ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 19-160

Teacher evaluation for accountability and growth: Should policy treat them as complements or substitutes?

David D. Liebowitz
University of Oregon

Teacher evaluation policies seek to improve student outcomes by increasing the effort and skill levels of current and future teachers. Current policy and most prior research treats teacher evaluation as balancing two aims: accountability and growth. Proper teacher evaluation design has been understood as successfully weighting the accountability and growth dimensions of policy and practice. I detail six assumptions underlying teacher evaluation for growth and accountability and assess their reasonableness in light of empirical evidence from the personnel economics, social psychology and management literatures. I simulate a set of teacher evaluation policies and find that those that treat evaluation for accountability and evaluation for growth as substitutes modestly outperform policies that treat them as complements. The teachers' rates of learning through evaluation and the labor market effects of evaluation are critical in determining its impact. I conclude with recommendations for the design of teacher evaluation policies.

VERSION: November 2019

Suggested citation: Liebowitz, David D.. (2019). Teacher evaluation for accountability and growth: Should policy treat them as complements or substitutes?. (EdWorkingPaper: 19-160). Retrieved from Annenberg Institute at Brown University: http://www.edworkingpapers.com/ai19-160

TEACHER EVALUATION FOR ACCOUNTABILITY AND GROWTH: SHOULD POLICY TREAT THEM AS COMPLEMENTS OR SUBSTITUTES?

David D. Liebowitz[a]

University of Oregon

November 2019

ABSTRACT

Teacher evaluation policies seek to improve student outcomes by increasing the effort and skill levels of current and future teachers. Current policy and most prior research treats teacher evaluation as balancing two aims: accountability and growth. Proper teacher evaluation design has been understood as successfully weighting the accountability and growth dimensions of policy and practice. I detail six assumptions underlying teacher evaluation for growth and accountability and assess their reasonableness in light of empirical evidence from the personnel economics, social psychology and management literatures. I simulate a set of teacher evaluation policies and find that those that treat evaluation for accountability and evaluation for growth as substitutes modestly outperform policies that treat them as complements. The teachers' rates of learning through evaluation and the labor market effects of evaluation are critical in determining its impact. I conclude with recommendations for the design of teacher evaluation policies.

Keywords: education policy, teacher evaluation, labor contracts, personnel management, simulation

[a] I thank Joshua Cowen, Julie Berry Cullen, Morgaen Donaldson, Joshua Goodman, Richard Murnane, John Papay, Eric Taylor and Marcus Winters for helpful feedback on this paper. All errors are my own. Please direct correspondence to David Liebowitz at daviddl@uoregon.edu. Department of Educational Methodology, Policy and Leadership, 5267 University of Oregon, Eugene, OR 97403.

Teacher Evaluation for Accountability and Growth: Should Policy Treat them as Complements or Substitutes?

Most organizations seek to design employee evaluation systems that encourage workers to put forth maximal effort, that permit differentiated rewards and sanctions for various performance levels, and that stimulate employee skill development through feedback and learning generated as part of the evaluation process. The purported mechanisms through which present-day teacher evaluation improves 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, and (d) feedback mechanisms to improve the skills of current teachers. Most consequential teacher evaluation policies attempt to achieve a blend of accountability and developmental goals. While these multiple goals are firmly part of modern teacher evaluation policies, researchers and policy makers have reflected surprisingly little on the interactions between these mechanisms.

I argue that a clear explication of the tenets underlying a high-stakes teacher evaluation policy is critical to estimating its likelihood of success. I outline the assumptions of modern teacher evaluation designs and integrate the microeconomics, management and social psychology literatures to test the validity of these assumptions. I then draw on this empirical evidence base to assess the effect of policies that emphasize growth or accountability aims of teacher evaluation. I extend prior simulation work from Cowen and Winters (2013) and Rothstein (2015) by allowing for teachers to improve their skills as a result of evaluation, by assessing the effect of evaluation policies on outcomes beyond test scores, and by explicitly modeling potential interactions between accountability pressures and skill development in teacher evaluation designs.

To begin, I describe the common components of modern teacher evaluation (Section I). I then explicate six assumptions underlying current teacher evaluation policies and assess the reasonableness of these assumptions in light of the empirical literature (Section II). I highlight that the theories of action and the empirical literature motivating incentive- and sanction-based evaluation policies are in potential conflict with theories of action and empirical evidence about employees' skill development. In Section III, I introduce a framework for an alternate model of teacher evaluation. I then simulate teacher evaluation policies and vary estimation parameters in order to reflect frameworks that emphasize growth or accountability (Section IV). Through these simulations, I attempt to model the effects of teacher evaluation policies that provide only development support, policies that combine accountability pressures with skill development supports, and policies that apply accountability pressures to one group of teachers and skill development supports to a different group. I describe the second policy design as treating evaluation for accountability and evaluation for growth as complements and the third design as treating them as substitutes. I conclude in Section V that an evaluation policy that treats evaluation for growth and accountability as substitutes is more likely, across various plausible scenarios, to produce improved student outcomes, though these differences are relatively modest in magnitude. Such an evaluation policy would be less resource intensive than one promoting purely growth and potentially more politically feasible than one emphasizing accountability and growth.

I. Teacher Evaluation Policy Goals and Practices

Modern teacher evaluation policies generally attempt to hold teachers accountable to standards for quality instruction and to create a process through which teachers can improve their skills. As Murnane and Cohen (1986) document, the traditional rationale for incentive-based merit pay, emerging from the microeconomics field of contracts, is that workers' preferences are not

perfectly aligned with their employers' and monitoring worker output and actions is difficult and costly. Instead, 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 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, as I discuss below the vast majority of modern teacher evaluation systems employ a mix of both "new" and "old" style appraisals of teacher effectiveness.

Others have highlighted the potential of teacher appraisal as an opportunity for skill development. This viewpoint 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 understanding, appraisal creates a formal structure for the supervisor to provide coaching and opportunities for the worker to self-reflect on ways to improve her skills. Modern teacher evaluation policies generally attempt to maximize both of these aims.

In response to incentives from the Obama administration's Race to the Top program in 2009, 44 state legislatures across the country implemented reforms to their teacher evaluation systems (Kraft, Brunner, Dougherty, & Schwegman, 2019). 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 (Jacobs & Doherty, 2015; 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 such as the Danielson Framework (1996) or the CLASS rubric (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 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.[1] 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 some dimensions of its educational system, its evaluation guidelines are typical of many states and provides a helpful policy example to which I return throughout the paper.

[1] 44 states implemented teacher evaluation ratings with at least three different categories (Ross & Walsh, 2019). Despite differences in nomenclature, most imply ratings of Ineffective, Needs Improvement, Proficient and Exemplary.

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), and as I discuss below, most educators continued to receive positive appraisal ratings. However, it is important to note that poor appraisals risk significant consequences in the majority of states. Over three-fifths (61 percent) of states instituted rules that led to the dismissal of teachers who were not rated Proficient and almost half (48 percent) of states use evaluation results to grant or revoke tenure (Steinberg & Donaldson, 2016).

To summarize the salient features of state and local teacher evaluation policies at the end of the 2010s: (a) on paper they set high standards for quality teaching; (b) their stated aims are for both development and accountability; and (c) the failure to receive a designation meeting high bars of proficiency should lead, ultimately, to dismissal.

II. Assumptions Underlying Modern Teacher Evaluation Systems

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. Different policy choices might emphasize accountability-incentive or growth systems. 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. In those cases where commentators consider the potential interactions between accountability and growth, they are implicitly understood to be mutually reinforcing; i.e., accountability motivates teachers to improve or

coaching supports justify potential dismissal for failure to improve. However, the assumptions underlying such joint-aim evaluation system have been to-date poorly explicated.[2]

For teacher evaluation with a focus on both accountability and incentives *and* the developmental process to be maximally effective to promote improvement in student learning outcomes, I argue that the following conditions must hold:

1. **Reliable and valid evaluations.** Evaluation ratings reliably distinguish between teachers who facilitate higher and lower levels of learning for students across a set of meaningful outcomes;
2. **Evaluations improve teaching through accountability/incentives.** Classification of teachers into different rating categories produces consequential outcomes for some (or all) teachers that promote either (a) improvement in teaching practices through increased effort and skill acquisition; (b) contingent rewards (pecuniary and non) for effective teaching; (c) re-assignment of some teachers into alternative positions that hold greater expectations for success; and (d) exclusion from the teaching profession for teachers deemed ineffective;
3. **Evaluations improve teaching through skill development.** Results of evaluations generate meaningful developmental supports for teachers and opportunities for highly rated teachers to both refine advanced practices and disseminate knowledge and skills;
4. **Evaluation outcomes do not overly tax or constrain the supply of teachers.** 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

[2] Firestone (2014) notes some of the operational problems with integrating internal incentives into evaluation systems with external accountability pressures. Hallinger, Heck and Murphy (2014; 2013) present direct and indirect empirical evidence on the effectiveness of high-stakes teacher evaluation, and discuss leadership and school culture obstacles to evaluation. Phipps and Wiseman (2019) and Phipps (2018) assess the effects of accountability and growth components of evaluation both separately and jointly on teaching practices and student outcomes. In this paper, I extend their work through an explicit analysis of the assumptions underlying different evaluation systems and consider the implications of treating evaluation for accountability and growth as substitutes rather than as complementary blends or requiring a tradeoff of one aim for the other.

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. **Evaluation improves quality of teachers in 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 into the profession; and
6. **Accountability and growth goals are complementary.** 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.

The preceding assumptions are not all strictly necessary to improve average student learning outcomes. For instance, unreliable evaluation ratings might still capture enough signal about teaching quality that their use would improve the overall distribution of teacher effectiveness as long as they avoided the harms outlined in the other assumptions. I argue that insofar as these assumptions underlie the design of modern teacher evaluation policies, an assessment of the evidence on them offers insight on the potential implicit tradeoffs in evaluation policy design. To the extent that policy makers are interested in maximizing the efficacy of teacher evaluation policy all six must hold. If alternative policy designs more readily satisfy some of the assumptions or do not require them, this would also be valuable for policy makers to know.[3] The remainder of this section synthesizes the evidence on these assumptions. In Section IV, I apply this empirical evidence to simulate various scenarios and policy designs to further test the assumptions.

[3] Note that, with slight adjustments, a similar list of assumptions can be generated for teacher evaluations policies that exclusively employ the tools of either accountability or growth. For instance, by dropping Assumptions 2 and 6 and revising Assumptions 4 and 5 to focus primarily on the labor market effects of dismissals, the assumptions can be repurposed for an evaluation system that intended to improve teaching performance exclusively through incentives and sanctions. Similarly, by eliminating Assumptions 3 and 6 and revising Assumptions 4 and 5 to focus on the labor market effects of attrition, the assumptions can be repurposed for a growth-only evaluation system.

Assumption 1. Evaluations generate reliable and valid ratings

The measurement and economics of education research literature have raised various methodological and substantive concerns about Assumption 1. Methodological concerns about the use of "new" style, piece-rate evaluation systems for teacher accountability purposes range from issues of student sorting (American Statistical Association, 2014; Ballou & Springer, 2015; Koedel & Betts, 2011; Rothstein, 2010, 2017),[4] to different ratings being assigned across different value-added models and tests (Guarino, Reckase, & Wooldridge, 2015; Papay, 2011), to the potential for the narrowing of the curriculum or gaming the test (Ballou & Springer, 2015). Substantively, 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, 2017). 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-score outcomes. However, as Staiger and Rockoff (2010) demonstrate, at all but the lowest levels of reliability, value-added-based dismissals can improve average teacher effectiveness.

