Teacher evaluation policies seek to improve student outcomes by improving teachers’ skills and increasing their effort. Most present-day evaluation policies in the United States aim to impose accountability pressures and provide supports that generate professional growth. Proper policy design has been understood as successfully weighting the accountability and growth dimensions of teacher evaluation. I detail the conditions that determine whether joint-aim teacher evaluation policies will improve student outcomes. I then assess the extent to which these conditions are likely to be met in light of the causal evidence base from the education, economics, social psychology and management literatures. Informed by this empirical review, I conclude with recommendations to more clearly delineate the accountability and growth aims of teacher evaluation.

Keywords: educational policy, teacher evaluation, labor contracts, personnel management
Workplaces tend to design human resource management systems to encourage employee effort, to permit recognition and sanction for different performance levels, and to improve skills through feedback generated as part of the evaluation process. The purported mechanisms through which present-day teacher evaluation design in the United States might improve the average quality of instruction are through (a) incentives to motivate teachers, (b) tools for retaining high-performing teachers and deselecting low-performing ones, (c) shifts to the labor market pool of prospective teachers as a result of the perceived attractiveness of the profession, and (d) feedback mechanisms to improve the skills of current teachers. As such, most teacher evaluation policies attempt to achieve a blend of accountability- and growth-oriented goals. While the two goals are firmly part of the design of present-day teacher evaluation policies, researchers and policy makers have infrequently reflected on the interactions between these aims.

In this empirical review, I argue that a clear explication of the tenets underlying teacher evaluation policies is critical to estimating their likelihood of success. I identify six conditions that determine whether teacher evaluation policies will produce improvements in average teacher effectiveness, which I define broadly as teachers’ contributions to better student outcomes. I assess the extent to which these conditions are satisfied in light of the education, microeconomics, management and social psychology literatures. I emphasize research that supports generalizable and credibly causal claims regarding the effects of teacher evaluation. My review highlights that the theory and evidence motivating incentive- and sanction-based evaluation policies are in potential conflict with theory and evidence justifying human resource strategies to develop employees’ skills. I conclude, as a result, that teacher evaluation policies that more clearly delineate the application of accountability pressures and growth supports across different groups of teachers may improve student outcomes more than current joint-aim policies.
This review proceeds in four parts. First, I describe historical trends and the current state of U.S. teacher evaluation policy. Second, I detail six conditions that determine the success of teacher evaluation policies. Next, I review the evidence on the extent to which these conditions are satisfied in practice. Finally, I conclude with a discussion about the relevance of this review to future teacher evaluation policy development.

Teacher Evaluation Policy in the United States

Historical Trends in Teacher Evaluation Policy, Practice and Research

Reformers have sought to implement rigorous controls on teachers’ practice for over a century. Proponents of scientific management and Taylorism attempted to improve the efficiency of teachers’ practices as early as the turn-of-the-century Progressive Era (Callahan, 1964; Mehta, 2013). In addition to classroom visit checklists, expert “school surveys,” and the development of standardized tests, Callahan documents the surprising finding that almost half of U.S. school districts sampled in 1918 used some form of a merit pay compensation system.

The traditional rationale for merit pay, emerging from the microeconomics field of contracts, is that workers’ preferences are not perfectly aligned with their employers’, and that monitoring worker output and actions is difficult and costly (Murnane & Cohen, 1986). Therefore, employers enter into a contract with workers in which employees receive additional pay for performance based on either the completion of a particular output or the subjective assessment of a supervisor. Murnane and Cohen present a stylized situation in which the justification of merit

---

1 I focus this review on teacher evaluation in the United States and draw empirical evidence from studies in this context. Some lessons in this review are broadly applicable. Others have limited relevance to national contexts where teacher accountability is entirely incentive- and sanction-based (e.g., Muralidharan & Sundararaman, 2011), where evaluation is conducted primarily to fulfill administrative requirements (e.g., OECD, 2019), or to teacher coaching schemes with no connection to the appraisal system (e.g., Murphy, Weinhardt, & Wyness, 2020)
pay is relatively straightforward: two workers unload boxes from a truck; one worker is paid more because she moves more boxes than the other worker.

Several theoretical objections exist to the application of contracts theory to public sector employees generally and teachers, particularly. A first order concern is that accurately measuring all desired student outcomes is difficult; it is even more difficult to know which teacher behaviors will lead to these outcomes; and most difficult of all is to observe the totality of teacher behaviors (Dixit, 2002; Murnane & Cohen, 1986). An additional challenge is that public institutions have multiple goals for which outcomes are unverifiable, and actions are only minimally observable (Dixit, 2002; Holmstrom & Milgrom, 1991). As a result, the agents of teacher evaluation policies (teachers) also must respond to multiple principals (parents, children, administrators, policy makers, etc.) which complicates teachers’ response to the action of any one principal. Further, the nature of teaching lends itself poorly to these sorts of incentives because it is particularly susceptible to shifting resources to measured outcomes at the expense of other desirable ends (Murnane & Cohen, 1986). Finally, Dixit (2002) notes that teachers are generally “motivated agents,” such that what primarily impedes their success is skill rather than will. Scholars in the field of education have further highlighted the difficulty in disentangling the skill of teachers from the situational contexts that appear to strongly influence their development (e.g., Kennedy, 2010).

As a result of these challenges, combined with the legal protections of tenure and the institutional cultures of schools, for much of the second half of the twentieth century, school practice and policy favored the use of teacher evaluation as a mechanism to improve the quality of instruction. A brief survey of the most influential practitioner-oriented school leadership guides reveals the prominent role of evaluation as a tool to improve teacher practice through observation, review of student work products, and targeted feedback (Bambrick-Santoyo & Peiser, 2012;
Marshall, 2009; Platt, Tripp, Ogden, & Fraser, 2000). In fact, the historically low rates of the assignment of teachers into evaluation categories indicating unsatisfactory performance (Weisberg, Sexton, Mulher, & Keeling, 2009) suggest that administrators rarely saw potential for the use of dismissal (or its threat) as a mechanism to improve average instructional performance.

The mid-twentieth-century enacted understanding of the role of teacher evaluation as an opportunity for professional growth is best understood through a separate literature on human resource development and management. Armstrong (2000) argues that the developmental aspects of appraisal are key to improving employee performance. In this conceptualization, borrowing Murnane and Cohen’s (1986) heuristics, the supervisor’s goal is to ensure that all truck loaders can carry more boxes at higher speeds. Appraisal creates a formalized opportunity for the supervisor to provide insights, coaching and opportunity for the worker to improve her box-carrying skills.

A confluence of forces over the last two decades of the past century, however, led to a renewed interest in the policy and academic communities for accountability-driven teacher evaluation policies. Policy and political shifts in the aftermath of the publication of *A Nation At Risk* (Berliner & Biddle, 1995; Mehta, 2013) coincided with new theories from the academy (see Tolofari [2005] in the context of New Public Management) to prompt a return to an emphasis on efficiency improvements in the public sector through greater use of outcome-based personnel management. Still, it was not until technologic and methodologic advances re-opened the policy window at the beginning of the 21st century that these efforts came to fruition in the United States.

