatisfactory evaluation in the year prior are evaluated at least once a year until a satisfactory evaluation is achieved. Tenured teachers with 10+ years of Poway experience are typically evaluated every three years.

Classroom observations. Classroom observations are structured and scored using the Continuum of Teaching Standards rubric, based on Charlotte Danielson's *Framework for Teaching* and the California Standards for the Teaching Profession. Each practice is scored on a four point scale: ineffective practice, approaching effective practice, effective practice, or accomplished practice. Observations are conducted by a school administrator. Each teacher is observed twice per evaluation period for a minimum of 60 minutes total. Results shared within three days during a post-observation conference.

Final evaluation ratings. Each teacher receives an overall rating of ineffective or effective.

Poway Professional Assistance Program. PPAP provides support for new teachers. PPAP's Teacher Consultants (TCs) both assess performance and provide one-on-one support to new teachers. Each TC serves for three years and must have at least five years of teaching experience. TC responsibilities include instructional materials, lesson planning, explaining curriculum, classroom observations, providing feedback, and modeling demonstration lessons.

Permanent Teacher Intervention Program. PTIP provides support for experienced teachers with an evaluation of unsatisfactory. In level one, a school administrator conducts formal observations, and a TC assists the teacher with practices such as lesson planning and classroom management. In level two, a TC continues to assist the teacher, but an evaluation team conducts formal observations. The team is comprised of a school administrator, district administrator, and a third individual selected by the district.

Consequences. A Peer Board of Review, composed of three union representatives and two district employees, makes final recommendations concerning contract renewal for new teachers and dismissal of permanent teachers.

Alternative evaluation. Teachers with five years of experience and a history of satisfactory evaluations may participate in a modified, self-directed professional growth cycle instead of evaluation. Teachers establish their own goals and provide a written reports to the principal.

Sources: Poway Unified School District and Poway Federation of Teachers (n.d., 2013, 2016), Poway Unified School District personal communication (March 9, 2018)

(2015) report results of the experiment, again measuring effectiveness by teachers' contributions to student test scores.

Chicago's EITP, like Cincinnati's TES, used multiple rubric-based classroom observations to evaluate teachers. Like Cincinnati, Chicago chose a rubric based on the *Framework for Teaching*, and observations were followed by a feedback conversation between teacher and evaluator. In Chicago, by contrast, school administrators were the only evaluators.

Steinberg and Sartain (2015) compare teachers in two groups of schools: 44 "treatment" schools randomly selected to begin EITP in 2008-09 and 49 "control" schools. During the 2008-09 school year teaching effectiveness was, on average, higher in the treatment schools than in control schools (the difference is statistically significant for reading tests but not math).

The higher effectiveness of treatment teachers, when compared to control teachers, continued in the years after 2008-09. That gap continued despite the fact that the "control" schools, at least nominally, began EITP in 2009-10. However, as Steinberg and Sartain (2015) report, institutional and financial support for EITP began to decline in early 2009 and the program was ended in the summer of 2010; the "control" schools who nominally began EITP in 2009-10 likely did not fully implement the new program. The results from 2009-10 are an example of why we should be skeptical of concluding that rubric-based observation programs will always improve teaching effectiveness. Plausible explanations for the encouraging results, like Cincinnati and Chicago, all involve teachers and schools investing meaningful effort in using the evaluation program as an opportunity for improvement.

One additional result on teacher turnover from the Chicago experiment is worth briefly mentioning. Sartain and Steinberg (2016) also find that EITP evaluations induced low-performing teachers to leave their schools at higher rates. This result parallels recent research from Houston where the introduction of a new evaluation system districtwide—an evaluation based on both classroom observations and student test scores—also resulted in higher turnover rates for low-performing teachers (Cullen, Koedel, and Parsons 2016). While such turnover is not a mechanism for improving the effectiveness of individual teachers, it can contribute to an improvement in average teaching effectiveness if the exiting teachers are replaced by relatively higher-performing teachers.

While the Cincinnati and Chicago cases are currently the best direct evidence—on the causal effect of multiple-observation rubric-based evaluation—there is other indirect evidence which is important to consider. Any true positive effect of rubric-based observations will presumably be stronger the more accurate or meaningful are the observation results (or the more accurate or meaningful teachers perceive them to be). Researchers are beginning to build understanding of, for example, the reliability of observation scores in different evaluation uses, and the extent to which non-random student-to-teacher assignments affect observation scores (Hill, Charalambous, and Kraft 2012, Ho and Kane 2013, Cohen and Goldhaber 2016, Gill et al. 2016, Steinberg and Garrett 2016, Bacher-Hicks et al. 2017). We have a reasonable understanding of the relationship between observation scores and teachers' contributions to their students' test scores (Kane et al. 2011, Kane and Staiger 2012), but still lack evidence of the relationship

between teachers' observation scores and longer-run student outcomes like college attendance and labor market success.

Cautions and Concerns about Classroom Observations

Before moving on to the next of our four themes, we briefly discuss some cautions and often-raised concerns about classroom observations.

One concern frequently raised about multiple-observation rubric-based evaluations is the substantial time and effort required. If evaluators do not have or devote sufficient time to observations—as many school principals feel when they are given primary or sole responsibility for conducting observations—the accuracy and meaningfulness of scores and feedback might understandably suffer (Kraft and Gilmour 2016). As mentioned above, this may explain some of the results in Chicago's experiment. By contrast, the Cincinnati Public Schools devoted substantial resources to TES, namely full-time peer evaluators who otherwise would be teaching their own students. Later in the paper we return to the issue of costs.

A second concern, or at least caution, is that classroom observation scores might reflect more than a teachers' performance observed in the classroom. First, when scoring a teacher, an observer may draw on what she already knows or believes about the teachers. We return to this topic later when discussing subjectivity in evaluations. Second, differences in observation scores may reflect differences in students in the observed class not differences in teaching; there is some new research demonstrating this (e.g., Gill, Shoji, Coen, and Place 2016, Steinberg and Garrett 2016), but there is still much to learn about when this arises and how to correct for it if at all. If observation scores do partly reflect students and not teachers then the usefulness of observations for teacher development, and the incentives of evaluation, will be muted.

A final concern sometimes raised about classroom observation scores is that "everyone passes." This if often true if we focus on only whether a teacher "passes" or "fails" on her final overall (or summative) evaluation score. But focusing on the final overall score ignores potentially rich micro-data that are gathered in multiple observations each producing a score for several (dozens) of practices. These rich micro-data are (likely) useful for teacher development purposes, even if they do not become part of a teacher's final official evaluation score.[3]

[3] A typical example of moving from micro-data to final score is as follows: The dozens of scores created in multiple observations for several practices are first averaged (perhaps with some weights). Then often the overall average is rounded off to the nearest integer score before determining "pass/fail" further lowering the implied threshold (e.g., an overall average score of 1.5 would be rounded to 2 which might well be passing on a 1-4 scale). In one example (Papay et al. 2017 Table I) the overall average score was 3.66 on a 1-5 scale with a standard deviation of 0.68, even though very few overall scores were below the failing cutoff of 3. Continuing the example, 41 percent of teachers had at least one skill scored below the passing threshold of 3, and the average number of skills below 3 was 2.4 out of 19.