A full consideration of whether piece-rate evaluation systems can reliably distinguish between teachers across multiple outcomes is beyond the scope of this piece. What the empirical evidence consistently finds, however, is that subjective "old"-style evaluation and objective "new"-style student-outcome-based appraisal align (Harris & Sass, 2014; Jacob & Lefgren, 2008;

[4] Though Chetty, Friedman and Rockoff (2017, 2014b) argue that concerns about student- and teacher-sorting and causal misattribution are unfounded.

Kraft, Papay, & Chi, 2018; Sartain, Stoelinga, & Brown, 2011); however, they do so imperfectly. 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 consistent across years (Doan, Schweig, & Mihaly, 2019).

While I do not attempt to formally test Assumption 1, I incorporate measurement error in value-add and evaluation scores as well as weak correlations across value-add outcomes in the simulations I perform when assessing the effects of evaluation policies on students and the teacher labor market.

Assumption 2. Evaluation ratings improve teaching through accountability and incentives

Despite the potential of accountability- and incentive-based evaluations as articulated in the contracts literature, the application of these models to the public sector have proved challenging. Murnane and Cohen (1986), Holmstrom and Milgrom (1991), and Dixit (2002) highlight that public institutions have multiple goals for which outcomes are unverifiable, and actions are only minimally observable. As Dixit notes, teachers are generally "motivated agents," such that what primarily impedes their success is skill rather than will.

Bergman and Hill (2018) and Pope (2019) examine the effects of the public release of teachers' value-added scores in Los Angeles. This event is a potential test of the pure accountability effect of piece-rate evaluation, i.e., Assumption 2(a), as it was not coupled with any feedback on teachers' practice. Pope finds that the public release of these ratings caused an increase in the effectiveness of teachers in the bottom quintile of the performance distribution on the order of 0.10 to 0.15 standard deviations. He finds no consistent (and some potential negative) effects throughout

the rest of the performance distribution. Bergman and Hill find that high-scoring students sort into classrooms of publicized, high-value-added teachers at the expense of lower-scoring students.

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 worse when there is no threat of evaluation accountability but that the increasing probability of evaluation does not affect test scores.[5] Thus 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, the design of modern evaluation policies generally satisfy the remainder of Assumption 2 in principle. While only one-fifth of school systems reward outstanding performance—Assumption 2(b)—the vast majority of evaluation systems assign teachers into one of several rating categories (Walsh, Joseph, Lubell, & Lakis, 2017). Assignment into low-performing categories generates meaningful consequences, including dismissal, satisfying 2(c) and (d). 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

[5] 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 some 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., Fryer, 2013; 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.

(Anderson, 2013; Kraft & Gilmour, 2017).[6] 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. In fact, the percentage of teachers rated Needs Improvement declined by 3 percentage points over these five years. 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 not only typical in assigning nearly universal positive ratings, but in fact does so at rates lower than two-thirds of sampled states. 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 last from 30 days to one school year in duration. Failure to make substantial progress towards Proficiency 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

[6] 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.

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).

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, the percent of teachers rated Unsatisfactory declined over time registering a trivial value of 0.3 percent (or 245 of 81,639 teachers evaluated) in 2016/17.

I hypothesize that the structure of states' and districts' policies 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 (in essence recognizing that Assumption 3 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 Assumption 5 does not hold). In yet a different interpretation, they may have understood that placing promising, but non-proficient, teachers in the Needs Improvement (or equivalent) rating would hamper their skill development due to reasons of psychological protection and stress-

induced performance failures (recognizing that Assumption 6 does not hold). All of these theories accord with Kraft and Gilmour's (2017) and Grissom and Loeb's (2017) findings that principals privately report substantially more teachers as performing below standards in low-stakes interviews than on high-stakes evaluations and Donaldson and Woulfin's (2018) conclusions that principals frequently modify 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 evaluation systems *as designed* created accountability pressures. However, the proportion of teachers dismissed *in practice*, was vanishingly small, limiting the potential for improving teaching quality through the replacement of ineffective teachers.

Assumption 3. Evaluation results improve teaching through skill development

Taylor and Tyler's (2012) work in Cincinnati and Phipps and Wiseman (2019) and Phipps (2018) in DCPS are, to my knowledge, the only studies that credibly estimates the causal impact of teacher evaluation on improvements in teachers' practice and student learning outcomes.[7] Relying on differential timing of the introduction of 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 well after the evaluation period, suggesting

[7] 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 on the Danielson (1996) Framework for Teaching rubric improved their 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 (2016) find in Tennessee (0.12 *SD*). Stecher and colleagues (2018) estimate evaluation effects on teacher practice as well but with a weaker causal claim given their comparison-group design, and the fact that evaluation policies were implemented unevenly and bundled with other human resource strategies.

that evaluations build skill rather than motivate. Importantly, for the context of this paper, 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.

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 .04 to .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 .06 and .03 SDs in reading and math. These magnitudes are smaller but similar to Taylor and Tyler (2012). Thus, there is suggestive evidence that teachers can improve their practices 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 teachers' instructional skill (Garet et al., 2017), violating the core premise of Assumption 3. In fact, 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). 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.

Assumption 4. Evaluation does not overly tax or constrain the supply of teachers

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).

On the supply side, 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 notes that the assumptions made on various parameters yield substantial variation in optimal dismissal rates, with suggested tenure denial rates ranging from 10 to 60 percent. Though Rothstein assumes a baseline linear risk-utility for teachers with respect to their earnings, Bowen, Buck, Deck, Mills and Shulls (2015) find that prospective teachers are substantially more risk averse than other professionals. Though Bowen and colleagues do not

quantify the utility function of the teachers in their study, this 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 (2019) 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.

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.

Assumption 5. Evaluation improves the quality of teachers in the labor market pool

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), 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 high-value-add teachers. Neither the projects in New York nor DC directly estimate the global effects of evaluation on student outcomes. In Houston, 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 (1951) 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 assumption 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. Given the lack of conclusive evidence in either direction, I vary these parameters across different simulations in Section IV to test the effects of various assumptions.

Assumption 6. Accountability and growth goals of evaluation either do not interact or are complementary

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 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 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.[8] Here, I find evidence that attaching high-stakes to a performance task, particularly punitive stakes, can impede performance. These

[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.

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.

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 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 supplement to evaluation for accountability.

III. An Alternate Model of Teacher Evaluation: Accountability for a Few, Growth for Most

Given the potential substitution effects between teacher evaluation for growth and accountability, an alternative system in which evaluation serves as a rigorous accountability floor for some and a developmental process for most, with clear distinctions between the two populations, 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. 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 assumptions articulated above could be either jettisoned or relaxed. The reliability and validity of evaluation

ratings (Assumption 1) would be most 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 (Assumption 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 (Assumption 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 (Assumption 5). The clear dividing line might return a sense of stability to risk-averse teachers and avoid some of the labor supply challenges of Assumption 4. One potential determinant of the success of such an evaluation policy design would be whether accountability and growth are, in fact, complements or substitutes (Assumption 6).

In the absence of the ability to empirically test the conditions in which individuals' performance deteriorates in the face of high-stakes accountability and incentive systems via exogenous policy variation, I attempt to simulate the labor market and student learning effects of various evaluation policies and their interactions with an underlying but unmeasurable population of teachers' risk sensitivities and work motivations.

IV. Simulation Evidence

I develop a stylized model of the teacher labor market that incorporates variability in teacher starting skill level, heterogeneity in teacher improvement patterns, differential attrition

patterns, and employee contracts that condition teacher employment on evaluation ratings. I draw from Winters and Cowen (2013) alternative deselection policy simulations and Rothstein's (2015) dynamic discrete choice model to construct a simulated teacher labor market and estimate the effect of a high-stakes evaluation policy on the supply-and-demand of teachers, on their skill development, and on various student-level outcomes. I draw plausible parameters on the preceding from the most current causal literature base. Where limited evidence exists, I estimate baseline, optimistic and pessimistic scenarios. Critically, the simulation introduces interactions between evaluation policies intended to promote human capital development and those that use human resource strategies to cull poor performing teachers. In so doing, I estimate the consequences of designing an evaluation system that treats teacher development and accountability as either complements or substitutes on a range of teacher and student outcomes.

In the main text of the article, I describe the basic structure of the simulation and the key differences across evaluation models. In Appendix A, I describe the full simulation process and reference Table A1 which includes all parameters from which I draw my results. I generate a starting pool of teachers with an experience profile representative of national averages. I assign teachers randomly distributed starting values for their latent ability to improve students' test score outcomes, with a mean of 0 and a standard deviation of 0.15, consistent with the empirical evidence (Chetty, Friedman, & Rockoff, 2014a; Rivkin, Hanushek, & Kain, 2005; Rothstein, 2010, 2015). I also assign a value-added estimate of teachers' ability to improve students' behavioral outcomes. Following Jackson (2018), their value-added for this behavioral index is weakly correlated with their value-add for test-score outcomes. Teachers' value-added contributions also depend on their

experience following Papay and Kraft (2015). Teachers improve rapidly in their first three years and then much more modestly through 25 years of experience.[9]

Teachers are evaluated yearly. Their evaluation score depends on a noisily observed annual estimate of their "true" latent value-added ability and a subjective observation score that is weakly correlated with their observed test-score value-added (Grissom & Loeb, 2017; Kraft, Papay, et al., 2018; Rockoff, Staiger, Kane, & Taylor, 2012). Teachers final evaluation rating is calculated as 20 percent of their standardized observed test-score value-add (objective "piece-rate" evaluation) and 80 percent of their standardized observation ratings (subjective "supervisor judgment" evaluation), a ratio reflective of many states' actual policies. The previous assumptions persist across all scenarios I estimate. I then vary parameters across three types of evaluation policies: Growth, combined Growth and Accountability, and Divided Growth and Accountability.

I assume that teachers will improve their skills from evaluation across all frameworks and scenarios. In the baseline case, following Taylor and Tyler (2012), and consistent with Phipps (2018), teachers improve by 0.11 *SD* in the first year they experience evaluation. Taylor and Tyler are able to observe teachers after initial evaluation and see no decline in their skills in the year following evaluation. I assume that more dosage of evaluation will lead to continued improvement over time, but with gradually diminishing returns to evaluation. The Optimistic and Pessimistic cases assume higher and lower bounds on total learning from evaluation.

I model a key distinction in the Combined Growth and Accountability framework where I incorporate evidence from above that some teachers may not improve under a personnel management framework that has a high-degree of accountability and growth. Here, I explicitly contrast effects of identical evaluation policies where one scenario assumes that accountability and

[9] I also simulate models in which teachers do not improve after their fourth year and find essentially identical results. This is because of slow rates of improvement after year 4 and high rates of natural attrition.

growth are Complements (B1) for all teachers, and another scenario that assumes they are Substitutes (B2) for particular teachers. In the Complements scenario, I assume all teachers who receive an evaluation improve following the same parameters as in the other two frameworks. In the Substitutes scenario, I assign half of all teachers who fall at or below the 80th percentile in their evaluation scores to not experience improvement through evaluation. There exists no empirical rationale for either the proportion of teachers who treat accountability and growth as substitutes or the evaluation percentile rank below which teachers experience accountability in their evaluation. I select these two values as reasonable, illustrative examples.

Each year, as a result of their annual appraisal I assign teachers a rank-ordered evaluation percentile. I use this approach rather than requiring consecutive poor evaluations following Winters and Cowen's (2013) finding that the latter evaluation scheme results in far fewer dismissals and student outcome gains due to year-over-year test score noise. Teachers are dismissed at varying rates across the scenarios with the smallest dismissal rates in the Growth framework, followed by the Divided scenario and finally the combined Growth and Accountability frameworks. I assume that once teachers are dismissed they do not return to the teacher labor market, though in current policy frameworks teachers can return to teaching by moving to different districts or states.