The advent of large-scale educational administrative datasets allowed researchers to document differences in the “productive” capacities of teachers with respect to their students’ improved learning outcomes. In the aftermath of pioneering work by Bill Sanders (1996; 1997) in
Tennessee, a large body of empirical work highlighted not only the large contributions that teachers make to student learning trajectories (e.g., Gordon, Kane, & Staiger, 2006; Hanushek & Rivkin, 2010; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004), but also the long-term causal labor market benefits to the students of highly effective teachers (e.g., Chetty et al., 2011). A separate set of high-visibility policy reports emerged simultaneous to these research findings documenting the near-universal positive appraisals assigned to teachers (Weisberg et al., 2009) along with recommendations to assign multiple ratings to teachers’ performance with significant consequences on employment decisions (TNTP, 2010).

These joint trends prompted researchers and policy makers to assess whether education systems might better serve students by improving the quality of the overall teaching pool through human resource management policies. These include rigorous evaluation systems for teachers leading to stronger incentives for teachers to perform and the deselection of ineffective teachers (e.g., Gordon et al., 2006; Hanushek, 2009; Jackson, Rockoff, & Staiger, 2014; Rockoff & Speroni, 2011; Rockoff, Staiger, Kane, & Taylor, 2012; Staiger & Rockoff, 2010).

**Present-Day Teacher Evaluation Policies**

Murnane and Cohen (1986) presciently categorize “new” and “old” style merit pay systems as those which provide additional compensation for either improved student outcomes (“new” piece-rate compensation) or those which rely on administrator observations (“old” subjective supervisor judgment). In fact, the vast majority of present-day teacher evaluation systems employ a mix of both “new” and “old” style appraisal of teacher effectiveness.

In response to incentives from the Obama administration’s Race to the Top (RTT) program in 2009, state legislatures across the country enacted laws aimed at increasing accountability standards for public school teachers through the implementation of high-stakes teacher evaluation...
systems. Between 2009 and 2016, 44 states and the District of Columbia implemented major reforms to their teacher evaluation systems (NCTQ, 2016; Steinberg & Donaldson, 2016). In almost all cases these reforms entailed adopting a common rubric for evaluating teachers’ performance with multiple rating categories, representing a shift away from the traditional Satisfactory/Unsatisfactory distinctions. All state reforms to teacher evaluation required that classroom observation of teaching practice be a part of a teacher’s final rating, and in most cases these reforms established a minimum frequency of classroom observations. In addition, many states required some or all teachers to be evaluated based on student-learning gains (either through formal measurements of students learning, through teachers’ contributions to students’ progress towards locally determined learning objectives, or both). Some states additionally included measures of whole-school performance or parent-, student-, and peer-surveys of teacher competency (Steinberg & Donaldson, 2016; Winters & Cowen, 2013).

Many states set a high bar for teachers’ instructional proficiency. States typically either adopted previously validated measures of instructional practice, e.g., the Danielson Framework for Teaching (1996), the CLASS rubric (2008), the Skillful Teacher (2008), or synthesized these frameworks into an original one for their evaluation systems. Figure 1 articulates the demonstrated skills required across rubric ratings for one of 30 elements in the Massachusetts Model System for Educator Evaluation Classroom Teacher Rubric: Well-Structured Units and Lessons. At the Proficient and above categories, the evaluation rubrics tend to articulate aspirational goals. It is the rare lesson that “implements a standards-based unit (…) with challenging tasks and measurable outcomes, appropriate student engagement strategies, pacing, sequence, resources, and grouping; purposeful questioning, and strategic use of technology and digital media, such that students are able to learn the knowledge and skills defined in state standards/local curricula” (DESE, 2018, p.
2). If classroom observations and artifact reviews reveal that only a few lessons, and by extension teachers, satisfy these rigorous criteria for proficiency, the implication is that most teachers should be rated at the Needs Improvement level. As I detail below, teachers who fail to earn an evaluation rating at the Proficient or higher threshold are subject to dismissal. Though Massachusetts is an outlier with respect to several dimensions of its educational system, its evaluation guidelines are typical of many states and districts and provides a helpful policy example to which I return throughout the paper.

Commentators have debated the extent to which modern teacher evaluation policies, either as written, as implemented, or as altered post-implementation, truly impose a higher-degree of external accountability. For example, while 43 states initially required teachers be rated on objective measures of student growth, nine states have since rescinded this requirement (Ross & Walsh, 2019). As I discuss in more detail below, most educators continued to receive positive appraisal ratings (Kraft & Gilmour, 2017). However, it is important to note that poor appraisals risk significant consequences; in over three-fifths of states, teachers who are not rated Proficient should, by policy, be dismissed and almost half of states use evaluation results to grant or revoke tenure (Steinberg & Donaldson, 2016). The Institute of Education Sciences (2014) judged the recent round of teacher evaluation reforms to have made strong shifts towards increased accountability.

To summarize the salient features of state and local teacher evaluation policies in the United States at the start of the 2020s: (a) on paper they set high standards for quality teaching; (b) their stated aims are for both teacher growth and accountability; (c) the failure to receive a designation meeting high bars of proficiency should, in principle, lead to dismissal; and (d) despite the new

---

2 38 states implemented teacher evaluation ratings with at least four different categories (Walsh et al., 2017). Despite differences in nomenclature, most imply ratings of Ineffective, Needs Improvement, Proficient and Exemplary.
policies, nearly all teachers continue to receive ratings above the Unsatisfactory or equivalent level.

**What Must Teacher Evaluation Accomplish to Improve Student Outcomes?**

The promise of mechanisms to evaluate teachers based on production led to widespread policy consensus on the potential value of higher-stakes teacher evaluation systems; however, the assumptions underlying the mechanisms by which these accountability and incentive systems work are poorly understood.