Connecting Individual Evaluation Results to Resources and Strategies for Improvement

One feature of an evaluation program is the resources and strategies for improvement that are available to teachers after they have been scored. What, for example, should a teacher do next if he scores low on "asking questions in class" and he wants to improve? The answer is not always an explicit or intentional part of an evaluation program's design, but it can be. One common approach is to have the school principal (or other evaluator) provide some strategies for improvement in a post-observation conversation.

In this section we give examples of evaluation programs, in California and beyond, where districts and states are being intentional and innovative about this feature. The first example is an approach used in Long Beach to make professional development resources easier to access and connect with evaluation results. We then summarize promising, though again sparse, research evidence on pairing teachers to provide support for improvement. In an example from Tennessee teachers are paired with a colleague who works in the same school. In another example teachers are assisted by a coach online.

Long Beach Unified's myPD Program

The first example comes from the Long Beach Unified School District and its myPD program—a program distinct from but related to LBUSD's teacher evaluation program. myPD software helps individual teachers create and carryout a personalized development plan, and update that plan over time. First, teachers decide which teaching practice(s) will be the focus of their improvement work. Specifically, teachers select practices from the California Standards for the Teaching Profession (CSTP). The software aids this decision by combining and analyzing information about the teachers' current performance, including the teacher's own self-assessments, student achievement data, evaluation results, and other sources. Second, the software suggests specific resources to the teacher, including traditional face-to-face PD courses, videos of other Long Beach teachers teaching, self-paced online courses, communities of teachers focused on the same practices, and other resources.

The CSTP make the link between myPD's resources and teacher evaluation results. Like myPD, LBUSD teacher evaluation is organized around the California Standards for the Teaching Profession (CSTP); teachers are given a separate rating for each of the six CSTP standards. One advantage of myPD is that it helps teachers make easy connections between their evaluation results and the several professional development resources described in the previous paragraph. The extent to which myPD is successful in its goals for improving teaching effectiveness will be borne out over time, but the thoughtful design is promising.

Long Beach Unified School District
Certified Personnel Evaluation and *myPD*

Key components. Based on the California Standards for the Teaching Profession. Multiple observations. Post-observation and end of year conferences. Evaluation results linked to professional development resources.

Frequency. Non-tenured teachers are evaluated every year. Tenured teachers are evaluated every other year. Teachers with 10+ years of experience and a history of satisfactory evaluations are evaluated every five years.

Classroom observations. Observations are structured and scored following the California Standards for the Teaching Profession (CSTP). The CSTP includes six standards and 5-7 specific practices for each standard. Each of the six standards are scored from 1-4, corresponding to unsatisfactory, developing, effective, or distinguished. Observations are conducted by school administrators. Each non-tenured teacher is observed three times per evaluation. Each tenured teacher is observed 1-3 times per evaluation.

After the observation. Post-observation conference to discuss strengths and weaknesses, and to provide recommendations for improvement as needed. If the teacher receives an unsatisfactory observation rating, she may request to be observed again by the evaluator and an administrator certified in the teacher's assignment area.

Final evaluation ratings. Each teacher receives an overall rating of unsatisfactory, developing, effective, and distinguished.

Linking evaluation to resources for improvement. Teachers access a curated menu of personalized resources through myPD. Available resources include more-traditional face-to-face professional development courses, self-paced online courses, and videos of other Long Beach teachers demonstrating effective practices. Resources are suggested to individual teachers are based on the teacher's goals, self-assessment, evaluation results, student data, among other things.

Support for new teachers. Full time mentors coach and support new teachers for two years. Teachers are placed in small mentoring groups based on teaching assignment area. Mentor conducts observations, provides feedback, models demonstration lessons, and hosts monthly seminars on topics such as lesson planning, classroom management, and support for English Learners.

Sources: Long Beach Unified School District (2014), Long Beach Unified School District and Teachers Association of Long Beach (n.d.)

Teacher Partnerships Linking Evaluation and Improvement Efforts

A second example, and a quite different approach, comes from the Tennessee Department of Education and its Instructional Partnership Initiative (IPI). The resources and strategies for improvement come in the form of a "partnership" pairing of two teachers who work in the same school. Partnerships are matched based on evaluation scores from rubric-based classroom observations; each pair includes one teacher who has scored particularly low in one or more of the rubric's practices, and a matched partner teacher who scores highly in the same practices.[4] The state provides each school a list of proposed one-to-one pairings, which principals are free to adjust. If a school principal wants to use IPI in her school, she introduces the partnership and gives them a charge to work tougher on improving each other's teaching, especially the practices used to match the partnership. Suggested partnership activities include discussing each other's evaluation results, observing each other teaching, discussing strategies for improvement, and following up on commitments and goals. One key motivation for the IPI approach is to maximize the individualization in supports provided for teachers' improvement efforts.

Do such teacher partnerships motivated and formed by evaluation results help improve teachers' effectiveness? Papay, Taylor, Tyler, and Laski (2017) conducted a field experiment where randomly-selected treatment schools used the Instructional Partnership Initiative. (The experiment was part of a pilot of IPI in one Tennessee school district.) Participating in a partnership did improve the effectiveness, on average, of the teachers who had previously scored low in one or more practices and had thus been matched with a partner by IPI. The experiment's primary measure of teaching effectiveness was teachers' contributions to student achievement test scores in reading and math. Teachers' rubric-based observation scores also improved in the specific practices where there was a strength-to-weakness match between the pair of teachers. Finally, teachers in IPI treatment schools reported more favorable opinions of the evaluation program.[5]

IPI's coworker partnerships are one approach which intentionally links evaluation results and one-on-one individualized support for developing teachers, but there are other similar approaches which use different people in the "partner" role. The first example is well known: Peer assistance and review programs, like those in Poway USD and San Juan USD, in which the peer teacher providing evaluation, feedback, and assistance is a formal job in the district. Indeed one-on-one individualized support may be an important cause of the positive effects of Cincinnati's TES described earlier. A second example is the MyTeachingPartner (MTP) service offered by the University of Virginia's Center for the Advanced Study of Teaching. MTP pairs each participating teacher with a trained MTP consultant. The foundation of MTP's classroom observations and

---

[4] More information can be found at http://team-tn.org/ipi/. In practice, "scored particularly low" means scoring less than 3 on a 1-5 scale where 3 is "At Expectations"; depending on the specific teaching practice, approximately 5-25% of teachers score less than 3 on the given practice.