In addition to teachers who leave the labor market due to dismissal, I assume various rates of attrition from the profession. I specify across all scenarios that all teachers leave teaching after 35 years of experience. I also specify an annual exit rate in which a large proportion of early-career teachers leave the profession with a declining probability as they gain experience.

Following evidence summarized in Winters and Cowen (2013), I also assume differential quality for teachers who leave. To draw attention to the differences in the effects of attrition across the evaluation systems, I emphasize the component of differential attrition related to accountability

pressures imposed by the evaluation framework. Therefore, I assume no differential attrition in the Growth scenario. In the Growth and Accountability scenario, I assume each year 25 percent of all teachers rated in the bottom 20 percent of the evaluation score distribution leave teaching. In the Divided scenario, I assume lower rates of attrition for teachers near the margin and a smaller margin around the dismissal threshold as a result of the clear division between accountability and growth purposes for evaluation.

Finally, I incorporate the possibility that some of these evaluation policy changes may alter the composition of latent skills in the supply of new teachers. Here, I have little empirical evidence from which to specify parameters since these counterfactuals are nearly impossible to observe over the long run. Thus, I impose reasonable bounds and use the simulation results as informative on the range of these effects. I assume no labor market changes under the Growth evaluation framework. In the optimistic scenario under the combined Growth and Accountability evaluation framework, I specify that entrants into the teaching profession will gradually improve in quality. I assume similar potential positive labor market effects for the Optimistic Divided evaluation framework because the same mechanisms by which labor market quality improvements would be purported to operate (Winters & Cowen, 2013) would be at work in the Divided framework as well.

Given evidence from Rothstein (2015) and Kraft et al. (2019) on the potential teacher shortages resulting from high-stakes evaluation, I also specify a pessimistic scenario in which as a result of fewer prospective teaching candidates in the labor market pool, hiring committees must select candidates from lower in the latent skill distribution. This occurs both through greater demand due to more vacancies and the potential for risk-averse teaching candidates to withhold their labor supply. I assume that the potential negative effects on the labor market supply of teachers will be more modest in the Divided scenario than the Growth and Accountability as fewer

teachers are dismissed, fewer are needed to fill vacancies, and risk averse teachers would observe less risk due to lower dismissal rates.

I begin the simulation with a starting pool of teachers with the traits described above. I create a yearly observed value-add score and an associated evaluation score. I estimate the average true test- and behavioral-value-add skills of teachers. Then, I assign groups of teachers to be dismissed and to attrite following the previous rules. For teachers who remain, I increase their true value-add score based on gaining experience and being evaluated. Their behavioral value-add scores increase in tandem with their test value-add scores, but remain weakly correlated. I then fill vacancies created through dismissal and attrition with novice teachers who have the previously defined characteristics such that the total number of teachers remains constant over years. I iterate the simulation over forty years.

In Figure 3, I present the results of the simulation. Each panel represents a different evaluation framework and includes Baseline, Optimistic and Pessimistic parameters. In all Baseline and Optimistic scenarios, teachers improve on average by 0.11 *SD* in their first year experiencing evaluation. This is a substantial gain, and is reflected in the sharp increases in average teacher value-added estimates in Year 2 (the first year post-evaluation) of the simulation. By contrast, early average value-added improvements are more gradual in the Pessimistic models. Panel A of Figure 3 suggests that under the assumptions of the Baseline Growth model, teachers' average value-added scores would be expected to improve by 0.17 SD over the first ten years of the evaluation policy, then gradually decline before stabilizing around 0.14 *SD* for the remaining years of the simulation. The most optimistic expectations of the Growth model predict an average improvement of 0.25 *SD* effectiveness, while in the most pessimistic projections, teachers would

only experience a peak average improvement of 0.08 *SD*. For ease of comparison across models, I present the results of these simulations numerically in binned years in Table 1.

In Panel B1 of Figure 3, I present the average improvements in teacher test-score value-added effectiveness under an evaluation policy attempting to combine Growth and Accountability goals for teachers where evaluation for growth and accountability are treated as complements. Strikingly, baseline results are similar to the Growth-only model, peaking at 0.16 *SD* improvements, despite very different evaluation strategies. While the Growth-only model improvements stem largely from improvements in more experienced teachers' skills through on-the-job and evaluation-based learning (see Figure 5), the combined Growth-and-Accountability framework accomplishes effectiveness improvements in the Baseline model through rapid dismissal of low-performing early-career teachers, coupled with learning from evaluation.

The Optimistic Growth and Accountability scenario makes strong assumptions about the potential for attracting higher-skilled teachers into the profession as a result of increased prestige or rigor. This scenario projects improvements in average effectiveness through 15 years of the policy at 0.26 *SD*, before a levelling off in gains. However, in the Pessimistic scenario the negative effects of limited growth from evaluation and the shrunken labor market talent pool results in quite modest gains in teacher effectiveness, plateauing at 0.05 *SD* improvement.

Panel B2 of Figure 3 reveals that if accountability and growth purposes for evaluation are, in fact, substitutes, an evaluation framework that attempts to accomplish both simultaneously will underperform. In the Baseline model, average value-added estimates plateau around 0.15 *SD*, below both the Growth and combined Complements framework, though differences are small in magnitude (0.01-0.03 *SD*). The Optimistic Substitutes scenario outperforms the Growth-only scenario as a result of the growing talent in the labor supply pool, thought it slightly underperforms

the Optimistic Complements (B1) framework. The Pessimistic Substitutes scenario is the worst of all, with improvements reaching a steady-state between 0.02 and 0.04 *SD*.

Panel C of Figure 3 presents results for an appraisal policy that evaluates some teachers with the purpose of accountability and others for supportive-, growth-oriented purposes. Whereas in the prior scenario teachers' improvement probabilities are uncertain, this evaluation framework explicitly treats evaluation for accountability and growth as substitutes due to the unknown interaction between growth and accountability. Baseline gains are slightly higher (0.17-0.19 *SD*) than those of the Growth framework and the combined Growth & Accountability - Complements framework and clearly outperform the Accountability & Growth – Substitutes framework. The estimates on the ceiling for the Optimistic scenario are slightly higher (0.28 *SD*) than those for both combined Growth and Accountability frameworks. The Pessimistic scenario performs equivalently to the Growth-only model and outperforms both Growth-and-Accountability models as shifts to the underlying labor market trump the effects of teachers' exit from the profession.

Figure 4 presents the analogous results for measures of teachers' value-added contributions to students' behavioral outcomes. I present the equivalent binned-year estimates in Appendix Table B1. The crucial insight from these results is that due to weak correlations across outcome measures and yearly measurement error, the effects of the evaluation policies on these outcomes are substantially attenuated and somewhat more noisy. Across all three evaluation frameworks in the most optimistic scenarios, impacts reach maximal values of 0.03 *SD*. Thus, the effects of any one of these different evaluation schemes on outcomes that are the most predictive of medium- and long-term educational success are small in substantive magnitude.

In Figure 5, I present the effect that each of these evaluation policies would have on the experience profile of teachers under the Baseline assumptions. As anticipated, under the Growth-

only evaluation framework (Panel A), teachers have much more experience in the simulation due to lower rates of dismissal and attrition. By contrast, the median level of experience after 7 simulation years in both Accountability and Growth policies (Panels B1 and B2) is five years and after 15 simulation years, the 75th percentile of experience never exceeds fifteen years. Panel C presents results from the divided-purpose evaluation system in which median experience remains fairly stable throughout, between 8 and 11 years of experience. The experience profiles resulting from each of the evaluation frameworks have significant budgetary implications.

Following Rothstein's (2015) baseline contract assumption, I specify that pay returns to experience are modeled as 0.015*t. In Figure 5, I also plot the implied budgetary costs of the experience profile that each of the evaluation policies could be anticipated to have. The values on the second y-axis are represented as proportions of employee costs above a contract that employs only first-year teachers. The starting distribution of actual teachers across the country implies a 19.2 percent added cost associated with experience given the current structure of teaching contracts and experience profile of teachers. The baseline Growth evaluation framework costs consistently between 20-25 percent more than an all-novice teaching force. The baseline Growth and Accountability evaluation framework declines rapidly in cost as more experienced teachers are replaced with early-career ones such that by year 11 and onwards, it never exceeds an additional 14 percent surplus cost. Finally, in the divided Growth / Accountability evaluation framework, the costs stabilize around 16-17 percent more than a novice-only teaching force.

V. Discussion

Previous research has rarely examined the extent to which the accountability and growth aims of teacher evaluation policy support or undermine each other. Similarly, teacher evaluation policy has not 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. In this paper, I examine the assumptions underlying treating teacher evaluation for growth and accountability as substitutes or complements by surveying the existing literature and presenting results of a simple simulation. This simulation extends existing evidence by explicitly considering the conditions under which teachers improve through evaluation, rather than through dismissal alone.

In Table 2, I summarize key insights from the simulation with respect to the assumptions embedded in different evaluation frameworks. In Panel A, I present evidence from prior empirical work on the reliability and external validity of teacher evaluation and value-added measures that I incorporate into the simulations. In Panel B, I share evidence from the results of the simulations on the assumptions embedded in modern evaluation systems.

First, given weakly correlated measures of teacher effectiveness, growth in one dimension of teacher effectiveness—either through accountability or growth—will result in much more modest growth in other dimensions of teacher effectiveness. Second, small but meaningful differences in anticipated teacher effectiveness result from different types of evaluation policies, though these differences depend on particular assumptions. Under baseline assumptions, average teacher effectiveness in a growth-only evaluation policy results in 0.01-0.03 worse average teacher effectiveness than a policy that imposes accountability pressures on some teachers and growth-only support for others. The effects of policies that impose accountability pressures and growth supports jointly on teachers depend in small part on the extent to which teachers experience growth and accountability as complements or substitutes. In the former case, teacher effectiveness would

be expected to improve at levels approximately the same as the divided scenario. If teachers experience evaluation for growth and evaluation for accountability as substitutes, joint-aim evaluation policies would result in 0.01-0.03 worse teacher effectiveness than divided-aim policies. Ultimately, it is difficult to observe the underlying probability of improving under high-stakes evaluation conditions. Disentangling the probability of improving as a result of evaluation from the mean effect of evaluation presents an identification challenge. Thus, policy makers advance joint growth-and-accountability evaluation schemes under a condition of uncertainty.

Third, an important aspect of the combined growth-and-accountability evaluation policies is that as a result of higher dismissal rates they would yield greater variability in teachers' effectiveness. I plot in Appendix Figure B1 the yearly standard deviations of teacher effectiveness. After an identical starting point, there are slight increases in the standard deviation of effectiveness for the growth- and divided-evaluation policies, and standard deviation increases of up to 0.04 *SD* for the combined evaluation policies. The substantive implication is that joint-aim evaluation policies will not only produce slightly lower levels of student learning gains, but that these gains will accumulate through a widening of inequalities across classrooms.

Fourth, high rates of attrition unrelated to teacher quality limit the potential benefits of evaluation. Winters and Cowen (2013) find that the introduction of ability-related attrition mutes the effect of a value-added based deselection policy. In the growth scenario, there is no ability-related attrition other than retirement, and (almost) no performance dismissal, so improvements depend entirely on teacher skill acquisition through experience and evaluation. However, high rates of natural attrition mean that after initial improvements in performance due to growth through evaluation, when many of these teachers attrite from the profession, irrespective of skill, the average performance regresses. When I reduce the ability-unrelated attrition to one-third of

baseline levels, teacher effectiveness in the growth scenario is at or above levels in all other scenarios. In the absence of successful strategies to dramatically reduce overall attrition, however, evaluation policies must rely on some form of ability-based exit from the profession (either dismissal or differential attrition) to maximize improvement from evaluation.