Over at least the past 35 years, commentators have discussed the strengths and weaknesses of teacher evaluation for accountability and evaluation for growth (e.g., Darling-Hammond, Wise, & Pease, 1983; Donaldson & Papay, 2015; Popham, 1988). They generally conceive of particular teacher evaluation policies as points along a continuum. A particular policy might apply more or less accountability pressures or growth supports. The sum total of the policy is understood as essentially the linear combination of the aspects of the policy that focus on accountability and rewards and those aspects of the policy that focus on professional growth. The goal of the policy maker is to appropriately balance the weights of these two dimensions of evaluation. While many have considered the possibility that the accountability and growth goals of teacher evaluation policy are either mutually reinforcing or are counterproductive to the other aim (e.g., Firestone, 2014; Hallinger, Heck, & Murphy, 2014; Lavigne, 2014; Moore Johnson, 2015; Papay, 2012), few have considered the possibility that evaluation policies might tailor their accountability or skill development emphasis to teachers based on their individual professional profiles. In this paper, I extend prior work through an explicit analysis of the conditions under which evaluation systems produce improved student outcomes, and I highlight the promise of evaluation policies that permit differentiated approaches dependent on teacher needs.
Teacher evaluation systems with a joint focus on both accountability/incentives and the skill development of teachers employ the following logic: evaluations increase effort and skill of current teachers and evaluations change the composition of the teaching force. Average teacher effectiveness improves both through improvements in incumbent teachers’ skills, deselection and selective attrition, and improvements in the skills of entrants to the profession due to perceptions of increased professional prestige. For such a theory of action to hold, a combination of the following six conditions should be true:

1. **Reliability and validity.** Evaluation ratings must reliably distinguish between teachers who facilitate higher and lower levels of learning for students on short-term outcomes, which are in turn predictive of meaningful longer-term outcomes;

2. **Accountability/incentive effects.** Classification of teachers into different rating categories produces consequential outcomes that (a) promote improvement in teaching practices through increased effort and skill acquisition and (b) inform employment decisions;

3. **Skill development effects.** Evaluations generate meaningful developmental supports that produce improvements in teachers’ skills;

4. **Overall labor demand and supply effects.** The supply of teachers seeking employment is minimally equal to the demand for teachers generated either by attrition or dismissal; prospective and current high-capacity teachers are able to accurately forecast their positive evaluation results and do not withhold their supply of labor to the teaching pool as a result of their risk aversion to potential future dismissal;

5. **Compositional effects on labor market.** Prospective teachers possess (or can quickly acquire) teaching skills equal or superior to departing teachers; candidate teachers with skills higher than those currently employed are able to gain entry to the profession; and
6. **Interactions between growth and accountability.** The structure by which evaluation policy promotes both accountability and developmental goals does not cause one mechanism to inhibit the effectiveness of the other and may generate complementarities.

One of (#2), (#3) or (#5) is a necessary, though not sufficient, condition for any evaluation policy to improve average teacher effectiveness. Said differently, evaluation must increase skills in incumbent teachers through either accountability (#2) or professional growth (#3), or evaluation must change the skill composition of the teacher workforce through differential attrition, deselection and changes to the skills of incoming cohorts of teachers (#5). As long as one of the three preceding conditions is met, sufficiency is an empirical concern that requires a comparison of the reinforcing or countervailing effects of the other conditions. For instance, if accountability pressures improve teacher practice (#2), unreliable evaluation ratings (#1) might still capture enough signal about teaching quality that their use would improve the overall distribution of teacher effectiveness. In this scenario, accountability pressures could improve average teacher effectiveness, even in the presence of teacher shortages (#4) or even if it constrained the skill development dimensions of evaluation (#6), as long as these secondary effects were smaller in magnitude than the main effects of individual teacher improvements from accountability.

An assessment of present-day, joint-aim teacher evaluation policies requires not only evaluating whether they satisfy sufficiency conditions for improving average teacher effectiveness, but also how they compare to policy counterfactuals. If alternative policy designs more readily satisfy some of the conditions or do not require them, this would be valuable for policy makers to know as well. The next section synthesizes the evidence on the extent to which these conditions hold in practice.
What is the Evidence on the Extent to which the Conditions for Optimal Joint-Aim Teacher Evaluation Policies Are Met?

In the following sections, I present evidence on the above conditions from an empirical review of research studies that support causal inferences. I define such studies as those which identify the effects of teacher evaluation through a source of exogenous variation in teachers’ exposure to different forms of evaluation. In practice, I restrict the definition of exogenous variation to studies employing randomized control trials, regression discontinuity, difference-in-differences or instrumental variable estimates. However, I am also interested in understanding measurement concerns (Condition 1), descriptive evidence on policy implementation, and causal evidence on professional accountability outside the field of education. As such, and given the interdisciplinary goals of my review, I also rely on snowball searches to supplement the systematic portion of my analysis.3

Condition 1. Reliability and Validity of Evaluations

The measurement and economics of education research literature have raised various concerns about the reliability, internal validity and predictive validity of teacher evaluation ratings. In this context, reliability (also described as “stability” in some contexts) implies that different measurements of teacher effectiveness should rate the same level of teaching performance consistently across outcome measures and over time. Internal validity means that teacher effectiveness ratings should be a function of teachers’ skill, rather than, for example, differences

3 In particular, I search ERIC, EconLit, and Google Scholar for the terms “teacher evaluation” AND “United States” AND “causal OR quasi-experimental” for items published since 2000. Of the 2,399 returns across the three databases, only 10 include distinct analyses of the effects of K-12 teacher evaluation in the United States in a context in which exogenous variation in exposure to different forms of supervisory evaluation practices exist. I exclude from these counts articles that focus exclusively on methodological questions of reliability and validity, but I include many of these results in my discussion of Condition 1. I include all 10 of these studies in this section though the full empirical review incorporates more than 75 studies, highlighting the value of literature searches using techniques other than database-reliant ones.
in the starting ability levels of the students they teach. Finally, to be predictively valid, teachers’ evaluation ratings should relate to meaningful later-life outcomes for their students. I categorize these issues as they pertain to either the use of student-outcome measures (e.g., test scores, attendance rates, high-school graduation) or to other measures of performance (e.g., observation ratings) in teacher evaluation frameworks.

Test-score-based teacher effects measures are moderately reliable; thus, for a particular teacher in a given year, her rating may be idiosyncratic. Whether as a result of different methods for estimating teachers’ value-added contributions (Guarino, Reckase, & Wooldridge, 2015) or due to different tests within the same subject area (Papay, 2011), teachers’ effectiveness ratings can be sensitive to the method for calculating them. However, the more years a teacher is observed, the more stable the estimates become (Goldhaber & Hansen, 2010). Further, Staiger and Rockoff (2010) demonstrate that value-added-based dismissals can improve average teacher effectiveness at all but the lowest levels of reliability; though their results ignore the labor market and cross-productivity effects of evaluation.

The central threat to the internal validity of student-test-score-based measures is the problem of student sorting. The key concern with student sorting is that if students with greater potential for large learning gains systematically end up in some teachers’ classrooms rather than others, estimates of teachers’ effects will be biased (American Statistical Association, 2014; Ballou & Springer, 2015; Rothstein, 2010, 2017). Others argue that concerns about sorting and causal misattribution are either overblown or can be solved through novel modeling strategies (Chetty, Friedman, & Rockoff, 2014a, 2017, 2014c; Koedel & Betts, 2011; Koedel, Mihaly, & Rockoff, 2015). Kane, Staiger and co-authors (2013; 2008) estimate bounds of the potential bias from student sorting. They find that teachers’ value-added estimates from a year in which students are randomly
assigned to classrooms are highly predictive of their next-year effectiveness, suggesting that standard value-added modeling approaches account for most of the sorting concerns. Koedel, Mihaly and Rockoff (2015) artfully review the methodological issues surrounding value-added modeling. A full discussion of the technical debate regarding the internal validity of teacher effect estimates falls beyond the scope of this paper. However, to the extent that these concerns are reasonable, they threaten the ability of an evaluation system which relies on such measures to improve student outcomes because they increase the risk that strong teachers are misidentified as weak due simply to the students they teach and vice-versa.

