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

David D. Liebowitz*
University of Oregon

July 2020

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 skill development. Proper teacher evaluation design has been understood as successfully weighting the accountability and professional growth dimensions of policy and practice. I develop a model of teacher effectiveness that incorporates improvement from evaluation and detail conditions which determine the effectiveness of teacher evaluation for growth and accountability at improving student outcomes. Drawing on empirical evidence from the personnel economics, economics of education and measurement literatures, I simulate the long-term effects of a set of teacher evaluation policies. I find that those that treat evaluation for accountability and evaluation for growth as substitutes outperform policies that treat them as complements. I conclude that optimal teacher evaluation policies would impose accountability on teachers performing below a defined level and above which teachers would be subject to no accountability pressure but would receive intensive instructional supports.

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

***JEL* codes**: I21, I28, J24, J41, J45

*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. This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors. Please direct correspondence to David Liebowitz at daviddl@uoregon.edu, Department of Educational Methodology, Policy and Leadership, 5267, University of Oregon, Eugene, OR, 97403.

1 Introduction

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 design might improve the average quality of instruction are through (a) incentives to motivate teachers, (b) tools for retaining high-performing teachers and deselecting low-performing ones, (c) shifts to the labor market pool of prospective teachers, and (d) feedback mechanisms to improve the skills of current teachers. Most consequential teacher evaluation policies attempt to achieve a blend of accountability and skill developmental goals. While these multiple goals are firmly part of the design of modern teacher evaluation policies, researchers and policy makers have reflected surprisingly little on the interactions between these mechanisms.

In this paper, I argue that a clear explication of the tenets underlying teacher evaluation policies is critical to estimating its likelihood of success. I develop a stylized model of teacher effectiveness that includes improvement from evaluation. My model's central contribution is to highlight that improvement in teaching effectiveness from evaluation depends not only on the independent effects of accountability pressures and skill development supports, but also on their interaction. From this model, I identify several conditions that determine the effectiveness of modern teacher evaluation policy design.

I then develop a set of simple simulations to assess the effect of policies that emphasize the growth or accountability aims of teacher evaluation. To do so, I draw on the current causal evidence base to select parameters related to my model's production function. I extend prior simulation work from Winters and Cowen (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. As Rothstein notes, models that incorporate selection and deselection effects into and out of the profession are ill-suited to evaluation by experiment (either random assignment or natural) because their effects take a decade or more to manifest. This fact motivates my simulation.

Through this simulation, I attempt to model the effects of teacher evaluation policies that (*i*) provide only developmental supports; (*ii*) policies that combine accountability pressures with skill development supports; and (*iii*) 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.

My simulation's central contribution is that, given reasonable parameter bounds derived from the causal literature, the potential benefits of an evaluation system that applies accountability and growth supports to all teachers are dwarfed by its potential risks. My simulations synthesize Winters and Cowen's (2013) conclusion that the hypothesized benefits of dismissing less effective teachers are muted by natural attrition with Rothstein's (2015) insights that a skill-constrained supply of prospective teachers influences optimal dismissal rates. Even if evaluation for growth and accountability have cross-productivity effects, the upper-bounds on these effects are relatively low. The most important determinants for improving average teacher effectiveness are the rate of skill development from ongoing supervisor feedback over several years and the labor supply effects of evaluation policies, both of which are empirically indeterminate. Thus, evaluation policies that jointly maximize skill development rates of current teachers and starting skill endowments of future teachers hold the most promise. As a result, I conclude that an evaluation policy that treats evaluation for growth and accountability as substitutes rather than complements is more likely, across various plausible scenarios, to produce improved student outcomes.

2 Teacher Evaluation Policy and Practice

Present-day teacher evaluation policies in the United States generally attempt to hold teachers accountable to standards for quality instruction and to create a process through which teachers can improve their skills.[1] As Murnane and Cohen (1986) document, the traditional rationale for accountability- and incentive-based appraisal, 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 ("new" piece-rate compensation) or the subjective assessment of a supervisor ("old" subjective supervisor judgment).

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 for the worker to self-reflect on ways to improve her skills. Policy makers generally attempt to maximize both the accountability and skill development aims in the design of teacher evaluation policies.

In response to incentives from the Obama administration's Race to the Top program in 2009, 44 state legislatures across the United States implemented reforms to their teacher evaluation systems (Kraft et al., 2019). These reforms both imposed greater accountability and formalized professional growth processes via teacher evaluation. In almost all cases the new policies entailed adopting a common rubric for evaluating teachers' performance with multiple rating categories, representing a shift away from the traditional Satisfactory/Unsatisfactory distinctions (Ross

[1] I focus on teacher evaluation in the United States and draw empirical evidence from studies in this context. The lessons of the theoretical model are broadly applicable. The empirical evidence that informs the simulation and the simulation itself are most relevant in national contexts where teacher appraisal policy has the joint-aim of applying accountability and providing instructional supports. These insights have less relevance to national contexts where teacher evaluation is entirely incentive- and sanction-based (e.g., Muralidharan and Sundararaman, 2011), where evaluation is conducted primarily to fulfill administrative requirements (e.g., OECD, 2019), or to teacher coaching schemes with no connection to the appraisal system (e.g., Murphy, Weinhardt & Wyness, 2020).

and Walsh, 2019). In most cases, state reforms to teacher evaluation required that classroom observation of teaching practice be a part of a teacher's final rating and established a minimum frequency of these observations. In addition, many states required some teachers 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 (Steinberg and Donaldson, 2016; Winters and Cowen, 2013). Teachers who fail to earn an evaluation rating at the Proficient or higher threshold are subject to dismissal.[2] Additionally, most states require educators to receive professional development or coaching in response to their evaluation results (Steinberg and Donaldson, 2016).