Finally, and perhaps most importantly, while the simulation reveals important outcome differences across evaluation frameworks, the two most significant influences on evaluation policies' comparative effectiveness are (a) how teachers improve from evaluation and (b) how, if at all, evaluation policies affect the labor supply and composition. The main differences distinguishing the pessimistic and optimistic scenarios from baseline comparisons across the evaluation frameworks are the rate of improvement from evaluation and the quality of new teachers. Whereas the differences *between* evaluation frameworks range from 0 to 0.05 *SD*, the differences between optimistic and pessimistic scenarios *within* evaluation frameworks are around 0.2 *SD*. Evidence on learning from evaluation is limited, with no evidence on how teachers learn from evaluation over a span longer than two years. The most recent estimates of the effects of high-stakes teacher evaluation reforms suggests that while it has decreased overall supply it may have increased overall quality (Kraft et al., 2019); thus, the general equilibrium effects of such policies remain indeterminate. These two insights suggest the need for better research on learning from evaluation and the value of policy that minimizes potential negative effects on teachers' labor supply and the future skill composition of prospective teachers.

Beyond the particulars of different evaluation policy effects on test-score and behavioral value-added estimates, the simulation offers insights on the political feasibility and cost of various evaluation frameworks. Surprisingly, under reasonable assumptions, evaluation policies that dismiss almost no teachers perform broadly equivalently to policies that dismiss between 10 and

20 percent of all teachers in a given year. Differences in average teacher effectiveness across rates of dismissal are similar in magnitude to those in Winters and Cowen (2013, p. 644). This bears important consideration for policy makers given the political objections to policies that dismiss large numbers of teachers and the potential for unfairness in evaluation systems that misidentify teachers as low-skill when they are not. However, evaluation policies that dismiss large numbers of teachers will ultimately employ a much less experienced teaching force. This has important budgetary implications. As Figure 5 indicates, combined accountability-and-growth evaluation frameworks would result in substantial human resource savings, up to six percentage points of total employee expenditures. These could be reinvested in teacher salaries to counteract potential negative effects on labor supply. On the other hand, divided accountability and growth policies would represent up to 3 percentage point cost savings, whereas growth-only evaluation policies add up to six percentage points to human resource budgets.

In general, I interpret these findings as providing suggestively positive results for a teacher evaluation system that imposes a relatively low accountability 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 receive coaching and other instructional supports. However, 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 conclude that policies that treat evaluation for accountability purposes and evaluation for growth purposes as substitutes (rather than treating them as complements or prioritizing only one of these aims) have the greatest likelihood of success, both in terms of student outcomes and political feasibility.

References

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. https://doi.org/10.3102/0162373716663646

American Statistical Association. (2014). *ASA Statement on Using Value-Added Models for Educational Assessment* (Vol. 2016). Alexandria, VA. Retrieved from http://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf%5Cnhttps://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf

Anderson, J. (2013, March 30). Curious grades for teachers: Nearly all pass. *New York Times*, p. A1. Retrieved from https://www.nytimes.com/2013/03/31/education/curious-grade-for-teachers-nearly-all-pass.html

Ariely, D., Gneezy, U., Loewenstein, G., & Mazar, N. (2009). Large stakes and big mistakes. *Review of Economic Studies, 76*(2), 451–469. https://doi.org/10.1111/j.1467-937X.2009.00534.x

Armstrong, M. (2000). Performance management. In R. Dransfield (Ed.), *Human Resource Management* (p. 121). Heinemann.

Ballou, D., & Springer, M. G. (2015). Using Student Test Scores to Measure Teacher Performance: Some Problems in the Design and Implementation of Evaluation Systems. *Educational Researcher, 44*(2), 77–86. https://doi.org/10.3102/0013189X15574904

Beer, M. (1987). Performance appraisals. In *Handbook of Organizational Behavior* (Lorsch, J., pp. 286–301). Englewood Cliffs, NJ: Prentice Hall, Inc.

Bergman, P., & Hill, M. J. (2018). The effects of making performance information public: Regression discontinuity evidence from Los Angeles teachers. *Economics of Education Review, 66*, 104–113. https://doi.org/10.1016/j.econedurev.2018.07.005

Blazar, D., & Kraft, M. A. (2017). Teacher and Teaching Effects on Students Attitudes and Behaviors. *Educational Evaluation and Policy Analysis, 39*(1), 146–170. https://doi.org/10.3102/0162373716670260

Bowen, D. H., Buck, S., Deck, C., Mills, J. N., & Shuls, J. V. (2015). Risky business: an analysis of teacher risk preferences. *Education Economics, 23*(4), 470–480. https://doi.org/10.1080/09645292.2014.966062

Brunner, E., Cowen, J. M., Strunk, K. O., & Drake, S. (2019). Teacher labor market responses to statewide reform: Evidence from Michigan. *Educational Evaluation and Policy Analysis, 41*(4), 403–425. https://doi.org/10.3102/0162373719858997

Burgess, S., Rawall, S., & Taylor, E. S. (2019). *Teacher peer observation and student test scores: Evidence from a field experiment in English secondary schools* (Working Paper). Cambridge, MA. Retrieved from https://scholar.harvard.edu/files/erictaylor/files/teacher-peer-obsv-brt-jan-19.pdf

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the Impacts of Teachers I:

Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review, 104*(9), 2593–2632. https://doi.org/10.1257/aer.104.9.2593

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2017). Measuring the Impacts of Teachers: Reply. *American Economic Review, 107*(6), 1685–1717. https://doi.org/10.1257/aer.20170108

Chetty, R., Friedman, J., & Rockoff, J. (2014b). Discussion of the American Statistical Association's Statement (2014) on Using Value-Added Models for Educational Assessment. *Statistics and Public Policy, 1*(1), 111–113. https://doi.org/10.1080/2330443X.2014.955227

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology, 74*(1), 130–135. https://doi.org/10.1037/0021-9010.74.1.130

Cullen, J. B., Koedel, C., & Parsons, E. (2019). The compositional effect of rigorous teacher evaluation on workforce quality. *Education Finance and Policy*, 1–85. https://doi.org/10.1162/edfp_a_00292

Danielson, C. (1996). *Enhancing Professional Practice: A Framework for Teaching*. Alexandria: ASCD.

Darling-Hammond, L., Wise, A. E., & Pease, S. R. (1983). Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research, 53*(3), 285–328. https://doi.org/10.3102/00346543053003285

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*(6), 627–668. https://doi.org/10.1017/CBO9781107415324.004

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management, 34*(2), 267–297. https://doi.org/10.1002/pam.21818

Derakshan, N., & Eysenck, M. W. (2009). Anxiety, Processing Efficiency, and Cognitive Performance. *European Psychologist, 14*(2), 168–176. https://doi.org/10.1027/1016-9040.14.2.168

DESE. (2012). *Model system for educator evaluation Part IV: Model collective bargaining contract language*. Malden. Retrieved from www.doe.mass.eduwww.doe.mass.edu

DESE. (2018). *Massachusetts model system for educator evaluation: Classroom teacher rubric*. Malden. Retrieved from http://www.doe.mass.edu/edeval/model/PartIII_AppxC.pdf

Dixit, A. (2002). Incentives and organizations in the public sector: An interpretive review. *Journal of Human Resources, 37*(4), 696–727.

Doan, S., Schweig, J. D., & Mihaly, K. (2019). The Consistency of Composite Ratings of Teacher Effectiveness: Evidence From New Mexico. *American Educational Research Journal*, 0002831219841136. https://doi.org/10.3102/0002831219841369

Donaldson, M. L., & Mavrogordato, M. (2018). Principals and teacher evaluation. *Journal of Educational Administration, 56*(6), 586–601. https://doi.org/10.1108/JEA-08-2017-0100

Donaldson, M. L., & Papay, J. (2015). Teacher evaluation for accountability and development. In Helen Ladd & Margaret Goertz (Eds.), *Handbook of Research in Education Finance and Policy* (2nd ed., pp. 174–193). New York: Routledge.

Donaldson, M. L., & Woulfin, S. (2018). From tinkering to going "rogue": How principals use agency when enacting new teacher evaluation systems. *Educational Evaluation and Policy Analysis, 40*(4), 531–556. https://doi.org/10.3102/0162373718784205

Ederer, F. P. (2010). Feedback and Motivation in Dynamic Tournaments. *Journal of Economics & Management Strategy, 19*(3), 733–769. https://doi.org/10.2139/ssrn.691384

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion, 7*(2), 336–353. https://doi.org/10.1037/1528-3542.7.2.336

Firestone, W. A. (2014). Teacher Evaluation Policy and Conflicting Theories of Motivation. *Educational Researcher, 43*(2), 100–107. https://doi.org/10.3102/0013189X14521864

Fryer, R. G. (2013). Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. *Journal of Labor Economics, 31*(2), 373–407. https://doi.org/10.1086/667757

Gagné, M., & Deci, E. L. (2005). Self-determination theory and work motivation. *Journal of Organizational Behavior, 26*(4), 331–362. https://doi.org/10.1002/job.322

Garet, M. S., Wayne, A. J., Brown, S., Rickles, J., Song, M., & Manzeseke, D. (2017). *The Impact of Providing Performance Feedback to Teachers and Principals (NCESS 2018-4001)*. Washington, DC.

Gershenson, S. (2016). Linking Teacher Quality, Student Attendance, and Student Achievement. *Education Finance and Policy, 11*(2), 125–149. https://doi.org/10.1162/EDFP_a_00180

Glewwe, P., Ilias, N., & Kremer, M. (2010). Teacher incentives. *American Economic Journal: Applied Economics, 2*(3), 205–227. https://doi.org/10.1257/app.2.3.205

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives, 25*(4), 191–210. https://doi.org/10.1257/jep.25.4.191

Goodman, S. F., & Turner, L. J. (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. https://doi.org/10.1086/668676

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. https://doi.org/10.1162/EDFP_a_00210

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). Can Value-Added Measures of Teacher Performance Be Trusted? *Education Finance and Policy, 10*(1), 117–156. https://doi.org/10.1162/EDFP_a_00153

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability, 26*(1), 5–

28. https://doi.org/10.1007/s11092-013-9179-5

Harris, D. N., & Sass, T. R. (2014). Skills, productivity and the evaluation of teacher performance. *Economics of Education Review, 40*, 183–204. https://doi.org/10.1016/J.ECONEDUREV.2014.03.002

Holmstrom, B., & Milgrom, P. (1991). Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, and Organization, 7*(special), 24–52. https://doi.org/10.1093/jleo/7.special_issue.24

Jackson, C. K. (2018). What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test-Score Outcomes. *Journal of Political Economy, 126*(5), 2072–2107.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101–136. https://doi.org/10.1086/522974

Jacobs, S., & Doherty, K. (2015). State of the States 2015: Evaluating Teaching, Leading and Learning.