While teachers’ effects on students’ test scores are predictive of some long-term outcomes, they align imperfectly with other important medium- and long-term outcomes. Teachers’ test-score value-added contributions are predictive of students’ longer-term college attainment and labor market outcomes (Chetty et al., 2011; Chetty, Friedman, & Rockoff, 2014b). As Chetty and co-authors note, their estimates emerge from a low-stakes context. As the stakes of student tests increase for teachers, incentives increase to narrow the curriculum or game the test (e.g., Ballou & Springer, 2015). Even discounting these Campbell’s Law (1979) effects, separate predictive validity concerns exist.

Recent evidence indicates that teachers’ impacts on test-score outcomes do not correlate well with teachers’ impact on other desirable outcomes such as student attendance (Gershenson, 2016) and non-cognitive skills such as resilience, growth mindset, self-efficacy and behavior in class (Blazar & Kraft, 2017; Kraft, 2019). In fact, Jackson (2018) finds that teachers’ contributions to student behavioral outcomes (measured as an index of suspensions, absences and GPA) are only weakly related to test-score outcomes ($r \approx 0.15$). Further, he finds that teachers’ behavioral value-added estimates are more predictive of high-school completion and long-run outcomes than their
value-add on test scores. This last point is critical in determining which measure of teacher performance is most valid to make an inference about teacher effectiveness. Evaluation schemes that incorporate near-term performance measures that most closely correlate with longer-term life outcomes are preferable both in prioritizing goals and in establishing face validity. Given these concerns about student-outcome-based measures of teacher effectiveness, present-day evaluation systems tend to rely on multiple measures of teacher effectiveness.

Multiple-measure evaluation systems are likely more reliable and valid instruments than value-added rank-order systems; though much depends on implementation. The empirical evidence consistently finds that observation-based evaluation and student-outcome-based appraisal align; however, they do so imperfectly (Harris, Ingle, & Rutledge, 2014; Harris & Sass, 2014; Jacob & Lefgren, 2008; T. J. Kane et al., 2013; Kraft, Papay, & Chi, 2020; Sartain, Stoelinga, & Brown, 2011). Cognizant of these imperfect correlations, policy makers have designed many evaluation systems to include multiple measures of teaching effectiveness that are bundled into a composite rating. While these composite measures return different rank orders of teachers depending on the weight assigned to each component of the system (Martínez, Schweig, & Goldschmidt, 2016; Steinberg & Kraft, 2017), they are generally internally consistent across years (Doan, Schweig, & Mihaly, 2019). However, if no variation in teacher observation ratings exists (see the following sub-section on Condition #2), multiple-measure evaluation systems will effectively revert to a rank-order test-score value-add based system because the lack of variation in non-test-based measures will lead to a performance distribution determined entirely by teachers’ test-score effectiveness.

In sum, while both “old”- and “new”-style evaluations may measure a similar underlying construct related to overall teacher effectiveness, substantial work remains to improve the validity
and reliability of this measurement. While no measurement will be perfectly valid, shortcomings in this dimension of teacher evaluation practices are of particular concern if punitive stakes remove a high-capacity or high-potential teacher from the profession.

**Condition 2. Accountability and Incentive Effects of Evaluation**

Some evidence highlights the potential of accountability pressures on teachers to improve teacher effectiveness, as measured by student test scores. Pope (2019) examines the effects of the public release of teachers’ value-added scores in Los Angeles. This event potentially tests the pure accountability effect of outcome-based evaluation (Condition 2(a)), as it was not coupled with any feedback to teachers. Pope finds that the release of these ratings caused an increase in the effectiveness of teachers in the bottom quintile of the performance distribution between 0.11 to 0.14 standard deviations. He finds no consistent (and some potential negative) effects throughout the rest of the distribution.

Phipps and Wiseman (2019) leverage random variation in District of Columbia teachers’ probability of being observed for accountability purposes and document small improvements in evaluators’ ratings of teachers’ practices as their probability of being observed increases. Phipps (2018) uses the same natural experiment to estimate the effects of evaluation probability on outcomes and finds students perform better when there is a threat of evaluation accountability (0.03-0.07 SDs) but that the increasing probability of evaluation does not affect test scores.4 Thus

---

4 Macartney, McMillan and Petro (2018) document increases in teacher value-added performance in North Carolina when larger proportions of their students are close to accountability-based proficiency thresholds on state exams. They argue that teachers respond with increased effort to more intensive accountability pressures. However, they explicitly discount the possibility for teacher learning-on-the-job. While not a direct test of teacher evaluation, these results suggest potential benefits to accountability pressures that are not tied to educator support. There is also a large adjacent body of literature on the effects of incentive pay on teacher performance (e.g., Goodman & Turner, 2013; Sojourner, Mykerezi, & West, 2014; Speroni et al., 2019). I do not fully explore the applicability of differential teacher compensation insights to broader evaluation strategies. However, the mixed nature of the evidence on merit pay for teachers is generally consistent with the results above.
evidence on 2(a) is mixed, but suggestive of the benefits of accountability pressures on the extensive, if not the intensive, margin particularly for the least effective teachers.

As I discuss above, modern evaluation policies generally satisfy the remainder of Condition 2 as designed. The majority of evaluation systems assign teachers into one of several rating categories (Walsh, Joseph, Lubell, & Lakis, 2017), and assignment into low-performing categories generates meaningful consequences, including dismissal. The extent to which the new evaluation systems introduced meaningful accountability as enacted, however, is less certain.

Even after the introduction of clear observational rubrics and rating categories, most teachers across the country continued to receive ratings above the standard of proficiency (Anderson, 2013; Kraft & Gilmour, 2017). In Figure 2, I present the evaluation ratings assigned to Massachusetts teachers from 2012/13 to 2016/17. Despite the high standards for Proficient teaching practice articulated in the standards above, fewer than 4 percent of teachers were evaluated below standard in the 2016/17 school year. The distribution of state-level evaluation scores in 2014/15 placed Massachusetts eighth of 24 states for the frequency with which teachers received ratings below Proficient (Kraft & Gilmour, 2017). Thus, Massachusetts is fairly typical in assigning nearly universal positive ratings. While there are many explanations for this phenomenon, ranging from school culture to lack of administrative capacity, one critical explanations is the design of the policy response to ratings below the Proficient level.