[5] The Tennessee Department of Education and its academic research partners are conducting ongoing research about IPI. In the years since the IPI pilot experiment described above, Tennessee has made IPI partnership recommendations and program materials available to all schools in the state, but the state does not require that schools use the program. About one of every five or six schools uses IPI (Papay, Goldring, et al. 2017), suggesting there is much to learn about why many principals do not use the program.

coaching is a detailed rubric called the Classroom Assessment Scoring System (CLASS) which is applied to videos the teacher records. In a random-assignment field experiment, MTP improved the teaching effectiveness of participating teachers (see Allen, Pianta, Gregory, Mikami, and Lun 2011, and a review of several MTP studies in Kraft, Blazar, and Hogan 2018).[6]

Using Multiple Evaluation Measures, Including Subjective Evaluation

Many modern teacher evaluation programs incorporate multiple measures of performance. A teacher's overall evaluation score might incorporate rubric-based classroom observation scores, student survey results, and scores based on student test scores like "value added measures." Intuitively, each component measures teaching in a different way focusing on different tasks and responsibilities, and thus the combination should better, even if still incompletely, measure the diverse responsibilities and performance of a teacher.

There is, unfortunately, little if any direct evidence on whether "multiple measures" teacher evaluation programs improve teaching effectiveness.[7] There are no (quasi-)experiments comparing a multiple measures design to, say, evaluation with just classroom observations. However, next we do discuss some indirect evidence relevant to thinking about multiple measures evaluation designs. That indirect evidence comes from an experiment in which school principals were provided "value added scores" for their schools' teachers—a new measure added to existing evaluation measures.

In this section we also discuss subjective evaluation in the context of multiple measures programs, and highlight the example of subjective evaluation in LAUSD. Yet, again we have no evidence to share on whether subjective evaluation improves or hinders teacher effectiveness.

An Experiment with Teacher Value Added Scores

There is some indirect evidence relevant to thinking about "multiple measures" evaluation which comes from an experiment involving teacher value added scores. "Teacher value added" is a now-common, even if sometimes confusing, short hand for one particular measure of teacher performance which might be used in an evaluation program. Specifically, value added scores measure a teacher's contribution to her assigned students' test scores in math, language arts, and other subjects. Our goal in this paper is not to address all the pros and cons of using value added scores in teacher evaluation programs, but rather to focus on the results of one experiment.

The evidence we discuss here comes from a random-assignment field experiment in the New York City public schools in the 2007-08 school year, and analysis by Rockoff, Staiger, Kane,

---

[6] CASTL does describe MTP as a coaching program. We have largely avoided the word "coaching" in this paper simply to avoid confusion with the much larger set of programs and research on teacher coaching. The examples of IPI, PAR, and MTP are, in a sense, coaching programs, but what sets them apart for our purposes is the explicit and intentional connections to structured evaluation. Kraft, Blazar, and Hogan (2018) provide an updated review of teacher coaching programs generally.

[7] There is research on the design of multiple measure evaluation programs, and research demonstrating other reasons—reasons besides improving effectiveness—why a multiple measures approach might be preferable.

and Taylor (2012). Randomly selected treatment principals were given value added score reports for the teachers in their school, along with training on how to read and interpret the reports. At the time of the experiment, school principals had not previously seen value added scores for their teachers.

Principals' actions demonstrated that the value added reports provided new and useful information the principals did not know before. First, value added scores changed principals' subjective evaluations of their teachers. Principals' post-experiment subjective evaluations were more strongly correlated with value added scores in treatment schools compared to the positive correlation in control schools which did not receive the reports. This change is consistent with the stated motivation for multiple measures programs: value added scores measure an aspect of teachers' job responsibilities that principals did not previously fully incorporate. But we can also turn the result around. If the school district had only known the value added scores, and never asked for principals' subjective evaluations, the district would have missed out on important information the school principal knew.

Additionally, the changes in principals' subjective evaluations were not a simplistic or naive adoption of the value added reports. Principals did not, for example, simply replace their own prior ranking of teachers with the value added ranking. The more years a principal had worked with a teacher, the less the principal's rating of the teacher changed in response to the value added reports. Similarly, the wider the statistical confidence interval on a teacher's value added score, the less the principal's rating changed in response to that score.[8]

The second result is that value added scores changed teacher turnover. In schools that received value added reports, teachers with relatively low math value added scores were more likely to leave the school at the end of the year, perhaps of their own choice or at the principal's prompting or a mix. This parallels the results from Houston and Chicago mentioned above (Cullen, Koedel, and Parsons 2016, Sartain and Steinberg 2016).

Did the value added score reports improve teachers' effectiveness in their work? Student test scores, especially math scores, were higher in treatment schools the year after treatment principals received value added reports.[9] The improvement in student achievement suggests either an improvement in teacher effectiveness—through a change in practices or effort—or an improvement in the way the school is managed, or some combination of the two. A low value added score alone would not tell a teacher what he needed to do to improve, but could well prompt him to seek out other resources and make new efforts to improve.

Subjective Evaluation

LAUSD's teacher evaluation program adds a different measure: the principal's subjective assessment. This subjective evaluation is implicit in LAUSD's approach. In practice, the principal

[8] The reports provided to principals included a visual representation of the score and its confidence interval due to sampling. Intervals were smaller when more student scores were available to include in the value added score.
[9] Differential turnover, described in the previous paragraph, may explain part but not all of the difference in test scores (Rockoff et al. 2012).

has responsibility to make a final evaluation of "exceed," "meets," or "below standard performance" for each teacher. Principals are asked to consider all evaluation measures, including classroom observations and contributions to student outcomes, but there is no formulaic combination of the scores. A principal can choose different final ratings for two teachers with otherwise similar scores; that difference would reflect the principal's subjective assessment. Such subjective principal ratings are rare in modern multiple-measure teacher evaluation programs.

Adding subjective assessments can improve evaluation, relative to evaluation which relies only on formulaic "objective" measures.[10] A principal's or supervisor's assessment can incorporate information about all of a specific teacher's job responsibilities, including those responsibilities not captured by formulaic measures. By focusing on only a few responsibilities, as is often the case, formulaic measures create an incentive for teachers to (a) *distort effort*: focus more effort than they otherwise would on those few responsibilities, or (b) *manipulate measures*: take actions, like teaching to the test, which raise the formulaic score but do not actually reflect effective teaching. Subjective evaluation can help reduce these distortion and manipulation problems. Additionally, a principal's or supervisor's assessment can incorporate information about unanticipated factors or changes in a teacher's responsibilities which are difficult to design into formulaic measures.

There are, nevertheless, reasons to be cautious about subjective evaluations. Subjective evaluation can, as we said above, bring into evaluation information which is otherwise difficult to measure; but that added information may or may not be relevant to a teacher's job. A principal's subjective evaluation may be influenced, intentionally or unintentionally, by personal biases for or against an individual teacher. This is a potential cost to weigh against the benefits. The potential cost or risk is, however, often mitigated by other features of the evaluation program. In Los Angeles, for example, principals are not asked to simply give their opinion, but instead are given specific guidelines within which to make their relatively-subjective assessment. Additionally, the more objective measures in LAUSD's evaluation program provide an opportunity to check for subjective evaluations which are far outside from the range expected under those guidelines.