Koedel, C., & Betts, J. R. (2011). Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique. *Education Finance and Policy, 6*(1), 18–42. https://doi.org/10.1162/EDFP_a_00027

Kraft, M. A. (2017). Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies. *Journal of Human Resources, 0*(0). https://doi.org/10.3368/jhr.54.1.0916.8265R3

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, 88*(4), 547–588. https://doi.org/10.3102/0034654318759268

Kraft, M. A., Brunner, E. J., Dougherty, S. M., & Schwegman, D. J. (2019). *Teacher evaluation reforms and the supply and quality of new teachers* (Brown University Working Paper). Providence, RI. Retrieved from https://scholar.harvard.edu/files/mkraft/files/kraft_et_al._teacher_evaluation_-_updated_feb_2019.pdf

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting The Widget Effect: Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness. *Educational Researcher, 46*(5), 234–249. https://doi.org/10.3102/0013189X17718797

Kraft, M. A., Papay, J. P., & Chi, O. L. (2018). *Teacher skill development: Evidence from performance ratings by principals* (Brown University Working Paper). Providence, RI. Retrieved from https://scholar.harvard.edu/files/mkraft/files/kraft_papay_chi_2018_teacher_skill_development.pdf

Liden, R. C., & Mitchell, T. R. (1985). Reactions to Feedback: The Role of Attributions. *Academy of Management Journal, 28*(2), 291–308. https://doi.org/10.5465/256202

Loeb, S., Miller, L. C., & Wyckoff, J. (2015). Performance screens for school improvement: The case for teacher tenure reform in New York City. *Educational Researcher, 44*(4), 199–212. https://doi.org/10.3102/0013189X15584773

Macartney, H., McMillan, R., & Petronijevic, U. (2018). *Teacher performance and accountability incentives* (NBER Working Paper Series No. No. 24747). Cambridge, MA.

Martínez, J. F., Schweig, J., & Goldschmidt, P. (2016). Approaches for Combining Multiple Measures of Teacher Performance. *Educational Evaluation and Policy Analysis, 38*(4), 738–756. https://doi.org/10.3102/0162373716666166

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., … Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development, 79*(3), 732–749. https://doi.org/10.1111/j.1467-8624.2008.01154.x

Muralidharan, K., & Sundararaman, V. (2011). Teacher Performance Pay: Experimental Evidence from India. *Journal of Political Economy, 119*(1), 39–77. https://doi.org/10.1086/659655

Murnane, R., & Cohen, D. (1986). Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and a Few Survive. *Harvard Educational Review, 56*(1), 1–18. https://doi.org/10.17763/haer.56.1.l8q2334243271116

Murphy, J., Hallinger, P., & Heck, R. H. (2013). Leading via Teacher Evaluation. *Educational Researcher, 42*(6), 349–354. https://doi.org/10.3102/0013189X13499625

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives.* Thousand Oaks, CA, US: Sage Publications, Inc.

Nagler, M., Piopiunik, M., & West, M. (2020). Weak markets, strong teachers: Recession at career start and teacher effectiveness. *Journal of Labor Economics, 38*.

Papay, J. P. (2011). Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates Across Outcome Measures. *American Education Research Journal, 48*(1), 163–193. https://doi.org/10.3102/0002831210362589

Papay, J. P., & Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics, 130*, 105–119. https://doi.org/10.1016/j.jpubeco.2015.02.008

Papay, J., Taylor, E., Tyler, J., & Laski, M. (2016). *Learning Job Skills from Colleagues at Work: Evidence from a Field Experiment Using Teacher Performance Data*. Cambridge, MA. https://doi.org/10.3386/w21986

Phipps, A. R. (2018). *Personnel contracts with production uncertainty: Theory and evidence from teacher performance incentives* (unpublished Working Paper). Charlottesville, VA.

Phipps, A. R., & Wiseman, E. A. (2019). Enacting the rubric: Teacher improvements in windows of high-stakes observation. *Education Finance and Policy*, 1–51. https://doi.org/10.1162/edfp_a_00295

Pope, N. G. (2019). The effect of teacher ratings on teacher performance. *Journal of Public Economics, 172*, 84–110. https://doi.org/10.1016/J.JPUBECO.2019.01.001

Popham, W. (1988). The dysfunctional marriage of formative and summative teacher evaluation.

Journal of Personnel Evaluation in Education, *1*(3), 269–273. https://doi.org/10.1007/BF00123822

Reinhorn, S., Moore Johnson, S., & Simon, N. (2017). Investing in Development: Six High-Performing, High-Poverty Schools Implement the Massachusetts Teacher Evaluation Policy. *Educational Evaluation and Policy Analysis*, *39*(3), 383–406. https://doi.org/10.3102/0162373717690605

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, *73*(2), 417–458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

Rockoff, J. E., Staiger, D. O., Kane, T. J., & Taylor, E. S. (2012). Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools. *The American Economic Review*, *102*(7), 3184–3213. https://doi.org/10.2307/41724631

Ross, E., & Walsh, K. (2019). *State of the States 2019: Teacher and Principal Evaluation Policy*. Washington, DC.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, *125*(1), 175–214. https://doi.org/10.1162/qjec.2010.125.1.175

Rothstein, J. (2015). Teacher quality policy when supply matters. *American Economic Review*, *105*(1), 100–130. https://doi.org/10.1257/aer.20121242

Rothstein, J. (2017). Measuring the Impacts of Teachers: Comment. *American Economic Review*, *107*(6), 1656–1684. https://doi.org/10.1257/aer.20141440

Roy, A. D. (1951). Some Thoughts on the Distribution of Earnings. *Oxford Economic Papers*, *3*(2), 135–146. Retrieved from http://www.jstor.org/stable/2662082

Sartain, L., Stoelinga, S. R., & Brown, E. (2011). *Rethinking teacher evaluation in Chicago*. Chicago. Retrieved from http://www.educationalimpact.com/resources/TEPC/pdf/TEPC_Report_Chicago_Research.pdf

Sass, T. R. (2008). *The stability of value-added measures of teacher quality and implications for teacher compensation policy*. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Sojourner, A. J., Mykerezi, E., & West, K. L. (2014). Teacher Pay Reform and Productivity. *Journal of Human Resources*, *49*(4), 945–981. https://doi.org/10.3368/jhr.49.4.945

Speroni, C., Wellington, A., Burkander, P., Chiang, H., Herrmann, M., & Hallgren, K. (2019). Do Educator Performance Incentives Help Students? Evidence from the Teacher Incentive Fund National Evaluation. *Journal of Labor Economics*. https://doi.org/10.1086/706059

Staiger, D. O., & Rockoff, J. E. (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives*, *24*(3), 97–118. https://doi.org/10.1257/jep.24.3.97

Stecher, B., Holtzman, D., Garet, M., Hamilton, L., Engberg, J., Steiner, E., … Chambers, J. (2018).

Improving Teaching Effectiveness: Final Report: The Intensive Partnerships for Effective Teaching Through 2015-2016. Santa Monica: RAND Corporation. https://doi.org/10.7249/RR2242

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy, 11*(3), 340–359. https://doi.org/10.1162/EDFP_a_00186

Steinberg, M. P., & Kraft, M. A. (2017). The sensitivity of teacher performance ratings to the design of teacher evaluation systems. *Educational Researcher, 46*(7), 378–396. https://doi.org/10.3102/0013189X17726752

Steinberg, M. P., & 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. https://doi.org/10.1162/EDFP_a_00173

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*. Retrieved from https://educationresearchalliancenola.org/files/publications/041217-Strunk-Barrett-Lincove-When-Tenure-Ends.pdf

Taylor, E. S., & Tyler, J. H. (2012). The Effect of Evaluation on Teacher Performance. *American Economic Review, 102*(7), 3628–3651. https://doi.org/10.1257/aer.102.7.3628

U.S. Department of Education National Center for Education Statistics. (2017). National Teacher and Principal Survey (NTPS), Public School Principal Data File, 2015–16 and Principal Follow-up Survey (PFS). Washington, DC.

Walsh, K., Joseph, N., Lubell, S., & Lakis, K. (2017). *Running in place: How new teacher evaluations fail to live up to promises*. Washington, DC. Retrieved from https://www.nctq.org/publications/Running-in-Place:-How-New-Teacher-Evaluations-Fail-to-Live-Up-to-Promises

Winters, M. A., & Cowen, J. M. (2013). Would a value-added system of retention improve the distribution of teacher quality? A simulation of alternative policies. *Journal of Policy Analysis and Management, 32*(3), 634–654. https://doi.org/10.1002/pam.21705

Tables

Table 1. Comparison of average teacher test-score value-added in year bins under growth- and accountability-oriented evaluation policies

	Baseline	Optimistic	Pessimistic
A. Growth			
Years 2-5	0.128	0.149	0.041
Years 6-10	0.165	0.230	0.076
Years 11-15	0.175	0.249	0.083
Years 16-20	0.169	0.243	0.076
Years 21+	0.145	0.216	0.054
B1. Growth & Accountability - Complements			
Years 2-5	0.121	0.147	0.039
Years 6-10	0.153	0.224	0.054
Years 11-15	0.162	0.258	0.065
Years 16-20	0.164	0.268	0.053
Years 21+	0.168	0.270	0.057
B2. Growth & Accountability - Substitutes			
Years 2-5	0.089	0.110	0.029
Years 6-10	0.130	0.197	0.039
Years 11-15	0.144	0.246	0.033
Years 16-20	0.150	0.264	0.029
Years 21+	0.154	0.270	0.029
C. Growth / Accountability Divided			
Years 2-5	0.126	0.151	0.043
Years 6-10	0.162	0.232	0.071
Years 11-15	0.174	0.266	0.079
Years 16-20	0.177	0.277	0.081
Years 21+	0.180	0.281	0.084

Notes: average test score value-added estimates derived from simulation described in Appendix A.

Table 2. Comparison of evaluation framework assumptions

Assumption	**Evidence**
A. Empirical evidence and policy frameworks	
#1 Reliable and valid evaluations	Yearly measurement error of 0.183 *SD*[a] *r* eval rating and VAM score = 0.3[b] *r* test-VAM & behav.-VAM = 0.15 *SD*[c]
#2 Current evaluations lead to consequences/incentives	Exists in most state evaluation frameworks, but poor ratings rare[d]
#3 Current evaluations lead to supports	Provided but often ineffective[e]
B. Simulation results	
#2 Accountability pressures improve effectiveness #3 Growth supports improve effectiveness	Across all evaluation frameworks, effect of improvement from evaluation has moderate to large effect on overall test-score value-added, but small effect on other student outcome value-added; magnitudes depend greatly on anticipated rate of improvement from evaluation
#4 Overall supply and demand effects	If accountability reduces total labor supply and this reduces overall quality of prospective teachers (by assumption), effects on overall test-score value-added are large
#5 Compositional effects on labor market	If current rates of attrition and ability of attriters and entrants to profession holds, high-stakes unlikely to shift compositional quality of teaching pool; however dismissal or increased attrition of low-performers and/or small improvements in novice teacher supply and selection could have larger effects
#6 Interactions between accountability and growth	Models that attempt to maximize joint goals of accountability and growth underperform models that explicitly distinguish these goals for different teachers if teachers do, in fact, treat them as substitutes

Notes: [a] Rothstein (2015), Sass (2008); [b] Harris & Sass (2014), Kraft, Papay & Chi (2018), Grissom & Loeb (2017), Rockoff et al. (2012); [c] Gershenson (2016), Jackson (2018), Kraft (2017); [d] Kraft & Gilmour (2017), Steinberg & Donaldson (2016); [e] Garet et al. (2017), Stecher et al. (2018)

Figures

	Unsatisfactory	Needs Improvement	Proficient	Exemplary
I-A-3. Well-Structured Units and Lessons	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.	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.	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.	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.