The Massachusetts state-wide evaluation model employs a typical accountability response to teachers deemed less-than-proficient through the evaluation process. Educators earning an Unsatisfactory rating are placed on an Improvement Plan. Improvement plans that last from 30 days to one school year in duration. Failure to make substantial progress towards Proficiency

---

5 New Mexico is the one exception to this pattern. However, as a result of massive political objections, the state began to dismantle its teacher evaluation policy in 2019.
should result, by the terms of the model contract language, in recommendation for dismissal to the superintendent. Most critically to understand the high-stakes nature of the evaluation process for teachers rated just shy of proficiency: if a teacher earns a Needs Improvement on either of the two instructionally focused Performance Standards in the teacher appraisal rubric, the model contract language places the teacher on a Directed Growth Plan of one-year in duration. If the educator does not earn a rating of Proficient at the end of the Directed Growth Plan’s duration (one year), the educator is placed on an Improvement Plan. Failure to make substantial progress towards proficiency should result as above in a recommendation for the teacher’s dismissal. Similarly, teachers at the end of their third year within a district must be rated Proficient on all four of the Performance Standards on the teacher rubric as well as Proficient overall to attain professional status and tenure and remain employed by the district (DESE, 2012).

As articulated, this contractual language should imply an increase in the proportion of teachers rated Unsatisfactory after the initial introduction of the evaluation policy. Teachers initially rated Needs Improvement who fail to improve on growth plans are, by policy, converted into the lowest category rating. However, as Figure 2 reveals, no such increase occurred. In fact, the proportion of teachers rated Unsatisfactory declined over this period.

I hypothesize that the policy structures may explain why some of the purported benefits of evaluation for accountability purposes did not materialize. Evaluators may have been deterred from assigning low ratings because the costs of pursuing teacher dismissal were too high in the face of uncertain prospects stemming from tenure law protections (recognizing that Condition 2(b) does not hold). Alternatively, evaluators may have recognized that moderate doses of accountability for marginally effective teachers quickly transformed into intense accountability pressures, including dismissal. As a result, evaluators may have rated teachers as Proficient even
if their subjective assessment of their teaching performance was below standards of proficiency to avoid their dismissal. They may have been particularly remiss to dismiss a marginally effective teacher if their projection of the range of skill levels from which they would be able to recruit for in the replacement market would not be equivalent to that teacher (recognizing that Condition 5 does not hold). In yet a different interpretation, they may have understood that placing promising, but non-proficient, teachers in the Needs Improvement rating would hamper their skill development due to reasons of psychological protection and stress-induced performance failures (recognizing that Condition 6 does not hold). All of these theories accord with findings that principals privately report substantially more teachers as performing below standards in low-stakes assessments than on high-stakes evaluations (Grissom & Loeb, 2017; Kraft & Gilmour, 2017) and Donaldson and Woulfin’s (2018) conclusions that principals frequently modified the requirements of the evaluation system to fit their strategic needs.

Thus, there is plausible evidence on the potential for accountability to improve effort for low-performing teachers, and modern evaluation systems as designed might create accountability pressures. However, in practice, the application of accountability pressures on individual teachers was uneven and generally minimal.

**Condition 3. Skill Development Effects of Evaluation**

Taylor and Tyler’s (2012) work in Cincinnati and Phipps (2018) in DCPS are, to my knowledge, the only studies that credibly estimate the causal impact of teacher evaluation on improvements in student learning outcomes. Relying on differential timing of the introduction of

---

6 Steinberg and Sartain (2015) evaluate the effects of an experimental rollout of teacher evaluation in Chicago on overall school outcomes. These estimates combine the effects of individual teacher skill improvements with compositional changes to the teaching force within schools and experience-based productivity increases. Nevertheless, their estimates are of nearly identical magnitudes to Taylor and Tyler (2012) in reading (0.10-0.13 SD), with imprecisely estimated positive coefficients in math. Burgess, Rawal and Taylor (2019) examine a peer observation scheme in England which they describe as “peer evaluation.” They find that teachers receiving feedback improved their
intensive evaluation practices, Taylor and Tyler estimate that students improve by 0.11 standard deviation units in math when taught by a teacher who has been evaluated compared to a similar teacher who has not been evaluated. They find that these gains in teacher effectiveness persist in the year following the evaluation period, suggesting that evaluations build skill rather than motivate. Importantly, their analyses focus on mid-career teachers for whom the stakes of the evaluation are “limited to promotions or additional tenure protection, or, in the case of very low scores, placement in the peer assistance program with a small risk of termination” (Taylor & Tyler, 2012, p. 3633). The skill gains observed for Cincinnati’s experienced teachers emerged in a relatively low-stakes context, with the probability of rewards far superseding that for sanction. Thus, I characterize the observed skill improvements as resulting primarily from the feedback and coaching aspects of evaluation, rather than its accountability dimensions.

As I describe above, Phipps (2018) and Phipps and Wiseman (2019) use differential timing in evaluation visits within observation windows to separately identify the effects of accountability pressures and returns to repeated feedback and coaching sessions as part of the evaluation process. Phipps looks at effects on student outcomes, whereas Phipps and Wiseman examine responses in teaching practices. Phipps and Wiseman find that each subsequent observation results in ratings improvements between 0.04 to 0.16 standard deviation units (SDs). Phipps observes that teachers receiving feedback under no accountability threat improve their value-add scores, which he interprets as improvement from evaluator advice. When scaled to a full year of feedback, these results imply value-added score gains of 0.06 and 0.03 SDs in reading and math. These magnitudes contributions to student test-score learning gains by roughly similar levels as teachers in Cincinnati did (0.07-0.09 SD). I consider these results informative to estimating the effect of teacher evaluation on student learning, but more similar in substance to instructional coaching (Kraft, Blazar, & Hogan, 2018) as these peer observations occur outside the formal evaluation process. The magnitude of these peer-coaching effects on student test-score gains are similar to those Papay, Taylor, Tyler and Laski (2020) find in Tennessee (0.12 SD).
are smaller but similar to Taylor and Tyler (2012). Thus, there is suggestive evidence that teachers can improve their skills and student outcomes in response to supervisor coaching.

While states generally require educators to receive professional development or coaching as a consequence of their coaching plans (Steinberg & Donaldson, 2016), evidence suggests that these interventions did not result in improvements in supervisor feedback or teachers’ instructional skills (Garet et al., 2017), violating the core premise of Condition 3. Other national evidence highlights that the introduction of higher-stakes evaluation had no detectable effect on teachers’ practices as they relate to student behavior in the classroom (Liebowitz, Porter, & Bragg, 2019). In the largest-scale assessment of higher-stakes teacher evaluation systems to-date, results suggest that even in locales where high levels of technical support and expertise exist, the policies’ overall effects on student achievement outcomes were effectively nil (Stecher et al., 2018).