Additional Research Evidence on Multiple Measures Approaches

Two final pieces of evidence worth noting come from the Measures of Effective Teaching (MET) study. First, Ho and Kane (2013), studying one MET district, collected rubric-based observation scores from each teacher's own principal and then asked other principals in the district to score the same video. One hypothesis was that a teacher's own principal might score the teacher differently, for better or worse, because the principal incorporates information from other interactions with the teacher. Contrary to that hypothesis, rubric scores from the teacher's own principal were quite similar to scores from other district principals. If evaluation program

[10] Perfectly "subjective" and perfectly "objective" are theoretical extremes on a continuum. Most real evaluation measures lie somewhere in between. Classroom observations and student surveys include scope for subjectivity, but are much less subjective than overall principal assessments like those in LAUSD's program. Value added scores are quite formulaic and thus "objective" after the scoring process is designed and reduced to computer code.

designers want subjective evaluations, or assessments of other responsibilities, the program should ask for those evaluations explicitly.

Los Angeles Unified School District
Educator Development and Support: Teachers (EDST)

Key components. Initial self-assessment and goal setting. Multiple, rubric-based classroom observations. Pre- and post-observation conferences. Optional mid-year and end-of-year teacher reflections. Principal subjective judgment in final ratings.

Frequency. Non-tenured teachers are evaluated every school year. Tenured teachers with less than ten years of experience are evaluated every other year. Teachers with 10+ years of experience and a history of satisfactory evaluations may extend the interval to up to five years.

Classroom observations. Observations are structured and scored using the LAUSD Teaching and Learning Framework (TLF), based on Charlotte Danielson's *Framework for Teaching*. The rubric includes over 50 specific teaching practices, but each teacher is evaluated on seven practices: three chosen by the district, three chosen by the teacher, and one jointly chosen by the teacher and evaluator. Each focus element is scored on a three-point scale: ineffective, developing, or effective. Observations are conducted by the principal or a principal appointee. Each teacher is observed twice during the evaluation year: One formal unannounced observation covering an entire lesson. Plus one 15-20 minute "growth plan visit" where the administrator collects evidence and provides feedback on a teacher-selected practice.

Before and after observations. Pre-observation conferences to provide feedback on the lesson plan prior to the observation. Post-observation conferences include feedback, reflection, and review of student work samples generated during the lesson; followed by discussion of next steps in the teacher's professional growth and development.

Final evaluation ratings. At the conclusion of the evaluation, the school principal rates each teacher as one of three ratings: exceeds standard performance, meets standard performance, or below standard performance. There is no formula which determines this final rating. The school principal makes a final judgment, and is instructed to consider the teacher's classroom observations, contributions to student outcomes, progress toward planning objectives, and other professional responsibilities.*

Sources: Los Angeles Unified School District (2016, 2017), Los Angeles Unified School District personal communication (March 7, 2018)

* The principal must sign and take responsibility for all evaluations, but may ask an assistant principal or instructional specialist to help conduct evaluations.

Second, the MET study shows how using multiple measures increases the reliability (reduces the volatility) of a teacher's overall evaluation score (Kane and Staiger 2011). Unnecessary volatility in evaluation scores mutes the incentives of otherwise well-designed evaluation programs. The more unreliable a score is, the less confident a teacher can be that a change in her effort or practices will be captured by, and rewarded by, the evaluation system. (We return to this topic in the next section.) Thus using multiple measures may help generate improvements in teacher effectiveness by strengthening the incentives to improve.

Attaching Consequences, Positive or Negative, to Evaluation Results

A final feature of an evaluation program is the consequences for the teacher of receiving low or high scores. The consequences can be positive: public recognition, a bonus, a salary increase, a promotion or new responsibilities. Among other California districts, Long Beach, San Bernardino, and San Francisco USDs provide bonus compensation based, at least in part, on ratings from the teacher evaluation program. In Los Angeles, teachers with sufficiently high evaluation ratings can apply to be a mentor teacher, a role which includes additional pay.

Of course the consequences can also be negative: termination, denial or postponement of tenure, extra or more-intensive evaluation. Among the districts we spotlight in this paper, in San Juan and San Jose USDs teachers with low evaluation ratings are provided peer assistance, but afterward can be dismissed for a failing evaluation rating. The San Bernardino USD withholds a teacher's regular advancement on the district salary schedule when the teacher's evaluation ratings are repeatedly failing.

In the remainder of this section we summarize evidence on this question: Can the consequences, positive or negative, attached to evaluation results improve or worsen teachers' effectiveness in the work of teaching? We focus on the two most prominent consequences in policy discussions: pay for performance and termination or denial of tenure status.

Pay for Performance Attached to Evaluation

The evidence on pay for performance is decidedly mixed. There are several high-quality experimental and quasi-experimental studies that demonstrate, yes, pay attached to evaluation results can improve teachers' performance. A very recent example is the federal government's 2010 Teacher Incentive Fund grants. Chiang et al. (2017) report improvements in student achievement in schools randomly assigned to participate in the TIF pay for performance programs; TIF paid bonuses based on student test score growth and classroom observation scores. But there are also several high-quality (quasi-)experimental studies that find no effect of pay for performance for teachers. One notable example is a random assignment experiment in Tennessee which provided bonuses as large as $15,000 based on teacher performance measured by contributions to student test scores. Springer et al. (2010) report no improvement in student achievement as a result of the bonuses. Neal (2011) provides a review of several other (quasi-) experimental research studies of pay for performance programs for teachers.

San Jose Unified School District
Teacher Evaluation System (TES)

Key components. Based on the California Standards for the Teaching Profession. Multiple observations and reflective conversations. Narrative-based evaluation feedback instead of ratings or scores. Teacher Quality Panel that oversees evaluation.

Frequency. Non-tenured teachers are evaluated every year. Tenured teachers are evaluated every three years.*

Classroom observations. Classroom observations are structured and scored using five standards adapted from the CSTP. Each standard is accompanied by a list of teacher practices that do and do not exemplify each standard. Evaluators use this list to provide teachers narrative feedback on each standard, instead of a rating or score. Non-tenured teachers are observed twice: once by an administrator, and once by a consulting teacher. Tenured teachers are observed twice by an administrator in the fall; and a third time in the spring if fall observations are not scored meeting standards. Evaluators also make drop-in visits during the year to all teachers. All observations last at least 45 minutes, are not required to be announced in advance, and are accompanied by a reflective conversation.

Final evaluation ratings. At the conclusion of the evaluation, the administrator rates each tenured teacher as meets standards or does not meet standards. For non-tenured teachers, the administrator and consulting teacher each make separate ratings and submit separate reports. The Teacher Quality Panel (TQP), a group comprised of three teachers and three administrators, reviews all unsatisfactory evaluations to ensure evaluation procedures were followed according to protocol.

Consequences. The Teacher Quality Panel (TQP) determines next steps for non-tenured teachers. The TQP also makes recommendations regarding permanent status, remediation, and dismissal. Teachers with a performance evaluation of "does not meet standards" are placed in the Teacher Assistance Program (TAP).

Teacher Assistance Program. TAP participants are evaluated by an administrator and a consulting teacher, and receive additional support from a mentor. Each TAP teacher's status is reevaluated after 90 days of participation: the teacher remains in TAP (typically another 90 days), exits TAP, or is recommended for termination.