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

Source: Massachusetts Department of Elementary and Secondary Education (DESE) (2018). Massachusetts Model System for Educator Evaluation Classroom Teacher Rubric

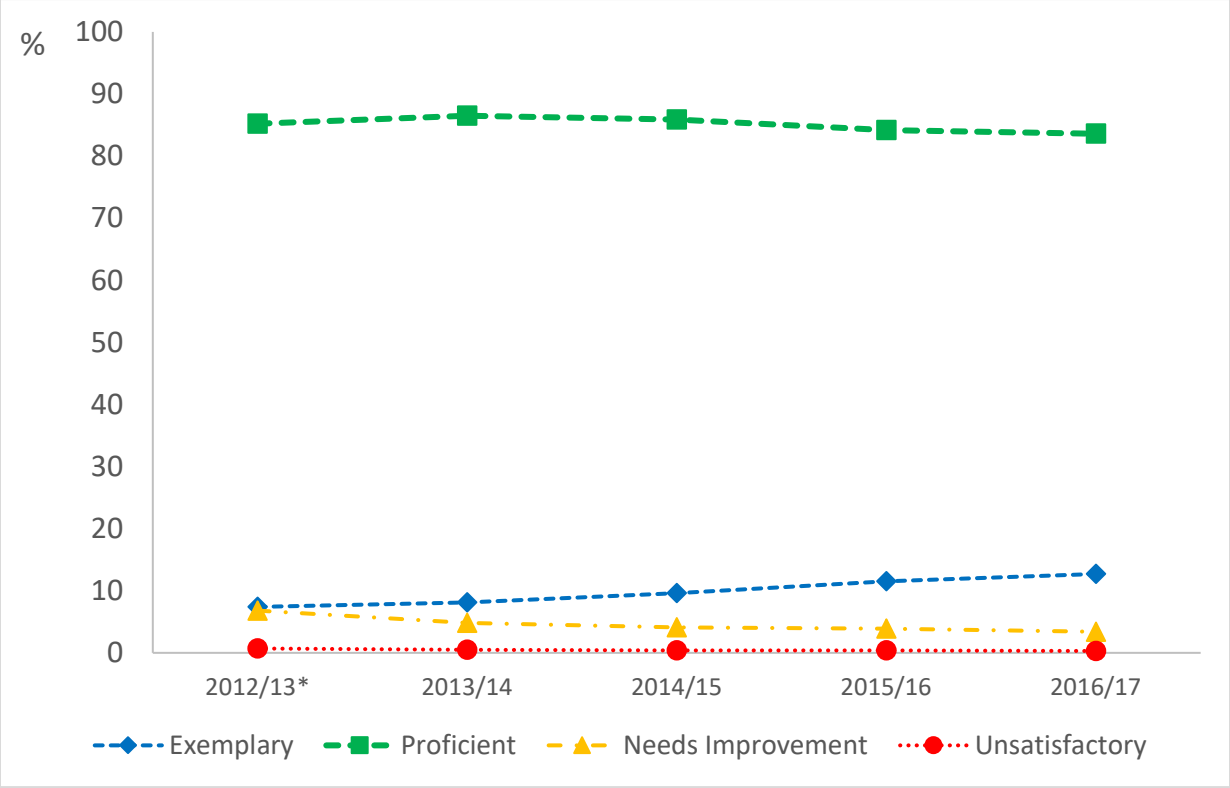

Figure 2. Percent of Massachusetts teachers rated in each of four summative rating categories

* only includes Race to the Top Districts

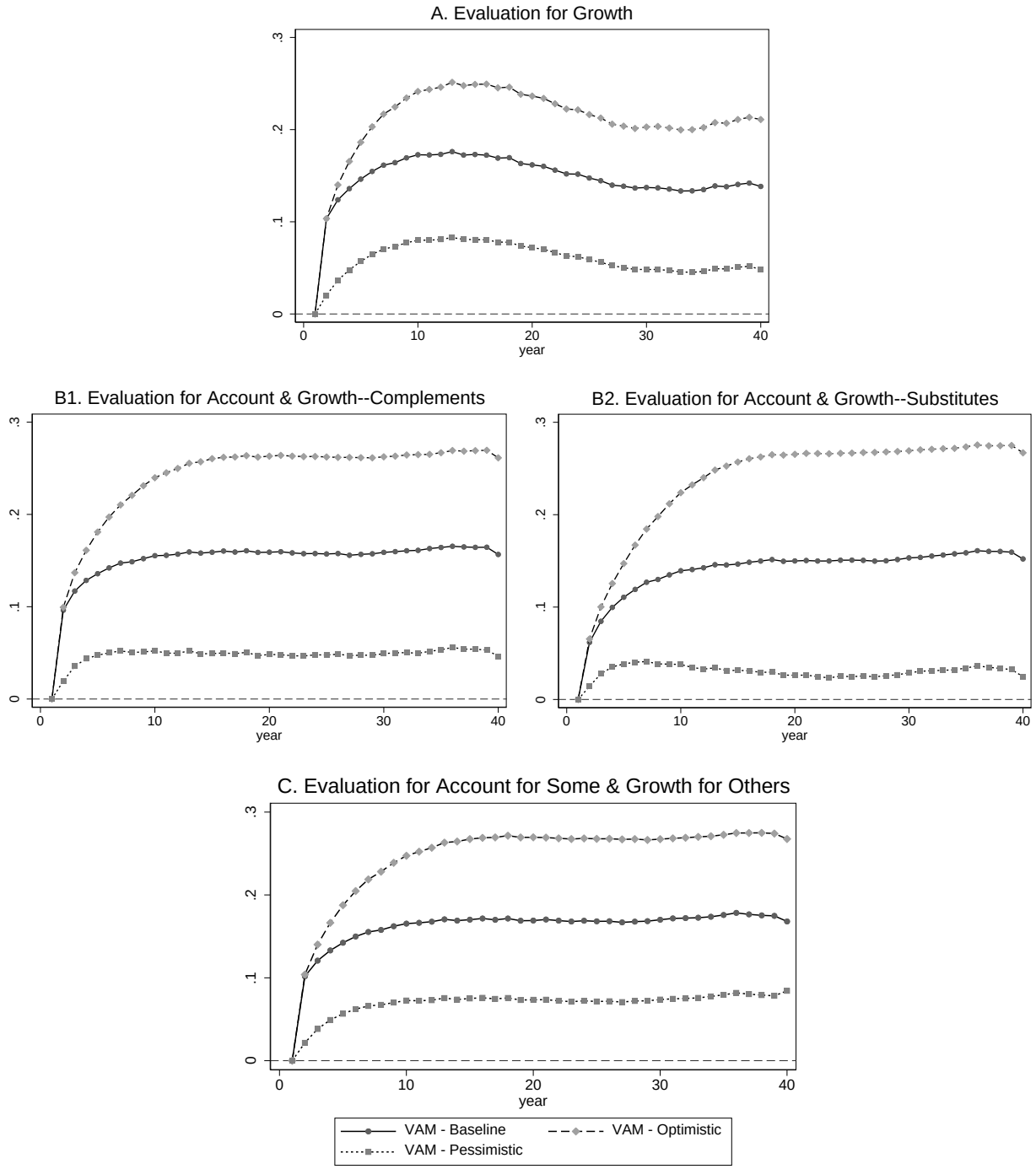

Figure 3. Average teacher test-score value-added profile over 40 simulated years under growth- and accountability-oriented evaluation policies

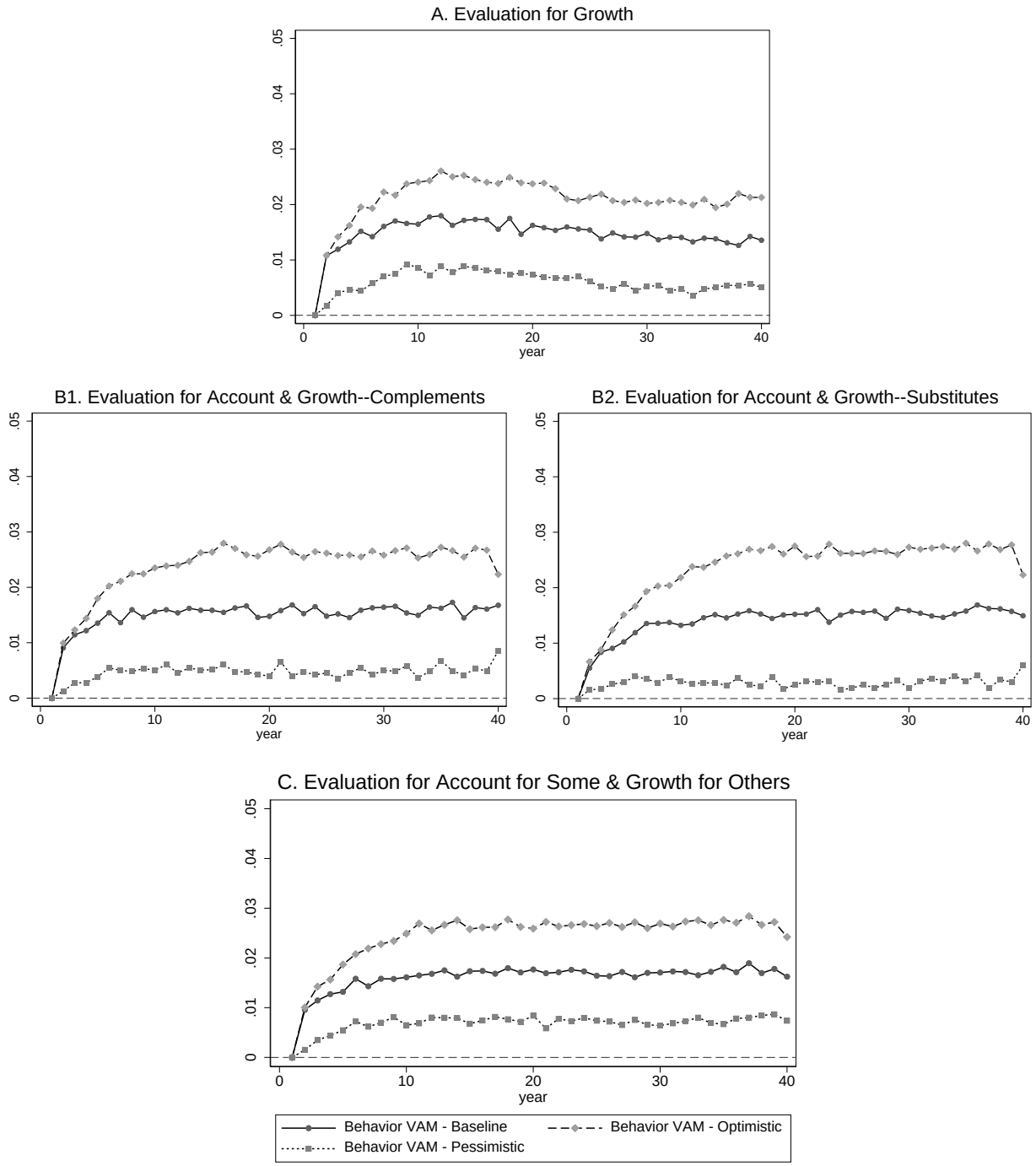

Figure 4. Average teacher value-added on index of behavioral outcomes (suspensions, absences and GPA) over 40 simulated years under growth- and accountability-oriented evaluation policies.