New teacher evaluation processes in the Boston Public Schools, the largest district in Massachusetts, were also implemented with low levels of intensity. Kraft and Christian (2019) review survey evidence and find that teachers perceived the district-wide implementation of evaluation to be weak and even fewer teachers report improving their instruction based on evaluative feedback. Further, Kraft and Christian leverage a staggered randomized professional development roll-out to estimate the causal effects of a training intended to improve evaluation quality. They find it produced no gains in either perceived evaluator quality, teacher performance or student outcomes. Thus, again, teacher evaluation may have the potential for generating skill development in principle, but in execution it is less clear that these results have been realized for the majority of teachers subject to present-day evaluation systems.
Condition 4. Overall Labor Demand and Supply Effects

Depending on the structure of evaluation systems, teacher evaluation policies may affect teacher turnover or the number of prospective teacher candidates. Such shifts might generate either gluts or shortages in the teaching labor market.

Several empirical studies and simulation evidence point to the effect that higher-stakes teacher evaluation policies have on the overall demand for teachers and prospective or current teachers’ willingness to enter or remain in the teaching labor market. On the demand side, Strunk, Barret and Lincove (2017) document an increase in the rate of teachers’ exit from the profession in the aftermath of the elimination of tenure protections in Louisiana. They further find that the most dramatic increase in exits come from low-performing schools. Similarly, reforms to teacher evaluation and tenure in Michigan resulted in little overall changes to rates of attrition, but higher exit rates for teachers assigned to hard-to-staff schools (Brunner, Cowen, Strunk, & Drake, 2019).

Other evidence on the supply side finds that prospective teachers may be more hesitant to enter the profession in the presence of higher accountability. Rothstein (2015) simulates a variety of merit pay and teacher dismissal policies in the context of a dynamic choice model in which teachers must assess their career prospects in and outside of the teaching force and make a decision about whether to enter or remain in the market. Rothstein finds that at high rates of tenure denial, substantial shortages emerge. Bowen, Buck, Deck, Mills and Shulls (2015) find that prospective teachers are substantially more risk averse than other professionals which provides suggestive evidence that potential teachers may be less likely to enter the labor pool if the profession were made less stable through higher-stakes evaluation policies. Much depends in these scenarios on difficult-to-observe risk preferences. Recent work by Kraft, Brunner, Dougherty and Schwegman (2020) tests the effects of the introduction of tenure reform and high-stakes evaluation policies.
They find that state-level changes in evaluation policies resulted in a decline in teachers receiving licensure and completing teacher preparation programs by up to 17 percent.\footnote{Here and for the subsequent condition, Massachusetts serves as an imperfect national representative given that, at least at the aggregate, it often reports an overall surplus of teaching candidates (Sutcher, Darling-Hammond, & Carver-Thomas, 2019) and more importantly no estimates of the effects of teacher evaluation on labor market outcomes exist.}

Given the preceding evidence, there is potential concern that teacher supply could be constrained under a condition of high-stakes evaluation. Such a shortage in the overall supply of teachers might result in declines in the overall quality of instruction unless the quality of teachers in the labor market pool improved.

**Condition 5. Compositional Effects on Labor Market**

In addition, to overall effects on the demand and supply of teachers, high-stakes teacher evaluation may alter the human capital skills of entrants to and exiters from the profession. Some work on the effects of high-stakes teacher evaluation finds that it increases average teacher effectiveness through changes to the composition of the teaching force. In Washington, DC (Adnot, Dee, Katz, & Wyckoff, 2017; Dee & Wyckoff, 2015; James & Wyckoff, 2020), for pre-tenure teachers in New York City (Loeb, Miller, & Wyckoff, 2015), and for low-performing teachers in Houston (Cullen, Koedel, & Parsons, 2019), higher-stakes evaluation systems led to a higher rate of exit for less effective teachers and greater rates of retention for highly-rated teachers. Neither the projects in New York nor DC directly estimate the global effects of evaluation on student outcomes; however, Adnot et al. find that teacher exits under IMPACT resulted in hires who improved student learning rates between 0.05 and 0.08 standard deviation units more than their departed colleagues. In Houston, on the other hand, the positive effects on the teacher labor market were small enough in magnitude that Cullen and co-authors find no detectable impact from the reform on student achievement.
Separate bodies of research find that when it is more difficult to secure a position in the labor market outside of teaching, more effective teachers enter the profession. For example, Nagler, Piopiunik and West (2020) apply the Roy occupational choice model to the teacher labor market in Florida. They find that during periods of economic contraction, teachers who enter the profession are more effective in raising student test scores than teachers entering at other moments. They conclude that when high-skill individuals assume that the returns to their skills in other sectors will be lower, they are more likely to enter the teaching profession since demand for these positions is counter-cyclical.

Depending on the specifics of a stringent evaluation policy, this might have two implications for the skill composition of the teaching pool. If teacher evaluation led to high rates of dismissals of individuals whose low skill levels in teaching were both accurately estimated and well-correlated with skill in alternate professions, it would lead to a decline in the skill level of the non-teaching labor market and an increase in the skill demands of the teaching labor market. One would expect that this would generate positive selection into teaching. On the other hand, if the higher rates of dismissal were of teachers who were either incorrectly identified as having low-skill or who had low teaching skill, but this was poorly correlated with skills outside teaching, such a policy would lead to negative selection. There would be a greater supply of workers in the non-teaching pool, including some with high skills, and a greater demand for teachers, but with added risk in the hiring decision and deselection only weakly linked to skills profiles.

What is evident from the brief review of the teacher labor market literature in this and the preceding condition is that while some empirical support suggests teacher evaluation can result in improvements in teaching quality, through deselection, differential attrition and teacher
improvement, the benefits on student learning outcomes are far from certain. Much depends on the values of underlying human capital development and labor market parameters.

**Condition 6. Interactions between Growth and Accountability**

Teacher evaluation polices that combine growth and accountability elements either assume that these goals must be balanced along a linear continuum, or that they present opportunities to reinforce each other and accelerate improvement in the skills of the teaching force. However, it is not self-evident that designing a system that, for a particular educator, holds high standards and assigns meaningful consequences for meeting or not meeting these standards can simultaneously support that educator’s growth. Even if the accountability component of evaluation is a helpful complement to its growth component for some teachers, the negative effects of high-stakes evaluation on teachers’ capacity to improve and on the composition of the labor market pool may swamp the potential benefits. In fact, while there is evidence that evaluators can distinguish between teacher effectiveness and that teacher deselection may improve the composition of the teaching pool, universally applied high-stakes accountability may risk labor market shortages and hamper the development of marginally effective teachers.

As the causal education literature is largely silent on the interaction effects of growth and accountability in teacher evaluation, I turn to experimental and other causal evidence from management, social psychology and behavioral economics. Here, I find evidence that attaching high-stakes to a performance task, particularly punitive stakes, can impede performance. These result from tradeoffs between short- and long-term motivation in external motivation schemes, negative effects generated by high-stakes situations, and motivational responses to stress.