Despite the prevalence of these joint-aim evaluation policies, there exist important theoretical and empirical reasons why accomplishing these two goals in practice may be challenging. High-stakes employee appraisal may generate "goal distortion" (the reallocation of effort from unmeasured to measured tasks) or the gaming of outcome measures (Holmstrom and Milgrom, 1991). Further, as Dixit (2002) notes, the agents of evaluation policies (teachers) also must respond to multiple principals (parents, children, administrators, policy makers, etc.) which complicates teachers' response to the action of any one principal. Gneezy, Meier and Rey-Biel's (2011) review of the evidence on teacher incentives finds them valuable in improving performance on tasks that require only the application of additional effort (e.g., Muralidharan and Sundararaman, 2011; Glewwe et al., 2010); however, they conclude that tasks that require development of skill do not improve in response to external motivation. 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.

The personnel economics literatures have also devoted substantial investigation into 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, there is suggestive evidence that accountability and skill development are competing aims within an employee management system as a result of goal distortion, the psychological effects of high stakes, resistance to feedback and reduced motivation.

Given the tensions between the accountability and skill development aspects of teacher evalua-

[2] Commentators have debated the extent to which present-day U.S. 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 and Walsh, 2019), and most educators continued to receive positive appraisal ratings (Kraft and Gilmour, 2017). However, it is important to note that poor appraisals risk significant consequences; in over three-fifths of states, teachers who are not rated Proficient should, by policy, be dismissed (Steinberg and Donaldson, 2016).

tion, policy makers and commentators have generally characterized evaluation policies as those which appropriately balance these two components (e.g., Darling-Hammond et al., 1983; Donaldson and Papay, 2015; Popham, 1988). Different policy choices might emphasize accountability and incentives or skill development. 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 the literature considers 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. In the next section, I do just that.

3 A Model of Teacher Effectiveness

My contribution begins with a simple model to describe an individual teacher's effectiveness:

$$Y_{it} = Y_{i0} + \mathrm{r}(exp) + (\pi G_{it}^{\rho} + (1-\pi)A_{it}^{\rho})^{1/\rho} + \zeta_{it} + \epsilon_{it} \tag{1}$$

Y_{it} represents teacher i's noisily estimated (ϵ_{it}) contribution to student gains on a particular outcome in time t, which may or may not correlate with a teacher's effects on other desirable outcomes. Y_{it} depends on the teacher's initial skill endowment (Y_{i0}), and a flexible concave function of returns to experience, $\mathrm{r}(exp)$, which might include informal on-the-job learning and formal professional development activities.

Improvements from evaluation are a product of coaching and skill development (G_{it}) and accountability pressures (A_{it}). Skills may be developed from direct feedback provided during post-observation sessions, self-reflection the part of an individual teacher catalyzed by the observation and feedback process or from professional development assigned as an outcome of the a teacher's formal evaluation. Accountability pressure may produce changes in effort and complement skill development. Accountability-induced improvements may result from the threat of sanction for poor performance, the desire to earn material or prestige-based rewards for improving performance, or from internalized motivation to earn a high (or improved) rating on a measure of performance.

A given evaluation policy in this model emphasizes growth (π) and accountability ($1-\pi$) on a scale normalized between 0 and 1. This is, of course, a heuristic simplification as it is not possible to characterize the precise emphasis of a given evaluation policy. Nevertheless rough approximates can provide bounds on the effectiveness of a particular policy choice.

ρ is a substitution parameter defined by the elasticity of substitution (σ) between skill development and accountability where $\rho = (\sigma - 1)/\sigma$. Readers may recognize the similarities between this model of improvement from evaluation and Arrow et al.'s (1961) specification of the constant elasticity of substitution (CES) function. Important insights include that the CES model nests a Cobb-Douglas production function ($\sigma = 1$), a linear von Neumann function where both factors are perfect substitutes ($\sigma \to \infty$), and a Leontief function where they are perfect complements ($\sigma = 0$). To preview later concerns, the relative effectiveness of evaluation that prioritizes skill development compared to accountability pressure, combined with the elasticity of substitution between the two dictates whether evaluation policy should simultaneously provide growth supports and impose accountability pressures on all teachers. If both are equally effective at improving skill development, the choice of π has no effect on student learning. If growth and accountability are perfect substitutes, the choice of π depends entirely on their independent effectiveness.