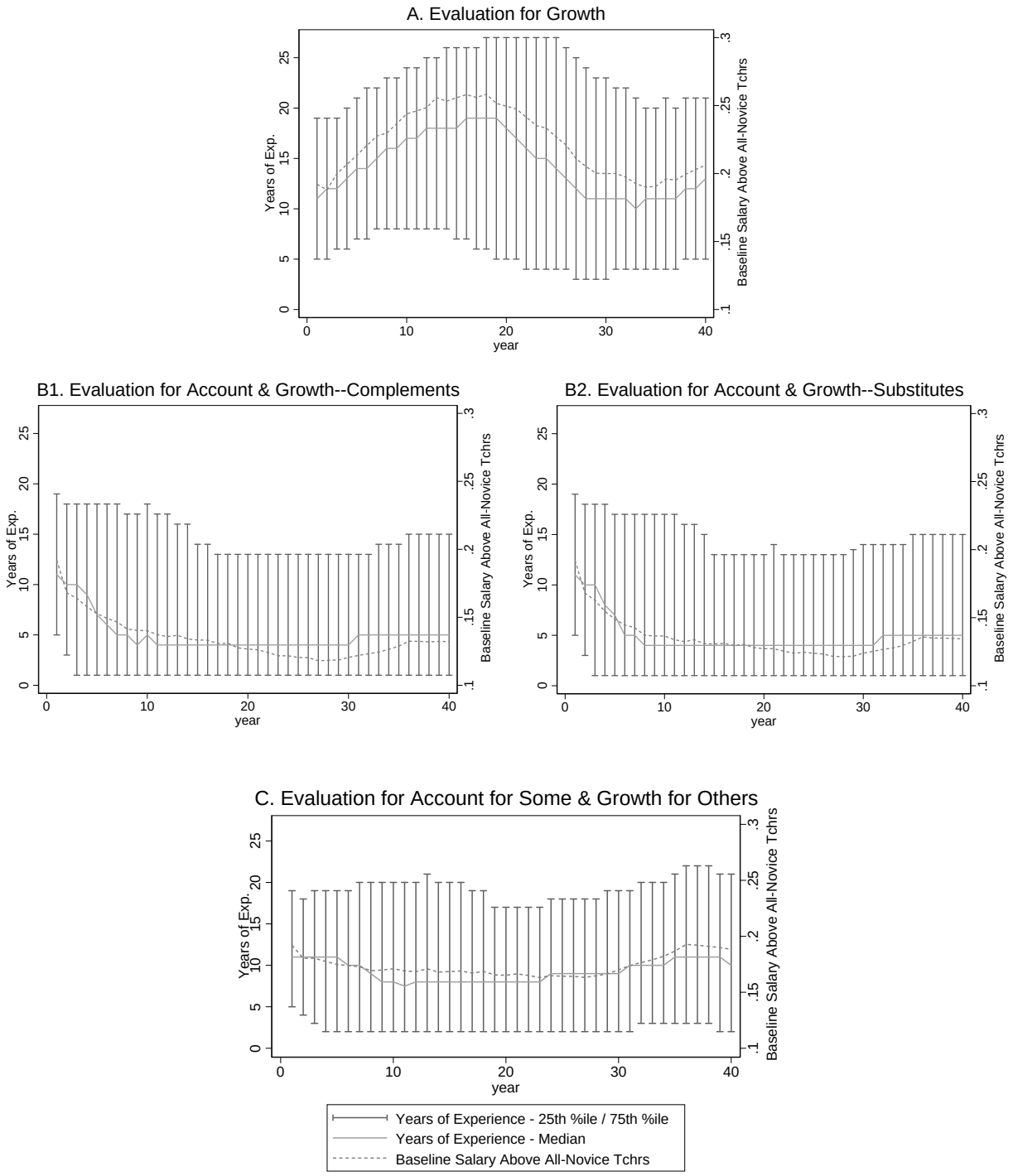

Figure 5. Teacher experience and experience-based compensation profiles over 40 years under growth- and accountability-oriented evaluation policies.

Appendix A. Simulation Description

I generate a starting pool of 25,000 teachers with an experience profile identical to those surveyed in the nationally representative 2015/16 National Teacher and Principal Survey (U.S. Department of Education National Center for Education Statistics, 2017), with experience capped at 35 years. I use the parameters from Appendix Table A1 to assign teachers starting values for their latent ability in improving students' test score outcomes. Following Rothstein (2015), Rivkin, Hanushek and Kain (2005), Rothstein (2010), Chetty, Friedman and Rockoff (2014a), I assign a test-score value-added standard deviation of 0.15 to all teachers. One of Winters and Cowen's (2013) simulation's key insights is that the effects of deselection based on value-add score depend in large part on value-added variance. As my interest is not in the effect of deselection *per se*, I hold teachers' value-added variation constant at a mid-range of empirical estimates in order to explore the interaction of accountability and growth in the evaluation process.

Drawing on Jackson (2018), I assign a value-added estimate of teachers' ability to improve students' behavioral outcomes, defined as students' school suspensions for behavioral infractions, absences from schools and their 9^{th}/10^{th} grade GPA. Following Jackson, their value-added for this behavioral index is weakly correlated (r=0.15) with their value-add for test-score outcomes and has a standard deviation of 0.10.

Teachers' value-added contributions varies with their experience. I draw from Papay and Kraft (2015) to assign returns to experience. Teachers in their first year have test-score value-added measures 0.06 *SD* below teachers in their fourth year. Those in their second and third years perform 0.03 and 0.01 *SD* below fourth year teachers respectively. Beyond the fourth year, teachers improve by 0.003 standard deviations each year up to 25 years of experience at which point their value-added gains are capped at 0.066 SD above fourth year teachers.

Teachers are evaluated yearly. Their evaluation score depends on a subjective observation score that is correlated with their yearly observed value-added. Following Rothstein (2015), I set the standard deviation of the noise in annual teacher value-added scores at 0.183. This captures the cross-year variability in observed teacher value-added. Subjective observation scores are correlated with test-score value-added at r=0.3, an upper bound of Kraft, Papay and Chi (2018), Grissom and Loeb (2017), and Rockoff et al. (2012). Teachers final evaluation rating is calculated as 20 percent of their standardized observed test-score value-add (objective "piece-rate" evaluation) and 80 percent of their standardized observation ratings (subjective "supervisor judgment" evaluation), a ratio reflective of many states' actual policies (Steinberg & Donaldson, 2016). All of the previous assumptions persist across all scenarios I estimate. I then vary parameters across types of evaluation policy and optimistic and pessimistic scenarios.

I assume that teachers will improve their skills from evaluation across all evaluation frameworks and scenarios. In Panel A, I present plausible parameters for their learning in a Growth evaluation framework. In the baseline case, following Taylor and Tyler (2012) teachers improve by 0.11 *SD* in the first year they experience evaluation. The magnitude of these improvements are also consistent with Phipps's (2018)estimates of 0.06 and 0.03 *SD* improvements n reading and math. Taylor and Tyler are able to observe teachers immediately after the evaluation year and see no decline in their skills in the year following evaluation. I assume in the baseline case that more dosage of evaluation will lead to continued improvement over time. I specify an increasing exponential decay function in which teachers improve most over the first few years of exposure to evaluation and then experience gradually diminishing returns to evaluation. In the baseline case, I specify that their additional growth after the first year asymptotes at 20 evaluation years at an additional 0.1 *SD* above the gains in their first year of evaluation. In the optimistic scenario, I allow them to improve by an additional 0.2 SD above their first year gains. In the pessimistic scenario, teachers improve asymptotically towards a total gain of 0.11 standard deviations around their 20th year of being evaluated. I use these same parameters for the Divided Growth / Accountability evaluation framework (Panel C). Estimates are not sensitive to multiple other parameterizations that assume teacher skill improves in some concave function.

In Panel B (Growth and Accountability), I incorporate evidence from above that some teachers may not improve under a personnel management framework that has a high-degree of accountability and growth. Here, I explicitly contrast effects of identical evaluation policies where one scenario assumes that accountability and growth are complementary for all teachers, and another scenario that assumes they are substitutes for particular teachers. In Panel B1, I assume all teachers who receive an evaluation improve following the same baseline, optimistic and pessimistic parameters as in the Growth framework. In Panel B2, I assign half of all teachers who fall at or below the 80th percentile in their evaluation scores to not experience improvement through evaluation. There exists no empirical rationale for either the proportion of teachers who treat accountability and growth as substitutes or the evaluation percentile rank below which teachers experience accountability in their evaluation. I select these two values as reasonable, illustrative examples.

Each year, as a result of their annual appraisal I assign teachers a rank-ordered evaluation percentile. I use this approach rather than requiring consecutive poor evaluations following Winters and Cowen's (2013) finding that the latter evaluation scheme results in far fewer dismissals and student outcome gains due to year-over-year test score noise. Even in pure Growth scenarios (Panel A), I preserve supervisor ability to dismiss teachers for negligence, chronic absence or failure to meet basic professional expectations by dismissing teachers observed in the bottom (1st) percentile of the performance distribution. In the Growth and Accountability scenario (Panel B), teachers are dismissed if they are in the bottom 5 percentiles in their first two years. Following, Rothstein (2015) they are dismissed if they fall in the bottom 20 percentiles in their tenure year. After their third year of teaching, they still face dismissal if they are in the bottom 10 percentage

points of the distribution of evaluation scores. In the Divided evaluation framework, I use the same parameters for teachers in their first two years. I assume that it would be politically feasible to create clearly defined categories of untenured, early career teachers who would be subject to stringent evaluation. Given the high-stakes nature of tenure-year dismissals, I assume that it would be possible to clearly divide and designate teachers for dismissal in only the bottom 10 percentiles of the evaluation distribution. After the tenure year, I assume that teachers in the bottom 5 percentiles of evaluation scores could be defined as needing improvement, and that failure to improve out of the 5^{th} percentile would result in their dismissal. I assume that once teachers are dismissed they do not return to the teacher labor market, though in current policy frameworks teachers can return to teaching by moving to different districts or states.

In addition to teachers who leave the labor market due to dismissal, I assume various rates of attrition from the profession. I specify across all scenarios that all teachers leave teaching after 35 years of experience, which aligns with choices made by all but 2 percent of teachers in the National Teacher and Principal Survey (2017). I also specify an annual exit rate following a Gompertz function in which teachers' probability for attriting is around 18.2 percent following their first year, 12.6 percent after their second year, falls to 4 percent by their fifth year, and has an annual hazard exit rate below 1 percent by year 9. These values roughly match attrition rates in Winters and Cowen (2013) drawn from the 2004/5 Teacher Follow-Up Survey; however they assume greater attrition after the first years of teaching and distinguish between attrition rates at different years of experience beyond year five. This is important in my analysis as I assume that teachers continue to improve until year 25. Additionally, it reflects the reality that teachers with more than five years of experience do, in fact, leave the profession.

Following evidence summarized in Winters and Cowen (2013), I also assume differential quality for teachers who leave. To draw attention to the differences in the effects of attrition across the evaluation systems, I emphasize the component of differential attrition related to accountability pressures imposed by the evaluation framework. Therefore, I assume no differential attrition in the Growth scenario. In the Growth and Accountability (Panel B) scenario, I assume each year 25 percent of all teachers rated in the bottom 20 percent of the evaluation score distribution leave teaching. In practice, because all teachers in the bottom 10 percent of the evaluation distribution are dismissed, this would mean that one-quarter of teachers who are on the margin of being dismissed (between the 10^{th} and 20^{th} percentiles) would attrite. In the Divided scenario, I assume lower rates of attrition for teachers near the margin and a smaller margin around the dismissal threshold as a result of the clear division between accountability and growth purposes for evaluation. Thus, in the Divided evaluation framework, I assume that 1 in 8 teachers (12.5 percent) between the 5^{th} and 10^{th} evaluation score percentiles will attrite.