---

8 Two small-scale, qualitative studies (Donaldson & Mavrogordato, 2018; Reinhorn, Moore Johnson, & Simon, 2017) are the only explicit discussions of which I am aware that attempt to understand mechanisms through which school leaders integrate or tradeoff the developmental and accountability aims in teacher evaluation.
**External and internal motivation.** A long tradition of social psychology research has attempted to understand the relative merits of external and internal motivation in task performance. Deci, Koestner and Ryan (1999) review 128 experimental studies of the effect of extrinsic rewards on intrinsic motivation. They find that the introduction of performance-contingent rewards reduces intrinsic motivation (-0.28 SD), though this effect was stronger for children than adult college students. Gagné and Deci (2005) extend the principle of cognitive evaluation theory to the workplace setting and synthesize evidence from multiple studies indicating that intrinsic motivation and self determination are more effective in predicting task persistence and skill development, whereas controlled motivation will yield poorer performance on tasks requiring autonomous motivation.

Gneezy, Meier and Rey-Biel’s (2011) review of the evidence on incentives in education finds them valuable to alter effort but not skill. With respect to teacher incentives, rewards on tasks that require only the application of additional effort result in improved teacher performance (e.g., Glewwe, Ilias, & Kremer, 2010; Muralidharan & Sundararaman, 2011); however, tasks that require development of skill do not improve in response to external motivation (Gneezy et al., 2011).

**High-stakes settings.** In fact, high stakes may produce negative outcomes, particularly when the stakes involved are large in nature. A large body of social psychology literature explores the effects of anxiety on cognitive performance (e.g., Derakshan & Eysenck, 2009; Eysenck, Derakshan, Santos, & Calvo, 2007). Ariely and co-authors (2009) find in lab experiments that the greater stakes attached to a task, the more performance deteriorates, and this is particularly the case in tasks that require higher-degrees of cognitive performance. Experiment 2 in their 2009 study compared effects of incentives on routine key-pressing tasks to challenging mental arithmetic tasks. They found that higher stakes result in better performance on the low-cognitive-
demand tasks and worse performance on high-cognitive-demand tasks. Eysenck and co-authors (2009; 2007) note that performance may not decrease when tasks are low-skill or when individuals are able to compensate for anxiety by increasing effort and processing resources. However, in their review of the literature, they note that for otherwise anxious individuals this proves often to be too challenging a task. Thus, absent knowledge of teachers’ psychological profiles, an appraisal system may be hard-pressed to differentiate conditions in which stress will produce positive or negative results.

**Resistance to feedback and reduced motivation.** Individuals in an employment setting in which poor performance may result in negative consequences may respond either by increasing effort and skill development or may attempt to preserve their psychological safety by dismissing or resisting supervisor feedback. The management and human resource literatures have devoted considerable attention to employee supervision. Early work by Cleveland and Murphy (1989; 1995) and Beer (1987) documented the widespread use of interim employee evaluations across industries and noted some of the tensions between its purpose for establishing work motivation and encouraging employee development. Liden and Murphy (1985) were one of the first to causally test the role of feedback on motivation in a small laboratory study. They found that negative feedback which assigned internal causes to poor performance demotivated experiment participants, while feedback that identified external sources as the reason for poor performance in feedback did not diminish motivation.

The personnel economics literatures have also devoted substantial investigations to single-stage and dynamic tournaments in the workforce setting in which employees compete over time to advance their careers or earn more. Ederer (2010) summarizes the typical tradeoffs associated with interim performance evaluations: revealing information on employee skill through evaluation
may increase motivation (and retention) among skilled employees, but may encourage decreased second-period effort among poorly rated employees. Ederer demonstrates that while a full-feedback evaluation model is more efficient than a no- or partial-feedback model, a full-feedback model nevertheless depresses lower-rated employees’ motivation and effort, particularly if it reveals information about employees’ abilities. Thus, some compelling theory and empirical evidence suggests that performance evaluation for growth might best be understood as a substitute for evaluation for accountability.

**Discussion**

The size of the potential benefits of joint-aim teacher evaluation scheme depends on the magnitude of the effects of the six different conditions detailed above. In Table 1, I summarize the evidence from this empirical review on the implementation of joint-aim evaluation policies and the magnitude of each condition’s contribution to student learning outcomes. There is substantial noise in yearly value-added measures and weak correlations between overall evaluation scores, student test-score outcomes and other important behavioral outcomes. The small number of credibly causal studies that attempt to isolate the potential benefits of accountability-driven or growth-oriented evaluation schemas estimate positive effects roughly equivalent in magnitude. These effects are in the range of 0.03 to 0.14 standard deviation units; though in all cases the underlying study designs do not account for the potential labor market effects or the potential reinforcing or counterproductive effects of joint-aim evaluation policies. The short-run labor market effects of evaluation policies on supply, demand and skill composition appear somewhat smaller in magnitude than the direct effects on current teachers.

Interactions between growth and accountability are more difficult to quantify. No studies explicitly test these phenomena in an authentic educational setting; the ones on which I report
emerge from laboratory experiments or the private sector. Thus, in Table 1 I present qualitative descriptors alongside the quantitative estimates for the other conditions. The evidence suggests that for tasks on the effort margin, accountability and incentives may improve performance. One implication of this finding may be that accountability-oriented teacher evaluation may be most productive for teachers determined to be the lowest-performing via a multiple-measure evaluation framework. Conversely, for tasks on the skill development margin, accountability may be counter-productive. Future research should do more to estimate the magnitude of the cross-productivity effects (positive or negative) for skill-acquisition tasks.

However, the extent to which the current empirical evidence base provides meaningful insight into the magnitude of these effects is still indeterminate. In addition to standard caveats about the generalizability of findings across various contexts, the evidence base I review faces additional limitations. Research on teacher evaluation and coaching frequently emerges from contexts in which there was intensive implementation. Additionally, the labor market effects related to Conditions (#4) and (#5) are almost always studied only over the short-term when, in fact, their theorized effects are likely to manifest over spans of a decade or longer. Thus, the magnitude of these effects depends greatly on assumptions about how and whether teachers improve from evaluation and the future labor market composition of teachers.

I argue that the primary contribution of this review is to highlight the extent to which the accountability and growth aims of teacher evaluation policy support or undermine each other. Teacher evaluation policy has rarely explicitly considered these interactions. In fact, for teachers practicing at levels falling below standards outlined in instructional performance rubrics the design of policy may explicitly promote conflict between these two aims. This may take the form of either
rating inflation or of accountability crowding out potential for growth among marginally effective teachers.