Consulting teachers. Consulting teachers are experienced teachers released full time from classroom teaching to evaluate approximately 30 non-tenured teachers. Tenured teachers can also request a consulting teacher if they receive an unsatisfactory evaluation.

Source: San Jose Unified School District (2015), San Jose Unified School District & San Jose Teachers Association (2016), San Jose Unified School District personal communication (March 9, 2018)

* In years when tenured teachers are not formally evaluated, they are informally evaluated during a professional growth cycle program.

Why are the results mixed? "Pay for performance" is a broad category of evaluation consequences; comparing two pay for performance programs may not be an apples-to-apples comparison. There is not space in this paper to discuss all of the details which might make different pay for performance designs different in their effect. For readers interested in more we suggest starting with Neal (2011). In the next few paragraphs we highlight some illustrative examples.

Pay for Performance: Team Bonuses

The first example highlights the difference between individual and team incentives. A 2007 experiment in New York City schools, studied by Marsh et al. (2012) and Goodman and Turner (2013) among others, paid bonuses to schools—a team of teachers—based primarily on student achievement scores. The bonus program had little if any effect on student outcomes in the average school. However, as Goodman and Turner (2013) show, the bonuses did increase teacher performance in relatively small schools, that is, schools with relatively few teachers. As the team gets larger, any one individual teacher's actions have less influence on the team's evaluation, and thus less influence on the bonus that individual teacher will receive. If the goal of a pay for performance program is to incentivize teachers, those incentives are muted in team (school) bonus programs and even more muted as the size of the team grows.

The team versus individual consideration is important for all evaluation programs, even if there is no bonus money attached. Some teacher evaluation programs, for example, score teachers based in part on student test scores in grades and subjects the teacher does not herself teach.

Pay for Performance: Uncertainty

The second example highlights the way uncertainty can get in the way of otherwise-well-designed incentive programs. For evaluation incentives to work well, the teacher must be reasonably certain that if she changes her behavior (e.g., increases her effort, learns and applies a new teaching practice) then her evaluation scores will go up. One simple, but important, example of uncertainty is the statistical uncertainty in test-based value-added scores for individual teachers. In studying Houston's ASPIRE program, Brehm, Imberman, and Lovenheim (2017) show quasi-experimental evidence which suggests value-added noise mutes ASPIRE's incentives for teachers.

Pay for Performance: Size of Bonuses

A third example highlights the size (amount) of the bonus or salary increase. A simple, perhaps obvious, reason a bonus program might not affect teacher performance is that the bonus is too small to elicit teachers' attention or effort.[11] Washington, DC's IMPACT evaluation program, however, is an example of notably large bonus amounts, and was the focus of a quasi-

[11] Even if a small bonus does in fact affect performance, the change in performance may be too small to detect statistically given limited power.

experimental study by Dee and Wyckoff (2015).[12] IMPACT scores teachers in four categories: ineffective, minimally effective, effective, and highly effective. Teachers rated "highly effective" receive a bonus of between $5,000-25,000, with the largest bonuses going to those teaching tested grades or high-need subjects in high-poverty schools. That bonus is larger than all studies reviewed by Neal (2011). Even more notably, teachers rated "highly effective" two consecutive years receive a permanent increase in salary, which could be as large as a 29 percent increase in earnings over 15 years.[13]

Dee and Wyckoff (2015) focus their attention on the teachers who had been rated "highly effective" for the first time but barely so. These teachers' job performance improved substantially as a result of IMPACT's financial incentives. Why focus on these teachers? They had the strongest incentive to increase their effort or effectiveness or both; a second consecutive "highly effective" was not certain but would bring a large salary increase. By contrast, top teachers—those scoring far above the cutoff between "effective" and "highly effective"—could be confident they would rate "highly effective" twice and thus had little or no incentive to change their practices as a result of IMPACT. Incentives were also muted, by similar logic, for teachers far below the "highly effective" cutoff. These contrasting incentives help researchers demonstrate the influence of strong (weak) incentives, but they are also a reminder that discontinuous bonus rules (e.g., a cutoff in evaluation scores or requiring two consecutive years) can create unintended consequences.

Tenure Decisions as a Consequence of Evaluation

We now switch focus to a common negative consequence attached to low scores: termination, or denial or postponement of tenure. These consequences are often a stated feature of teacher evaluation programs, even if the use of the consequence is uncommon. Again Washington, DC's IMPACT is a notable counter example. Teachers rated "ineffective" are immediately dismissed, and so are teachers rated "minimally effective" in two consecutive years. Dee and Wyckoff (2015) calculate that 3.8 percent of all DCPS teachers were dismissed by these IMPACT rules (during the school years 2010-11 and 2011-12).

As with the financial incentives, Dee and Wyckoff (2015) focus their attention on the teachers who faced the strongest dismissal threat; teachers who had been rated "minimally effective" for the first time but barely missed reaching the "effective" category. These teachers' job performance improved substantially as a result of IMPACT's dismissal threat, much like their higher-performing colleagues who improved because of financial incentives. Put differently, the IMPACT case is evidence that negative consequences—the threat of dismissal—can lead to improvements in teacher job performance. Additionally, 30 percent of teachers rated "minimally

[12] The description of IMPACT in this paper is for 2009-10 through 2010-12, the period studied by Dee and Wyckoff (2015).

[13] Functionally teachers were given credit for 3-5 additional years of experience and a master's degree when determining where they were on the district's salary schedule.

effective" for the first time voluntarily quit, and the probability of voluntarily leaving rose as a teacher's score got close to "ineffective" (Dee and Wyckoff 2015).

As mentioned earlier in the paper, such turnover is not a mechanism for improving the effectiveness of individual teachers, but it can contribute to an improvement in average teaching effectiveness if the exiting teachers are replaced by relatively higher-performing teachers. Adnot, Dee, Katz, and Wyckoff (2017) focus on measuring this effects of replacing teachers who left because of IMPACT, and find that student achievement did increase as a result of IMPACT-induced turnover.

One other study of IMPACT is important to note. Adnot (2016), like Dee and Wyckoff (2015), focuses on teachers who had barely missed being rated "effective" for the first time; teachers for whom the threat of dismissal was strongest if they did not improve. Adnot (2016) also finds that IMPACT improved these teachers' performance, as measured by rubric-based classroom observations. Those improvements, however, appear to be concentrated in a subset of the teaching practices measured in the rubric, specifically practices where the rubric itself provides easier-to-follow descriptions of what to do to score higher. These results are an example of a common unintended consequence of evaluation programs: evaluatees focus more effort on aspects of their job which more readily increase their evaluation scores, sometimes to the detriment of other important responsibilities. We should thus be cautious about interpreting IMPACT's effects on evaluation scores as improvements in all important aspects of teaching.

Our final case of evaluation consequences is tenure decisions. Denying (deferring) tenure is quite similar to dismissal, but tenure decisions are more commonly listed as consequences in formal teacher evaluation programs. New York City's recent experience provides informative evidence on tenure.