Finally, I incorporate the possibility that some of these evaluation policy changes may alter the composition of latent skills in the supply of new teachers. Here, I have little empirical evidence from which to specify parameters since these counterfactuals are nearly impossible to observe over

the long run in practice. Thus, I impose reasonable bounds and use the simulation results as informative on the range of these effects. I assume no labor market changes under the Growth evaluation framework. In the optimistic scenario under the combined Growth and Accountability evaluation framework, I specify that entrants into the teaching profession will gradually improve in quality, such that by the end of the 40-year simulation window they will on average perform 0.2 *SD* better than novice teachers at the start of the simulation. I assume similar potential positive labor market effects for the Optimistic Divided evaluation framework because the same mechanisms by which labor market quality improvements would be purported to operate (Winters & Cowen, 2013) would be at work in the Divided framework as well.

Given evidence from Rothstein (2015) and Kraft et al. (2019) on the potential teacher shortages resulting from high-stakes evaluation, I also specify a pessimistic scenario in which as a result of fewer prospective teaching candidates in the labor market pool, hiring committees must select candidates from lower in the latent skill distribution. This occurs both through greater demand due to more vacancies and the potential for risk-averse teaching candidates to withhold their labor supply. In the pessimistic Growth and Accountability scenario, I assume that the average quality of the starting pool will decline in exponential fashion to 0.2 *SD* worse than novice teachers at the start of the simulation. I assume that the potential negative effects on the labor market supply of teachers will be more modest in the Divided scenario as fewer teachers are dismissed, fewer are needed to fill vacancies, and risk averse teachers would observe less risk due to lower dismissal rates. Thus, in the Divided pessimistic scenario, I assume that the average quality of the starting pool of teachers will decline in exponential fashion to 0.1 *SD* worse than the novice teachers at the start of the simulation.

I begin the simulation with the initial skills of teachers defined as a mean zero, standard deviation of 0.15 test value-added score. I create a yearly observed value-add score and an associated evaluation score. I estimate the average true test- and behavioral-value-add skills of teachers. Then, I assign groups of teachers to be dismissed and to attrite following the rules above. For teachers who remain, I increase their true value-add score based on gaining experience and being evaluated based on the rules above. Their behavioral value-add scores increase in tandem with their test value-add scores, but remain weakly correlated. I then fill vacancies created through dismissal and attrition with novice teachers who have the previously defined characteristics such that the total number of teachers remains constant over years. I iterate the simulation over forty years.

Table A1. Values of Key Parameters

Parameter	Common Parameters
SD of teacher test value-added	0.15
r behavior VAM-test VAM	0.15
SD of behavior-index VAM	0.10
SD of noise in observed VAM	0.183
r observed VAM-eval. score	0.3
Experience effect on value-add productivity	-0.06 if $t=1$ -0.03 if $t=2$ -0.01 if $t=3$ 0.003(t) if $3<t<26$

		A. Growth		
Parameter	Function	Baseline	Optimistic	Pessimistic
Learning from evaluation	$VAM = \alpha(1 - e^{-\beta y^{\gamma}})$	+.11 if $y=1$ α=0.1, β=0.2, γ=1 if y>1	+.11 if $y=1$ α=0.2, β=0.2, γ=1 if y>1	α=0.11, β=0.2, γ=1
Dismissal criteria		pct_eval≤1	pct_eval≤1	pct_eval≤1
Attrition				
General	$P(Attrite\|Yrs\_exp) = 1 - \alpha e^{-be^{-c(yrs\_exp)}}$	α=1, b=0.3, c=0.4; $t>35$	α=1, b=0.3, c=0.4; $t>35$	α=1, b=0.3, c=0.4; $t>35$
Scenario specific		NA	NA	NA
Labor market supply changes		NA	NA	NA

		B1. Accountability & Growth – Complements		
		Baseline	Optimistic	Pessimistic
Learning from evaluation	$VAM = \alpha(1 - e^{-\beta y^{\gamma}})$	+.11 if $y=1$ α=0.1,	+.11 if $y=1$ α=0.2,	α=0.11, β=0.2, γ=1

		β=0.2, γ=1 if y>1	β=0.2, γ=1 if y>1	
Dismissal criteria		pct_eval≤5 if t<3 pct_eval≤20 if t=3 pct_eval≤10 if t>3	pct_eval≤5 if t<3 pct_eval≤20 if t=3 pct_eval≤10 if t>3	pct_eval≤5 if t<3 pct_eval≤20 if t=3 pct_eval≤10 if t>3
Attrition				
General	$P(Attrite\|Yrs\_exp) = 1 - \alpha e^{-be^{-c(yrs\_exp)}}$	α=1, b=0.3, c=0.4; *t*>35	α=1, b=0.3, c=0.4; *t*>35	α=1, b=0.3, c=0.4; *t*>35
Scenario specific		25% of pct_eval≤20	25% of pct_eval≤20	25% of pct_eval≤20
Labor market supply changes	$VAM = \alpha(1 - e^{-\beta y^{\gamma}})$	NA	α=0.2 β=1/100 γ=2	α=-0.2 β=1/100 γ=2

B2. Accountability & Growth – Substitutes

		Baseline	Optimistic	Pessimistic
Learning from evaluation	$VAM = \alpha(1 - e^{-\beta y^{\gamma}})$	P(+.11 if *y*=1 α=0.1, β=0.2, γ=1 if y>1 | Base) = 0.5 if pct_eval≤80	P(+.11 if *y*=1 α=0.2, β=0.2, γ=1 if y>1 | Optim) = 0.5 if pct_eval≤80	P(α=0.11, β=0.2, γ=1 | Pessim) = 0.5 if pct_eval≤80
Dismissal criteria		pct_eval≤5 if t<3 pct_eval≤20 if t=3 pct_eval≤10 if t>3	pct_eval≤5 if t<3 pct_eval≤20 if t=3 pct_eval≤10 if t>3	pct_eval≤5 if t<3 pct_eval≤20 if t=3 pct_eval≤10 if t>3
Attrition				
General	$P(Attrite\|Yrs\_exp) = 1 - \alpha e^{-be^{-c(yrs\_exp)}}$	α=1, b=0.3, c=0.4; *t*>35	α=1, b=0.3, c=0.4; *t*>35	α=1, b=0.3, c=0.4; *t*>35
Scenario specific		25% of pct_eval≤20	25% of pct_eval≤20	25% of pct_eval≤20

Labor market supply changes	$VAM = \alpha\left(1 - e^{-\beta y^{\gamma}}\right)$	NA	α=0.2 β=1/100 γ=2	α=-0.2 β=1/100 γ=2

C. Divided Accountability / Growth

		Baseline	Optimistic	Pessimistic
Learning from evaluation	$VAM = \alpha\left(1 - e^{-\beta y^{\gamma}}\right)$	+.11 if y=1 α=0.1, β=0.2, γ=1 if y>1	+.11 if y=1 α=0.2, β=0.2, γ=1 if y>1	α=0.11, β=0.2, γ=1
Dismissal criteria		pct_eval≤5 if t<3 pct_eval≤10 if t=3 pct_eval≤5 if t>3	pct_eval≤5 if t<3 pct_eval≤10 if t=3 pct_eval≤5 if t>3	pct_eval≤5 if t<3 pct_eval≤10 if t=3 pct_eval≤5 if t>3
Attrition				
General	$P(Attrite\|Yrs\_exp) = 1 - \alpha e^{-be^{-c(yrs\_exp)}}$	α=1, b=0.3, c=0.4; t>35	α=1, b=0.3, c=0.4; t>35	α=1, b=0.3, c=0.4; t>35
Scenario specific		12.5% of pct_eval≤10	12.5% of pct_eval≤10	12.5% of pct_eval≤10
Labor market supply changes	$VAM = \alpha\left(1 - e^{-\beta y^{\gamma}}\right)$	NA	α=0.2 β=1/100 γ=2	α=-0.1 β=1/100 γ=2

Appendix B. Additional Figures and Tables

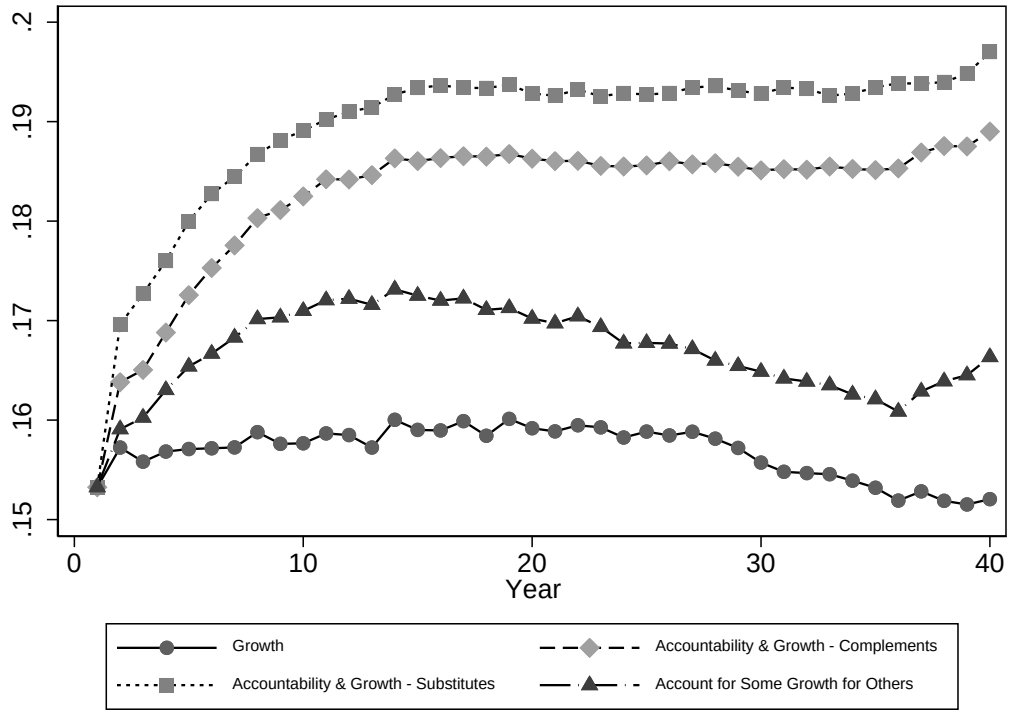

Figure B1. Standard deviations of test-score value-added, over 40 simulated years under baseline growth- and accountability-oriented evaluation policies

Table B1. Comparison of average teacher value-added on index of behavioral outcomes (suspensions, absences and GPA) in year bins under growth- and accountability-oriented evaluation policies

	Baseline	Optimistic	Pessimistic
A. Growth			
Years 2-5	0.013	0.015	0.004
Years 6-10	0.016	0.022	0.008
Years 11-15	0.017	0.025	0.008
Years 16-20	0.016	0.024	0.008
Years 21+	0.015	0.022	0.006
B1. Growth & Accountability - Complements			
Years 2-5	0.012	0.014	0.003
Years 6-10	0.015	0.022	0.006
Years 11-15	0.016	0.025	0.006
Years 16-20	0.016	0.027	0.005
Years 21+	0.017	0.027	0.006
B2. Growth & Accountability - Substitutes			
Years 2-5	0.008	0.011	0.002
Years 6-10	0.013	0.020	0.003
Years 11-15	0.015	0.025	0.003
Years 16-20	0.015	0.027	0.003
Years 21+	0.015	0.027	0.003
C. Growth / Accountability Divided			
Years 2-5	0.013	0.015	0.004
Years 6-10	0.017	0.023	0.008
Years 11-15	0.018	0.027	0.008
Years 16-20	0.017	0.028	0.008
Years 21+	0.018	0.028	0.008

Notes: average behavioral outcome value-added estimates derived from simulation described in Appendix A