Given the potential that a teacher evaluation system that imposes accountability on all teachers may impede some teachers’ growth, an alternative system in which evaluation serves as an accountability floor for some and a developmental process for most, with clear distinctions between the two populations of teachers, may resolve some of these tensions. In such an evaluation system, the large majority of teachers would be subject to an evaluation scheme directed exclusively towards professional growth. This portion of the evaluation scheme would offer targeted supports and opportunities for mentorship depending on teacher appraisals. Formal ratings (either quantitative scores or qualitative performance-level assignment) of teachers in this category need not occur and teachers might instead receive a narrative review of their practice. A much smaller group of teachers, falling below a bright line threshold would participate in a separate type of evaluation scheme in which the primary focus was on accountability for performance improvement. While some supports for growth might exist for educators in this range, teachers who did not improve within a defined period would be subject to reassignment or termination.

For such an evaluation framework to be maximally effective, several of the conditions articulated above could be either jettisoned or relaxed. The reliability and validity of evaluation ratings (Condition 1) would only be critical for teachers performing below or near the accountability floor. Given strained administrative capacity to conduct rigorous evaluations across the teacher performance distribution, greater attention could be allocated around the accountability margin. Multiple measure systems that incorporate student learning outcomes, observations, surveys and other measures could concentrate their efforts to achieve validity and reliability at the threshold point. Accountability pressures (Condition 2) would matter only for teachers near or
below the floor—those who Pope (2019) finds are most responsive to these pressures. Teachers performing above the floor would not need to be assigned ratings as long as supports for their professional development were guaranteed (Condition 3). Such a model might achieve the same theoretical benefits of positive selection into the profession as it would discourage those who projected themselves as unlikely to exceed the accountability threshold from entering the labor pool (Condition 5). The clear dividing line might return a sense of stability to risk-averse teachers and avoid some of the labor supply challenges of Condition 4. A potential determinant of the success of such an evaluation policy design is whether accountability pressures and growth supports increase or decrease the marginal benefit of the other (Condition 6), which motivates my call for further research in this area.

However, there are clear implementation challenges to a divided-aim teacher evaluation policy. Distinguishing between teachers who receive accountability pressures as opposed to growth supports will remain difficult to accomplish with perfect reliability and validity. Teachers in such an evaluation scheme may be skeptical that they are not being evaluated for accountability purposes; thus compromising many of the potential internal motivation and psychological safety gains from above. Perhaps most critically, gains from growth supports of the magnitude documented above (0.03 – 0.13 SD) occurred in contexts where there were financial and time resource commitments to regularly observe teachers, hold post-observation conferences, share written and verbal feedback and identify specific areas for skill development. Current evaluators may lack both the time and capacity to engage in such practices. On the other hand, an evaluation approach that limits accountability to a small, defined group of (predominantly) early-career teachers may be more likely to achieve political support than one that applies accountability pressures to all.
In general, however, I interpret the results of this empirical review as providing suggestively positive evidence for a teacher evaluation system that imposes a performance floor, under which teachers would be subject to accountability pressures, and above which teachers would be given clear signals that they were subject to no accountability but would intensive receive coaching and other instructional supports.
References


Table 1
Assessment of Conditions for Successful Joint-Aim Growth and Accountability Evaluation Policies

<table>
<thead>
<tr>
<th>Condition</th>
<th>Empirical findings and policy implementation evidence</th>
</tr>
</thead>
</table>
| #1 Reliable and valid evaluations | Yearly measurement error of 0.183 SD<sup>a</sup>  
Correlation of evaluation rating and VAM score = 0.3<sup>b</sup>  
Correlation of test-score VAM and behavioral-VAM = 0.15<sup>c</sup> |
| #2 Evaluations create accountability; improve effectiveness | Exists in most state evaluation frameworks, but poor ratings rare<sup>d</sup>  
One year test-VAM improvements: 0.03 – 0.14 <sup>SD<sup>e</sup></sup> |
| #3 Evaluations lead to supports and skill development; improve effectiveness | Evaluations, feedback and professional development occur but ineffective<sup>f</sup>  
One year test-VAM improvements: 0.03 – 0.13 <sup>SD<sup>g</sup></sup> |
| #4/5 Labor market effects | In short term, differential attrition improves teacher effectiveness: 0.05 – 0.08 <sup>SD<sup>h</sup></sup>  
Over longer term, risk of teacher shortages; effects on labor market indeterminate<sup>i</sup> |
| #6 Interactions between accountability and growth | Accountability pressures and skill development may complement low-skill tasks;  
Suggestive evidence about negative interaction effects in higher-skill tasks<sup>j</sup> |

Notes: <sup>a</sup>Rothstein (2015), Sass (2008); <sup>b</sup>Harris & Sass (2014), Harris et al. (2014), Kraft, Papay & Chi (2020), Grissom & Loeb (2017), Rockoff et al. (2012); <sup>c</sup>Gershenson (2016), Jackson (2018), Kraft (2019); <sup>d</sup>Kraft & Gilmour (2017), Steinberg & Donaldson (2016); <sup>e</sup>Phipps (2018), Pope (2019); <sup>f</sup>Garet et al. (2017), Stecher et al. (2018); <sup>g</sup>Papay et al. (2020), Phipps (2018), Steinberg and Sartain (2015), Taylor & Tyler (2012); <sup>h</sup>Adnot et al. (2017), Dee & Wyckoff (2015); Loeb, Miller & Wyckoff (2015); Cullen, Koedel & Parsons (2019); <sup>i</sup>Rothstein (2015), Kraft, Dougherty, Brunner & Schwegman (2020), Nagler, Piopiunik and West (2020); <sup>j</sup>Gneezy, Meier & Rey-Biel (2011); Ariely et al. (2009); Liden & Murphy (1985); Ederer (2010).
| Unsatisfactory | Needs Improvement | Proficient | Exemplary
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Delivers individual lessons rather than units of instruction; constructs units of instruction that are not aligned with state standards/local curricula; and/or designs lessons that lack measurable outcomes; fail to include appropriate student engagement strategies; and/or include tasks that mostly rely on lower level thinking skills.</td>
<td>Implements lessons and units of instruction to address some knowledge and skills defined in state standards/local curricula with some elements of appropriate student engagement strategies; but some student outcomes are poorly defined and/or tasks are not challenging.</td>
<td>Adapts as needed and implements standards-based units comprised of well-structured lessons with challenging tasks and measurable outcomes; appropriate student engagement strategies; pacing; sequence; resources; and grouping; purposeful questioning; and strategic use of technology and digital media; such that students are able to learn the knowledge and skills defined in state standards/local curricula.</td>
<td>Adapts as needed and implements standards-based units comprised of well-structured lessons with challenging tasks and measurable outcomes; appropriate student engagement strategies; pacing; sequence; resources; and grouping; purposeful questioning; and strategic use of technology and digital media; such that all students are able to learn and apply in authentic contexts the knowledge and skills defined in state standards/local curricula. Models this practice for others.</td>
</tr>
</tbody>
</table>

**Figure 1.** Example expectations for teacher practice across four performance levels

Figure 2. Percent of Massachusetts teachers rated in each of four summative rating categories
* only includes Race to the Top Districts