A central concern I seek to understand in this paper is the substitutability of evaluation for growth and accountability and their cross-productivity effects. An elasticity of substitution less (greater) than one implies growth and accountability are gross complements (substitutes) and influences the productivity of different choices of π.[3] Additionally if increasing one of the factors reduces (increases) the marginal benefit of the other, formally if their cross-partial derivative is less (greater) than zero $\left(\frac{\partial^2 Y}{\partial A \partial G} = \frac{\partial^2 Y}{\partial G \partial A} < 0\right)$, it implies that π cannot be chosen based on the independent effectiveness of evaluation for growth and accountability.[4]

In a particular year, mean teacher effectiveness is then given by:

$$\bar{Y}_t = \frac{\Sigma_{i=1}^{n} \left[Y_{i,t-1}^{return} + \mathrm{r}(exp)^{return} + \left(\pi G_{i,t-1}^{\rho,return} + (1-\pi) A_{i,t-1}^{\rho,return} \right)^{1/\rho} \right]}{n} + \bar{Y}_{t(0)}^{\text{entry}} \tag{2}$$

The mean of returning teachers' effectiveness in the prior year ($Y_{i,t-1}^{return}$), combined with annual improvements from experience and evaluation, and the starting effectiveness of entrants to the profession produces the average teacher effectiveness in any given year. For a given prospective teaching candidate, entry into the profession depends on whether the expected utility of entry into the teaching profession (tch) is higher than entry into all other possible career paths, (d): $\mathrm{E}[\mathrm{u}(tch)] - \mathrm{E}[\mathrm{u}(d)] > 0$. The expected utility for a prospective teacher is a function of their anticipated wages ($\mathrm{w}(t)$), their projected skill trajectory ($\mathrm{f}(Y_{it})$), the interaction of their projected skill trajectory with the incentive and consequence structure of teacher human resource policy (δ), and idiosyncratic initial preferences for teaching (ε_{i0}):

$$\mathrm{E}[\mathrm{u}(tch)] = \mathrm{E}\left[\mathrm{w}(tch) + \mathrm{f}\left(Y_{it}\right) + \mathrm{g}\left(Y_{it} \times \delta\right)\right] + \varepsilon_{i0} \tag{3}$$

δ can be understood as a policy response to a rank-ordered annual measure of teacher effec-

[3] I borrow the concept of "gross" complements (and substitutes) from Acemoglu (2002). In this context, I mean it to indicate that the marginal return to A_{it} increases in response to an increase in the productivity of G_{it}, holding the productivity of A_{it} constant, if and only if $\sigma < 1$, and vice versa. I assume that $\sigma \geq 0$, as a negative value would imply a concave function or that less skill development/effort would result in increased performance. The γ from the numerator of the exponent surrounding the standard CES function is unnecessary in my formulation as there is no direct analog for returns to scale in the teacher evaluation production function.

[4] For completeness, ζ_{it} represents a vector of contextual factors (e.g., class size, starting student achievement, community violence, etc.) and may include interactions with other terms. Teacher effectiveness may be improved by altering these contexts, but I hold them fixed throughout the analysis to focus on the effects of teacher evaluation itself.

tiveness based on evaluation scores. Below a given threshold under a particular δ teachers are subject to dismissal, above a threshold they may earn pecuniary or other rewards.

Teachers' exits (and the associated $\bar{Y}_t^{exit}$) depend on the opposite of expectation conditions at entry: $\mathrm{E}[\mathrm{u}(tch)] - \mathrm{E}[\mathrm{u}(d)] < 0$ plus whether a teacher is formally dismissed, a function of δ. The skill level of entrants (Y_{i0}^{entry}) depends, in turn, on $\mathrm{E}[\mathrm{u}(tch)]$. The expected utility of teaching declines for those who project future evaluation-based sanctions and increases for those projecting rewards. Through this mechanism, my model includes skill-biased attrition and retention.

From the above, we can derive in a straightforward fashion the conditions that determine whether, and the extent to which, teacher evaluation with a focus on accountability and the skill development process will improve student learning outcomes.

First, the larger the noise component of teachers' value-added metric (ϵ_{it} in Equation 1), the lower the reliability in annual evaluation ratings. Relatedly, the weaker the predictive validity of the test-score measure for other teacher-influenced student outcomes, the less informative it is as a measure of overall teaching quality. As the stakes of a given measure increases, the more likely the measure is to be corrupted (e.g., Figlio, 2006; Neal and Schanzenbach, 2010). As a result, to the extent that $\bar{Y}_t$ depends on teacher dismissal decisions made under δ, lower values of the reliability of Y_{it} and of its covariance with other important measures will produce lower values of $\bar{Y}_{t(VAM)}$ and $\bar{Y}_{t(1 \cdots N)}$, respectively.

Second, either a positive value of A_{it} or G_{it} in Equation 1 is a necessary condition for an individual teacher's effectiveness to improve as a result of evaluation. This is not, however, a sufficient condition as this depends on the value of ρ.

Third, the interactions between evaluation practices that promote both accountability and developmental goals determine the total factor productivity of teacher evaluation at the individual level. Formally, this implies that if the cross-partial derivative of the production function is less than zero, or said differently, if the elasticity of substitution (σ) between skill development and accountability pressures is greater than one, then total factor productivity will be decreased as a result of an increased focus on one of the two evaluation factors.