Beginning in 2009-10 the NYC DOE changed how tenure decisions would be informed by evaluation results (Loeb, Miller, and Wyckoff 2015 provide an excellent summary). In short, informed by a teacher's evaluation scores, the district provided a tenure recommendation to her principal; the principal could decide against that recommendation but would have to provide a written rational to the superintendent.[14] Both the district and principal had access to each teacher's classroom observation scores, prior principal subjective evaluations, and other information long used by NYC, plus new teacher-value added scores.

The NYC tenure changes had clear effects. In the years leading up to 2009-10, nearly 19 of 20 teachers were approved for tenure. In the years after, that fell to just 11 of 20. Most of this change was teachers whose probationary period was extended; tenure denials increased from 2 percent to 3 percent (Loeb, Miller, and Wyckoff 2015). In summary, the NYC change made tenure

[14] In 2009-10 the district recommendations were explicitly "tenure in doubt" and "tenure likely." In 2010-11 these changed to measure-specific recommendations like "low value add" is an "area of concern" or "high value add" is "notable performance." In 2011-12 the recommendations changed again to four options: "highly effective" and "effective" (favoring tenure), "developing" (favoring extension/deferral), or "ineffective" (favoring denial).

decisions a meaningful consequence—a real threat—much like DC's increased dismissal threat. Many teachers who had their probationary period extended left their schools.[15]

Did the new tenure consequences affect teachers' job performance? In a new quasi-experimental analysis, Dinerstein and Opper (2018) show that the new tenure-evaluation process did increase the measured effectiveness of non-tenured teachers in their pre-tenure years. Effectiveness is measured by teachers' contributions to student test scores. So, in other words, the new tenure rules led to higher test scores for students in non-tenure teachers' classrooms; however, these test score gains faded out more quickly than expected. Together these results suggest, as the authors note, that non-tenured teachers may have found ways to improve their evaluation scores which do not reflect meaningful improvements in teaching effectiveness.

One final note, pay for performance or the threat of dismissal are commonly-discussed consequences for the design of teacher evaluation systems. But these consequences are not required for evaluation to improve teachers' effectiveness in their work. Neither the Cincinnati nor Chicago cases, as examples, involved bonuses or dismissal, yet those programs produced improvements. There are many other consequences or incentives that teachers may feel, even if they are not explicitly stated by the evaluation program. One category is sometimes called "career concerns": the incentives to perform well in evaluations to improve one's chances of promotions or future job offers from other employers. These or other incentives may have been on teachers' minds in Cincinnati, Chicago, and other examples discussed in this paper.

Opportunity Costs of Evaluation Programs

Teacher evaluation programs can be costly. The most notable example is modern multiple-observation rubric-based classroom evaluations. The observing, scoring, devising suggestions for improvement, pre- and post-meetings, etc. together require considerable time and effort. If the evaluator is the school principal the budgetary costs may be hidden (responsibilities can change without salary changes). If the evaluator is a distinct job, like a peer evaluator, the budgetary costs are easier to see. But budgetary costs are not the only costs, and likely not the most important costs.

Modern observation-based evaluations, especially those which include formal or informal coaching, carry important opportunity costs. We should ask the question: What would the principal (peer evaluator) have been devoting her time and effort to if she had not been asked to carry out the evaluation? A principal may neglect or delegate other responsibilities. Many peer evaluators would have remained in their own classrooms teaching their own students. These opportunity costs can be sizable, but that does not necessarily mean investing time and effort in evaluation is the wrong choice.

[15] Recent research from Louisiana emphasizes the importance of receiving tenure or loss of tenure as consequences of evaluation. In 2012 Louisiana effectively eliminated tenure, but for two subsequent years did not yet begin evaluating teachers. Strunk, Barrett, and Lincove (2017) show that simply the threat of losing tenure, before evaluation began, increased teacher turnover by 20 percent.

San Juan Unified School District
System of Professional Growth

Key components. Based on the California Standards for the Teaching Profession, especially nine "essential elements." Initial self-evaluation and meeting. Multiple, rubric-based classroom observations. Pre- and post-observation conferences. Two reflective conversations to capture practices difficult to observe, such as lesson planning and data analysis.

Frequency. Non-tenured teachers are evaluated once a year. Tenured teachers are evaluated every other year. Tenured teachers with 10+ years of experience evaluated every three years.

Classroom observations. Observations are scored using a rubric that consists of nine elements from the CSTP (elements are the level below the six CSTP standards). Notably, the San Juan rubric includes many concrete strategies for teachers listed for each element. Each element is scored from 1-3, corresponding to not meeting standards, developing, or meeting standards. Observations for non-tenured teachers are conducted by administrators. Observations for tenured teachers are conducted by peer teachers when possible. All observations are announced in advance and last at least 40 minutes. Observers also conduct drop-in visits to collect additional evidence.

During observations. The district's objective during observations is not simply to score the teacher's practices, but rather to develop (strengthen) the teacher's ability to self-evaluate and develop her own plans for improvement; observers focus on skillfully asking questions.

Before and after observations. During a pre-observation conference, the teacher and evaluator discuss goals and the focus of the observation. During a post-observation conference, teachers select, share, and reflect on evidence of student learning.

Final evaluation ratings. Each teacher receives one of three overall ratings: not meeting standards, approaching standards, or meeting standards.

Consequences. To be considered in good standing, non-tenured teachers must score approaching standards, and tenured teachers mush score meeting standards. Teachers who do not (are not on track to) receive a rating of meeting standards are placed in Advisory. In Advisory, teachers work on an improvement plan receive support from a peer advisor for at least two hours per week. Teachers who receive a rating below meets standards after Advisory are referred to Peer Assistance and Review (PAR). After one full year of PAR, a teacher may be dismissed for an unsatisfactory evaluation.

Additional details. Teachers with more than five years of experience may apply to serve a four-year term as a peer facilitator. Peer facilitators are released full-time from teaching and receive continuing education credits.

Source: San Juan Unified School District and San Juan Teachers Association (2017), San Juan Unified School District personal communication (March 9, 2018)

A simple example demonstrates how the benefits of evaluation programs can quickly grow larger than the costs. First, let's be concrete about the opportunity costs of a peer evaluator. Imagine the peer evaluator is hired from among the district's top-quartile math teachers (the top 25 percent most effective teachers), and she is then replaced in her classroom with a teacher from the bottom quartile.[16] The students taught by the replacement teacher would score, on average, about 0.20 standard deviations lower in math than they would have if they had instead been taught by the teacher newly promoted to peer evaluator. The total loss is 0.20 times the number of students: as few as 20 for an elementary teacher or as many as 150 for middle and high school teachers.[17]

This substantial loss of math achievement can nevertheless be small relative to potential math achievement gains in the classrooms of teachers evaluated and coached by the new peer evaluator. Imagine that participating in a multiple-observation rubric-based peer evaluation does increase teaching effectiveness: specifically that students score 0.05 standard deviations higher in math, on average, than they would have if their teacher had not participated in the evaluation program. The opportunity cost of the peer evaluator would be "paid off" if she worked with just four evaluatees (0.05 * 4 = 0.20). In many such programs peer evaluators work with at least several teachers and often two dozen or more. What's more, if the improvements in evaluatee effectiveness continue into future years, as they did in Cincinnati, the return on investment grows large quickly as the now-more-effective teacher teaches many classes over the years of her subsequent career.[18]

A different potential opportunity cost of evaluation programs is that they might shrink or degrade the pool of applicant teachers. They might also expand or improve the pool. The GDTFII teacher survey asked what influenced their job application decisions; one of every seven teachers chose "how my work would be evaluated" among the top three influences.[19] For example, if a

---

[16] The example in this paragraph is taken from Taylor and Tyler (2012). In practice the school may be able to replace the top-quartile new peer evaluator it lost with an equally or only somewhat less effective teacher who was moved from a different classroom or hired away from another school. In the end, however, the district will need to hire a new teacher to replace the peer evaluator. Making the replacement more effective only *reduces* the opportunity costs.