Fourth, in order for average teacher effectiveness to improve, the necessary conditions are that either A_{it} or G_{it} is positive, or that evaluation improves the quality of teachers in the labor market ($\bar{Y}_{t(0)}^{entry} > \bar{Y}_t^{exit}$). In other words, prospective teachers possess (or can quickly acquire) teaching skills equal or superior to departing teachers. Y_{i0}^{entry} depends on $\mathrm{E}[\mathrm{u}(t)]$, total vacancies, and employers' ability to select for Y_{i0}^{entry}. $\bar{Y}_{t(0)}^{entry}$ in will decline if insufficient teachers are available to fill vacancies and $\mathrm{E}[\mathrm{u}(t)]$ in Equation 3 will decline if $\mathrm{E}[\mathrm{g}(Y_{it} \times \delta)]$ is negative or the weight of this term increases and prospective high-skill teachers are risk averse.

Simply stated, evaluation must increase skills in incumbent teachers through either accountability pressures or professional growth, or evaluation must change the skill composition of the teacher workforce through differential attrition, deselection and changes to the skills of incoming cohorts of teachers. As long as one of the three preceding conditions is met, sufficiency is an empirical concern that requires a comparison of the reinforcing or countervailing effects of the other conditions. For instance, if accountability pressures improve teacher practice, unreliable evaluation ratings might still capture enough signal about teaching quality that their use would improve the overall distribution of teacher effectiveness. In this example, accountability pressures could improve average teacher effectiveness, even in the presence of teacher shortages or negative cross-productivities to the skill development dimensions of evaluation, as long as these secondary effects were smaller in magnitude than the main effects of individual teacher improvements from accountability.

An assessment of present-day, joint-aim teacher evaluation policies requires not only evaluating whether they satisfy sufficiency conditions for improving average teacher effectiveness, but also how they compare to policy counterfactuals. If alternative policy designs more readily satisfy some of the conditions or do not require them, this would be valuable for policy makers to know as well.

4 Simulation Set-Up and Parameters

I develop a simulated 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's (2013) alternative deselection policy simulations and Rothstein's (2015) dynamic discrete choice model to construct a simulated teacher labor market and estimate the effects of several evaluation policies. I select plausible parameters on the preceding from the most current causal literature base. Where limited evidence exists, I use reasonable bounds across 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 increase effort and 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.

I first present simulation results that isolate the effects of the above conditions for effective evaluation independently and then combine them to assess the impacts of three stylized, multiple-measure teacher evaluation policies: a policy model focusing exclusively on providing growth-oriented supports, a model attempting to combine growth- and accountability-oriented strategies similar to modern evaluation policy designs, and a model dividing these aims across groups of teachers.

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 (2017), with experience capped at 35 years.

I use the parameters from Table 1 to assign teachers initial values for their latent ability in improving students' test score outcomes (Y_{i0}). Following Rothstein (2015), Rivkin, Hanushek and Kain (2005), Rothstein (2010), and Chetty, Friedman and Rockoff (2014), 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.

I also 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 (2018), and roughly similar to Gershenson (2016) and Kraft (2019), 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.

I create a yearly observed value-add score and an associated evaluation score. Teachers' annual 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. Following Sass (2008) and 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 (2020), 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 and 80 percent of their standardized observation ratings, a ratio roughly reflective of many states' actual policies (Steinberg and Donaldson, 2016).

Teachers' value-added contributions improves with experience, r(exp). I draw from Papay and Kraft (2015) to assign returns to experience. Teachers in their first year have test-score value-added measures 0.07 *SD* below teachers in their fourth year. Those in their second and third years perform 0.04 and 0.02 *SD* below fourth year teachers, respectively. Beyond the fourth year, teachers improve by 0.003 standard deviations each year up to 15 years of experience at which point their value-added gains are capped at 0.033 *SD* above fourth year teachers.[5]

In my model, teachers' value-added measures improve as a result of being evaluated. Drawing on Taylor and Tyler (2012) and Pope (2019), I specify in baseline scenarios that teachers improve by 0.11 *SD* following the first year they experience evaluation as a result of either instructional supports or accountability, respectively.[6] Taylor and Tyler observe teachers in the year after

[5] 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 natural attrition of teachers with 20+ years of experience.

[6] Other estimates of the effect of teacher evaluation are broadly consistent. 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 in reading (0.10-0.13 *SD*), with imprecisely estimated positive coefficients in math. Pope (2019) detects an increase in teacher effectiveness of 0.11-0.14 *SD* for teachers in the bottom quintile of the distribution from a public release of teacher performance data. Phipps (2019) finds that the potential for an accountability-based observation increases student performance by between 0.03-0.07 *SD*. Similarly, he finds that teachers improve from feedback under no accountability threat by 0.03-0.06 *SD*. Burgess, Rawal and Taylor (2019) examine a peer observation scheme in England which they describe as "peer evaluation." They find that teachers receiving feedback improved their contributions to student test-score learning

their initial evaluation and see no decline in their skills. I assume that more dosage of evaluation will lead to continued improvement over time, but with 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. I assume that more dosage of evaluation will lead to continued improvement over time, but with gradually diminishing returns to evaluation following the function in Table 1. With a given asymptote and a concave function specified, the exact functional form of diminishing returns to evaluation does not affect any of the substantive interpretations that I draw.

Teachers depart the profession as a result of either voluntary exits or dismissal. I specify across 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 defined in Table 1 in which teachers' probability for attriting is around 16.5 percent after their first year, 11.4 percent after their second year, 7.8 percent after their third year, 5.3 percent after their fourth year and then a cumulative 8.9 percent between years 5 and 15. I also specify that once teachers reach 20 years of experience, a sufficient number will attrite (with no skill bias) such that cohorts of 20+ years of experience will not exceed 1.5 percent of the total teaching workforce.[7]

In addition to longevity-based attrition, teachers attrite in differential ways based on their skill. The logic model underlying skill-biased attrition is that teachers choose to leave as a result of updating their beliefs about their fit with the profession from on-the-job experiences, from evaluator feedback, or from their expectations regarding future evaluation outcomes. I assume all three forces are at play. I draw evidence from Winters and Cowen (2013) and Adnot et al. (2017) to specify rates of differential skill-biased attrition. Adnot et al. identify that attriters have value-added scores 0.05-0.08 *SD* below their non-attriting peers with the same experience. These effects are similar substantively to Dee and Wyckoff (2015), Loeb, Miller and Wyckoff (2015), and Cullen, Koedel and Parsons (2019); though none of these studies estimate the global effect of teacher evaluation on student learning outcomes.

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 as these peer observations occur outside the formal evaluation process. The magnitude of these peer-coaching effects on student test-score gains are similar to those Papay, Taylor, Tyler and Laski (2020) find in Tennessee (0.12 *SD*). Together, I take this body of work to imply that both the accountability and coaching dimensions of evaluation produce one-year gains in teacher effectiveness of between 0.03 and 0.14 *SD* in the United States and similar national policy contexts. There is also a large adjacent body of literature on the effects of incentive pay on teacher performance (e.g., Goodman and Turner, 2013; Sojourner et al., 2014; Speroni et al., 2019). I do not fully explore the applicability of differential teacher compensation insights to broader evaluation strategies. However, the estimated magnitude of the effects of merit pay in similar national policy contexts is generally consistent with the results above.

[7] These values roughly match attrition rates in Winters and Cowen (2013) drawn from the 2004/5 Teacher Follow-Up Survey; however, my approach allows me to 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 15. More importantly, it reflects the reality that teachers with more than five years of experience do, in fact, leave the classroom; and therefore, more closely mirrors the actual distribution of teaching experience in the NTPS Survey. In fact, failure to model mid-career attrition leads simulations to overstate longer-term outcomes as the composition of the teaching force accrues greater experience over time. Even in cases where there are no benefits to experience after the initial five years of the teaching career, if simulations ignore mid-career attrition they will under-represent the proportion of these early career teachers in later iterations of the simulation.

Minimal empirical evidence exists on the medium- to long-term effects of evaluation policies on labor supply; thus the reliance on my simulation method. Prospective teachers may decide to enter (or not) the teaching profession based on whether a more rigorous evaluation system is attractive to them (Kraft et al., 2019). Additionally, increases in dismissals may result in greater supply needs, which might result in teacher shortages or the need to select from lower in the skill distribution (Rothstein, 2015). I draw on Rothstein's simulation evidence about the potential for teaching shortages under high-threshold dismissal policies and on Kraft et al.'s assessment of the mid-term effects of higher-stakes teacher evaluation on teacher licensure candidates and reporting of staffing difficulties to circumscribe the bounds of labor market effects. I classify a change in average starting effectiveness on the order of 0.05 *SD* as small and a change on the order of 0.15 *SD* as large. In all cases I assume that such changes in the skill levels of entrants will take several years to fully manifest following the function outlined in Table 1. As with improvements from evaluation over time, the exact functional form matters less than the assumption that labor market changes manifest in an exponential fashion, improving (worsening) slowly at first and then more rapidly in later years.

Each year, as a result of their annual appraisal I assign teachers a rank-ordered evaluation percentile. Depending on the particulars of a given simulation, I assign teachers below a given multiple-measure evaluation percentile to be dismissed. 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.

For teachers who remain, I increase their true value-add score based on gaining experience and being evaluated based on the rules of a particular scenario. 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.

5 Simulation Results

5.1 Condition-specific results

In Appendix Table A1, I present the parameters used to estimate the effects of various conditions of a growth- and accountability-based evaluation system.

The reliability of teachers' annual value-added contributions influences the effects of a value-added-only evaluation (and dismissal) policy. In Panel A of Figure 1, I begin by replicating Rothstein's (2015, fig. 6(A)) main result. Like Rothstein, I find that dismissing teachers in the bottom 20 percent of the value-added distribution in their tenure year improves mean true value-added by 0.02-0.03 *SD* when the standard deviation of the noise in annual value-added estimates is 0.183. As expected a near-doubling of the standard deviation decreases the benefits of this policy by half, whereas reducing annual noise to 0.05 *SD* increases the benefits of such a dismissal policy on student outcomes to 0.06 *SD*. Importantly, however, incorporating noisily estimated value-added scores into a multiple-measure evaluation system eliminates all benefits to dismissing the bottom 20 percent of teachers in their third year. The test-score measurement errors combine with the weak correlation between observations and test scores and the increasingly inexperienced teaching corps to produce no changes in student outcomes.