[17] Some readers may want to go a step further and convert this math achievement loss into lost future student earnings as adults, or other measures of future student success, as in Chetty, Friedman, and Rockoff (2014). Test-score units are sufficient for our comparison of costs and benefits. Converting to future earnings would not change the example materially.

[18] This example ignores the question of whether repeated evaluation year after year would produce additive performance improvements year after year. Evaluation program designers should not expect large additive gains year after year for a given teacher.

An improvement of 0.05 does seem plausible the first time a teacher is evaluated; 0.05 is much smaller than the improvement estimated in the (quasi-)experimental studies in Cincinnati, Chicago, Tennessee, and with MyTeachingPartner. However it also seems unlikely evaluation would add an additional 0.05 year after year; the marginal returns must be diminishing under the hypothesized mechanisms, and the standard deviation in teacher effects is only 0.15-0.20. Still, even much smaller gains from evaluation are quickly multiplied by the number of evaluatees assigned to an evaluator, and the number of classes the evaluatee will teach in subsequent years. Finally, evaluation designers could only invest in intensive evaluation periodically, say every five years for a given teacher.

[19] One out of seven is meaningful, but "evaluation" was far from the most cited influences. 91 percent of teacher cited location, 47 percent cited the kinds of students they would be teaching, and 40 percent cited salary and benefits.

district reduces the chances of tenure, novice teachers looking for work may choose not to apply to that district, especially if there are other similar districts nearby with high tenure rates. That choice may be less likely, however, for an applicant who is certain her performance will result in tenure even if the average tenure rate is low. As with turnover, this potential effect of evaluation does not improve or degrade individual teachers' effectiveness, but can improve or degrade the average quality of teaching in the district by changing the composition of the applicant pool. Other features of evaluation programs may similarly affect the teacher applicant pool. We do not have space in this paper to go further into these considerations, but for interested readers we suggest Rothstein (2015).

Conclusion

Many California school districts, along with states and districts around the country, are making new investments in their teacher evaluation programs. A widely held goal is to create an evaluation program which helps individual teachers become more effective in the work of teaching. This paper has highlighted four common features of teacher evaluation programs, and summarized available research evidence on whether those features promote or hinder improvements in teaching effectiveness.

In general, relevant research evidence is scarce, but in some cases promising. There are, for example, promising examples of the benefits of using multiple, rubric-based observation programs, or the benefits of connecting evaluation results to specific plans and resources for improvement. The careful (quasi-)experimental evaluations of these examples make the ideas more promising than ideas without any evidence. Designs using the evidence discussed above are more likely to result in benefits to teaching effectiveness, but the limits of the evidence should remind us to not be surprised if the results do not completely carry over to other jurisdictions.

California school districts have an opportunity to make meaningful progress on the design of teacher evaluation programs. Teachers see room for improvement. Half of California teachers feel the evaluation program they are subject to is primarily, mostly, or entirely about grading teachers for accountability. They feel the goal of helping teachers improve is at best a secondary purpose or, for some, not a part of evaluation in their school. The other side of the coin, however, is that half of teachers do feel evaluation is primarily, mostly, or entirely about helping teachers improve their teaching. That suggests many examples of success throughout the state, examples from which other California schools and districts can borrow and learn. Moreover, unlike many other states, in California teacher evaluation is the responsibility of each district, leaving substantial latitude to innovate like many of the districts we highlighted in this paper.

References

Adnot, M. (2016). Effects of incentives and feedback on instructional practice: Evidence from the District of Columbia Public Schools' IMPACT teacher evaluation system. Working paper, available from author.

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis, 39*(1), 54-76.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333*(6045), 1034-1037.

Bacher-Hicks. A., Chin, M., Kane, T., & Staiger, D. (2017). An evaluation of bias in three measures of teacher quality: value-Added, classroom observations, and student surveys. NBER Working Paper 23478.

Brehm, M., Imberman, S. A., & Lovenheim, M. F. (2017). Achievement effects of individual performance incentives in a teacher merit pay tournament. *Labour Economics, 44*, 133-150.

Chetty, R., Friedman, J., & Rockoff, J. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review, 104*(9), 2633-2679.

Chiang, H., Speroni, C., Herrmann, M., Hallgren, K., Burkander, P., & Wellington, A. (2017). *Evaluation of the teacher incentive fund: Final report on implementation and impacts of pay-for-performance across four years*. (NCEE 2018-4005). Washington, DC: U.S. Department of Education.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher, 45*(6), 378-387.

Cullen, J. B., Koedel, C., & Parsons, E. (2016). The compositional effect of rigorous teacher evaluation on workforce quality. NBER Working Paper 22805.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching.* Alexandria, VA: ASCD.

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management, 34*(2), 267-297.

Dinerstein, M., & Opper, I. (2018). Does incentivizing value added make it more or less meaningful? Working paper, available from authors.

Doe v. Antioch, MSN15-1127 (Superior Court of the State of California, Contra Costa County 2016).

Gill, B., Shoji, M., Coen, T., and Place, K. (2016). *The content, predictive power, and potential bias in five widely used teacher observation instruments.* Washington, DC: U.S. Department of Education.

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National Board Certification as a signal of effective teaching. *Review of Economics and Statistics, 89*(1), 134-150.

Goodman, S., & Turner, L. (2013). The design of teacher incentive pay and educational outcomes: Evidence from the New York City bonus program. *Journal of Labor Economics, 31*(2), 409-420.

Grissom, J. A., & Loeb, S. (2017). Assessing principals' assessments: Subjective evaluations of teacher effectiveness in low-and high-stakes environments. *Education Finance and Policy, 12*(3), 369-395.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56-64.

Ho, A., & Kane, T. (2013). *The reliability of classroom observations by school personnel.* Seattle: Bill & Melinda Gates Foundation.

Holtzapple, E. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education, 17*(3), 207–219.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). *Identifying effective classroom practices using student achievement data. Journal of Human Resources, 46*(3), 587-613.

Kane, T. J. & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill & Melinda Gates Foundation.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*.