I do not vary the reliability of test-score value-added in simulations of complete evaluation systems, but it is notable that a multiple-measures evaluation system that relies solely on dismissing the lowest performers is unlikely to improve student test-score outcomes.

When teachers improve their effectiveness as a result of evaluation, either through skill development or increased effort, it substantially increases average value-added measures, independent of any dismissal or labor market effects. In Panel B of Figure 1, I specify a uni-dimensional effect of evaluation that combines the skill development and accountability factors of evaluation. I calibrate my models with a first simulation in which there is no improvement from evaluation, no skill-bias in attrition or entrance, no dismissals and teachers exit and enter the profession following current patterns. As expected, there is no change in mean value-added. Following Taylor and Tyler (2012) and Pope (2019), I then specify a scenario in which teachers improve by 0.11 *SD* in the first year they experience 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 third simulation, I allow them to improve by an additional 0.2 *SD* above their first year gains. In the last, teachers improve asymptotically towards a total gain of 0.11 standard deviations in their 20^{th} year of being evaluated. As is evident, the amount of teacher improvement from evaluation is critical in understanding the effect of any evaluation policy, and plausible differences in assumptions produce average changes in future teacher effectiveness ranging from 0.08 to 0.25 *SD*.

Skill-biased attrition has a relatively small effect on overall effectiveness due to high natural rates of attrition. In Panel C of Figure 1, I draw on evidence from Winters and Cowen (2013) and Adnot et al. (2017) to specify rates of differential skill-biased attrition. In this panel, I do not specify teachers actually be dismissed to avoid masking the effects of differential attrition with those of a particular dismissal policy. I assume standard returns to experience, no improvement from evaluation, and no changes to entrants in the teacher labor markets. In the first case, I specify that teachers in the bottom 10 percent of the evaluation score distribution have a 1 in 6 probability of attriting on top of the general attrition rate defined above. As a result, attriters in this scenario have value-added scores 0.05 SD below their non-attriting peers with the same experience level—equivalent to the lower bound of estimates from Adnot et al. In the second scenario, teachers in the bottom 10 percent have a 1 in 2 chance of attriting, which results in attriters having value-added scores 0.08 *SD* below their non-attriting cohort peers—an upper bound of Adnot et al. These result in value-added gains over time of 0.02 and 0.05 *SD*, respectively.

Reducing early-career attrition alone is not sufficient to increase teacher effectiveness. When I dramatically reduce the rate of early-career exits (years 5 and earlier) by two-thirds, it has no effect on student outcomes. This is because there is no skill bias to the attrition in this scenario, and I preserve mid-career attrition in years 20+ so that the overall experience profile of teachers remains the same. Increasing early-career teacher retention must be accompanied by skill-biased attrition and a shift to the overall experience profile to accomplish improved outcomes.

Changes in the starting skill endowments of entering teachers have a larger overall effect than higher dismissal rates for poor-performing teachers. High natural rates of attrition mean that higher-skilled entrants are unlikely to be constrained from entering the profession even absent any evaluation-based dismissals. In Panels D and E of Figure 1, I simulate the effect of changes in skills of new entrants to the teaching profession. In Panel D, I specify that the skills of new teachers will ultimately increase or decrease by small (± 0.05 *SD*) or large amounts (± 0.15 *SD*). I assume that such changes in the skill levels of entrants will take several years to fully manifest following a growth function articulated in Table 1. As Panel D suggests, holding all else constant and only altering the prospective skills of entrants to the profession produces substantial differences in teaching effectiveness.

The differences in overall teaching effectiveness are more modest when altering the dismissal threshold (Panel E) than when changing the skill distribution of new entrants (Panel D). In Panel E, no teachers improve from evaluation and there are standard rates of return to experience. I assume across these scenarios that there will be modest skill-biased attrition (-0.05 *SD*). I then vary the rate of dismissal in each evaluation policy and the effects this will have on the skills of new entrants. First, I assume that the bottom 5 percent of teachers in the evaluation score distribution are dismissed and this results in small improvements (+0.05 *SD*) in the skill profile of new entrants. In the following simulation, I assume this will result in small declines (-0.05 *SD*) in skills of new entrants. Next, I specify that the bottom 20 percent of teachers will be dismissed with small improvements in new entrants. The final scenario assumes that when the bottom 20 percent of teachers are dismissed this will result in small declines in the skills of new entrants. As Panel E indicates, the difference between policies that dismiss only 5 percent as opposed to 20 percent of all teachers after their third year are negligible.

Finally, I vary the substitutability of evaluation for developmental and accountability purposes and find that the bounds of these effects are fairly small. If growth and accountability are equally effective ($G_{it} = A_{it}$) or are perfect substitutes ($\sigma \to \infty$), then the choice of π is either irrelevant or can be altered without any loss in productivity. As such, I focus on instances in which $G_{it} \neq A_{it}$ and where $\sigma \neq \infty$. Further, I fix attrition rates, learning from experience and the skill of labor supply. No teachers are dismissed. The goal of this exercise is to observe how altering the balance of an evaluation policy towards the more effective evaluation function produces different improvements in teacher value-added depending on whether the two production function factors (growth and accountability) are complements or substitutes. I fix the more effective evaluation mechanism (in this set-up it could be either growth or accountability) as improving teacher effectiveness by 0.13 *SD*. This is in line with the upper bounds of the Pope (2019) accountability-based improvements and the Papay et al. (2020) growth-based improvements. I set the less effective evaluation mechanism at 0.03 *SD*, in line with the lower bounds of Phipps (2018). In line with the baseline assumptions from Panel B, I assume teachers will continue to experience another 0.13 (or 0.03) *SD* improvement over the next 20 years. I start by stipulating that the evaluation scheme favors the less successful mechanism to improve teacher effectiveness and assign a value of 0.1 to π. In the first scenario, I assume that growth and accountability are gross substitutes ($\sigma = 1.4$), and in the second that they are gross complements ($\sigma = 0.6$).[8] Then, I shift the weight each evaluation framework assigns to the more successful mechanism to improve teacher effectiveness by assigning a value of 0.9 to π. As expected, if teacher evaluation for growth and accountability are substitutes, the shift towards the more successful factor increases the total factor productivity by 0.01-0.02 *SD*s above the case in which they are complements.

5.2 Full evaluation policy results

Having explored how each of the conditions independently contribute to teaching effectiveness, I now assess the global effects of full evaluation policies. I present results from a series of simulations that combine the components of an evaluation system outlined above into three stylized comprehensive evaluation systems: a policy model focusing exclusively on providing growth-oriented supports (Growth-only), a model attempting to combine growth- and accountability-oriented strategies similar to modern evaluation policy designs (Combined Growth and Accountability), and a model dividing these aims across groups of teachers (Divided Growth and Accountability). The distinguishing characteristics are in the emphasis each scheme places on providing coaching and support to teachers, imposing accountability on teachers, integrating the two or applying these components separately to different groups of teachers.

I choose to highlight a Growth-only evaluation system as it is most similar in its teacher dismissal rates to historical and current-day dismissal practices. However, even large-scale, well-implemented teacher evaluation systems have failed to provide meaningful instructional supports that improve teaching practice (Garet et al., 2017; Stecher et al., 2018). Thus, there remains an untested opportunity to develop a low-accountability, high-support evaluation system. I focus on Combined Accountability and Growth policies as these are typically what reformers have in mind when they propose imposing additional accountability on teachers' practice.[9]

Given the potential cross-productivity and substitution effects between teacher evaluation for growth and accountability, an alternative system in which evaluation serves as an accountability floor for some and a developmental process for most, with clear distinctions between the two populations of teachers, may resolve some of these tensions. I explore such an evaluation scheme in these simulations under the heading of a Divided Accountability and Growth approach. In such an evaluation system, the large majority of teachers would be subject to an evaluation scheme directed exclusively towards professional growth. This portion of the evaluation scheme would offer targeted supports and opportunities for mentorship depending on teacher appraisals. Formal ratings of teachers in this category need not occur and teachers might instead receive a narrative review of their practice. A much smaller group of teachers, falling below a bright-line

[8]The evidence base does not allow me to set the values of σ empirically. Nevertheless, there is a lower-bound to the complementarity of growth and accountability; CES set-ups require a value of σ greater than 0. I use these two values to provide rough approximations of the bounds of the effects of complementarity and substitutability on evaluation frameworks.

[9]I do not focus on an evaluation policy that relies on rank-ordered dismissal or incentives alone as these are typically outside the scope of United States policy debates.

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 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 would only be critical for teachers performing below or near the accountability floor. Given strained administrative capacity to conduct rigorous evaluations across the teacher performance distribution, greater attention could be allocated around the accountability margin. Multiple measure systems that incorporate student learning outcomes, observations, surveys and other measures could concentrate their efforts to achieve validity and reliability at the threshold point. Accountability pressures would matter only for teachers near or below the floor—those whom 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. 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. The clear dividing line might return a sense of stability to risk-averse teachers and avoid labor supply challenges.

I present parameters for these scenarios in Appendix Table A2. In the Growth-only scheme, 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 (1^{st}) percentile of the performance distribution. Beyond these 1^{st} percentile teachers, I conceive it would be possible to create an evaluation system with no threat of sanction; perhaps even avoiding performance rating labels altogether in favor of narrative summaries of teachers' practice.

In the Combined Growth and Accountability scenario, teachers are dismissed if they are in the bottom 5 percentiles in their first two years and in the bottom 10 percentiles in their tenure year. After their third year of teaching, they still face dismissal if they are in the bottom 5 percentage points of the distribution of evaluation scores.