Kraft, M. A., & Gilmour, A. (2016). Can principals promote teacher development as evaluators? A case study of principals' views and experiences. *Educational Administration Quarterly, 52*(5), 711-753.

Loeb, S., Miller, L., & Wyckoff, J. (2015). Performance screens for school improvement: The case of teacher tenure reform in New York City. *Educational Researcher, 44*(4), 199-212.

Long Beach Unified School District. (2014). *Certified personnel evaluation.* Long Beach, CA: Author. Retrieved from https://www.documentcloud.org/documents/2068661-teacher-eval-longbeachform2015.html

Long Beach Unified School District & Teachers Association of Long Beach. (n.d.). *K-12 teachers contract: effective through June 30, 2018.* Long Beach, CA: Authors. Retrieved from http://talb.org/wp-content/uploads/2017/08/TALBK-12Contract-EffectiveThrough6-30-2018-2016-17reopeners.pdf

Los Angeles Unified School District. (2016). *LAUSD Teaching and Learning Framework.* Los Angeles, CA: Author. Retrieved from

https://achieve.lausd.net/cms/lib/CA01000043/Centricity/Domain/433/2016%202017%20TLF%20Booklet.pdf

Los Angeles Unified School District. (2017). *2016-2017 EDST final evaluation report: Administrator handbook.* Los Angeles, CA: Author. Retrieved from https://achieve.lausd.net/cms/lib08/CA01000043/Centricity/Domain/433/2016%202017%20EDST%20Final%20Eval%20Report%20Handbook.pdf

Los Angeles Times. (May 20, 2013). California won't get relief from No Child Left Behind law. Byline: Howard Blume.

Los Angeles Times. (September 26, 2016). Court refuses to mandate test scores in teacher evaluations. Byline: Howard Blume.

Marsh, J. A., Springer, M. G., McCaffrey, D. F., Yuan, K., Epstein, S., Koppich, J., Kalra, N., DiMartino, C., & Peng, A. (2011). *A big apple for educators: New York City's experiment with schoolwide performance bonuses.* Santa Monica, CA: RAND Corporation.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33-53.

Milanowski, A. T., & Heneman, H. G. (2001). Assessment of teacher reactions to a standards-based teacher evaluation system: A pilot study. *Journal of Personnel Evaluation in Education, 15*(3), 193-212.

Milanowski, A. T., Kimball, S. M., & White, B. (2004). The relationship between standards-based teacher evaluation scores and student achievement: Replication and extensions at three sites. Consortium for Policy Research in Education, University of Wisconsin, Working Paper TC-04-01.

Neal, D. (2011). The design of performance pay in education. In *Handbook of the Economics of Education Volume 4*, Hanushek, E. A., Machin, S., & Woessmann, L. (eds), 495–550. Amsterdam: North-Holland, Elsevier.

Papay, J., Taylor, E., Tyler, J., & Laski, M. (2017). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. Working paper, available from authors.

Papay, J., Goldring, E., Grissom, J., Laski, M., Patrick, S., Taylor, E., & Tyler, J. (2017). Encouraging compliance without mandates: The challenge of take-up in a voluntary state-sponsored professional learning initiative. Working paper, available from authors.

Poway Unified School District & Poway Federation of Teachers. (n.d.). *Agreement between Poway Unified School District and Poway Federation of Teachers: July 1, 2012 – June 30, 2015.* Poway, CA: Authors. Retrieved from https://www.powayusd.com/PUSD/media/PSS/Employment/Certificated/pss_PFT_Contract-2012-2015.pdf

Poway Unified School District & Poway Federation of Teachers. (2013). *A comprehensive peer support and peer review program.* Poway, CA: Authors. Retrieved from https://www.powayusd.com/PUSD/media/PSS/PPAP/PPAPHandbook.pdf

Poway Unified School District & Poway Federation of Teachers. (2016). *Draft Continuum of Teaching Standards.* Poway, CA: Authors. Retrieved from http://www.powayusd.com/PUSD/media/PSS/PPAP/PUSD-Continuum-of-Teaching-Standards.pdf

Rockoff, J., Staiger, D., Kane, T., & Taylor, E. (2012). Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review, 102*(7), 3184-3213.

Rothstein, J. (2015). Teacher quality policy when supply matters. *American Economic Review, 105*(1), 100-130.

San Jose Unified School District. (2015). *Teacher evaluation system handbook.* San Jose, CA: Authors. Retrieved from http://documentcenter.meetingcaddie.mobi/Events/1b542d45-a9e1-43ac-920b-bf3f492a00ef/Documents/33542afc-d52c-4d5a-b565-84d9ff005f7c.pdf

San Jose Unified School District & San Jose Teachers Association. (2016). *Agreement between San Jose Unified School District and San Jose Teachers Association 2016-2019.* San Jose, CA: Author. Retrieved from http://sanjoseteachersassociation.org/wp-content/uploads/2014/08/2016-2019-SJTA-CBA-v1608.pdf

San Juan Unified School District & San Juan Teachers Association. (2017). *Collective bargaining agreement between San Juan Unified School District and San Juan Teachers Association.* Carmichael, CA: Authors. Retrieved from http://www.sjta.org/docs/SJTA-COLLECTIVE-BARGAINING-CONTRACT-1618-Revised-July-1-2017.pdf

Sartain, L., & Steinberg, M. P. (2016). Teachers' labor market responses to performance evaluation reform: Experimental evidence from Chicago Public Schools. *Journal of Human Resources, 51*(3), 615-655.

Springer, M.G., Ballou, D., Hamilton, L., Le, V., Lockwood, J.R., McCafrey, D., Pepper, M., & Stecher, B. (2010). *Teacher pay for performance: Experimental evidence from the Project on Incentives in Teaching.* Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: understanding the landscape of teacher evaluation in the post-NCLB ear. *Education Finance and Policy, 11*(3), 340-359.

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis, 38*(2), 293-317.

Steinberg, M., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy, 10*(4), 535-572.

Strunk, K. O., Barrett, N., & Lincove, J.A. (2017). *When tenure ends: The short-run effects of the elimination of Louisiana's teacher employment protections on teacher exit and retirement.* New Orleans, LA: Education Research Alliance for New Orleans.

Strunk, K. O., Cowen, J., & Goldhaber, D., Marianno, B., Kilbride, T. & Theobald, R. (2018). "It is in the contract: How the policies set in teachers' unions' collective bargaining agreements vary across states and districts." *Educational Policy, 32*(3), 280-310.

Strunk, K. O. & Reardon, S. (2010). "Measuring union strength: A partial independence item response approach to measuring the restrictiveness of teachers' union contracts." *Journal of Educational and Behavioral Statistics, 35*(6), 629-670.

Taylor, E. & Tyler, J. (2012a). The effect of evaluation on teacher performance. *American Economic Review, 102*(7), 3628-3651.

Taylor, E. & Tyler, J. (2012b). Can teacher evaluation improve teaching? Evidence of systematic growth in the effectiveness of midcareer teachers. *Education Next, 12*(4), 78-84.