In the Divided evaluation framework, I use the same thresholds for dismissal. I assume that it would be politically feasible to create clearly defined categories of untenured, early career teachers within 5 percentiles of these thresholds (bottom 10 percent in Years 1 and 2 and bottom 15 percent in Year 3) who would be placed on notice of their risk for dismissal, subject to evaluation and assigned summative rating scores. After the tenure year, I assume that teachers in the bottom 10 percentiles of evaluation scores could be defined as needing improvement, and that failure to improve raise performance higher than the 5^{th} percentile would result in their dismissal. For teachers above these bandwidths, I conceive it would be possible to implement a growth- and support-focused evaluation system. I assume that once teachers are dismissed they do not return to the teacher labor market.

In all Baseline scenarios, I assume that the effects of evaluation are to improve effectiveness by 0.11 *SD* in the first year in which teachers are evaluated and then by up to another 0.10 *SD* over the subsequent 20 years of being evaluated. As noted above, estimates are not sensitive to multiple other parameterizations that assume teacher skill improves in some concave function. In these baseline scenarios, I assume that improvement from evaluation is equivalent across accountability and growth ($G_{it} = A_{it}$) and these two dimensions are perfect substitutes ($\sigma \rightarrow \infty$). For the Growth-only scenario, I set π at a value of 0.9 in favor of G_{it}, I assume standard rates of attrition, but no skill-biased attrition given that accountability pressures are typically given as rationale for labor market effects of evaluation. For the Combined Growth & Accountability scenario, I assume an equal distribution of emphasis across growth and accountability purposes of evaluation ($\pi = 0.5$). I further assume that there will be small improvements in the skills of entrants to the teaching profession (+0.05 *SD*) and a small degree of skill-biased attrition (+0.05 SD). For the Divided Growth & Accountability scenario, I assign a value of 0.9 to π for teachers outside of the windows for potential dismissal and a value of 0.1 for teachers inside these windows. I assume the same effects on the labor market and on skill-biased attrition as in the Combined scenario.

Under baseline assumptions, average teacher effectiveness improves more in Combined and Divided evaluation schemes than under Growth-only policies. Panel A of Figure 2 indicates that under the assumptions of the Growth-only model, teachers' average value-added scores would be expected to improve by 0.17 *SD* over the first ten years of the evaluation policy and stabilize at an improvement of 0.19 *SD*. For ease of comparison across models, I present the results of these simulations numerically in binned years in Table 2. In the baseline case, I assume that evaluation for accountability and growth have an equal effect, so there are no gains from shifting focus; therefore the results for the Combined and Divided evaluation frameworks are identical. In these two scenarios, I estimate value-added improvements of 0.18 *SD* in the first ten years of the evaluation policy and then a stabilization of gains from evaluation at around 0.26 *SD*, or a realized benefit above the Growth-only policy of 0.01 to 0.07 *SD*.

For each optimistic scenario, I assume that the effects of evaluation are to improve effectiveness by 0.13 *SD* via the mechanism which that particular accountability policy emphasizes for most teachers and by 0.03 *SD* on the dimension of evaluation which is de-emphasized. As such, in the Growth-only accountability policy, I assume that evaluation improves teacher effectiveness by 0.13 *SD* when it focuses on support and growth and by 0.03 *SD* when it focuses on accountability.[10] In the Combined Growth and Accountability policy, I stipulate the reverse. As the optimistic scenarios make the most favorable, reasonable assumptions for each evaluation framework, I stipulate that in the Divided framework teachers on both sides of the dismissal threshold are subject to evaluation supports or pressures that most improve their effectiveness. Thus, I set π equal to 0.9 for teachers outside of the dismissal threshold and equal to 0.1 for those at risk for dismissal. Across scenarios I assume that evaluation for growth/support and evaluation for accountability purposes are gross substitutes ($\sigma = 1.4$). Under the Combined and Divided Growth and Accountability evaluation frameworks, I specify that skill-biased attrition

[10] As a consequence of selecting rates of improvement from evaluation from the empirical literature that average to 0.08 *SD* (or 0.03 *SD* less than the 0.11 I assume in the baseline scenarios), initial improvements in effectiveness are below those of the baseline scenario. If I scale these improvements to the total improvement in the Baseline scenario, I observe an initial rate of improvement similar to the baseline models with the same relative differences across evaluation frameworks.

will occur equivalent to upper bounds (+0.08 *SD*) of the empirical evidence in Adnot et al. (2017).

I also stipulate that skills of entrants will improve across all scenarios, though via different mechanisms. In the Growth scenario, I theorize that the intensive supports will increase the attractiveness of the profession and 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 0.05 *SD* better than novice teachers at the start of the simulation. I assume a larger upper limit of 0.15 *SD* in the Combined and Divided scenarios. 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 in Winters and Cowen (2013) would be at work in the Divided framework as in the Combined scenario.

Both the Growth-only and Divided evaluation models outperform the Combined approach using the most optimistic, reasonable assumptions for each scenario. As Panel B of Figure 2 shows, rates of improvement in the Growth-only model outstrip those in the Combined and Divided models in the first 10 years because they take advantage of the large emphasis on the factor that is the most effective. After 10 years, however, the Divided model emerges as clearly superior generating returns over 0.35 *SD* from the start of the simulation and 0.05 to 0.1 higher than the Growth-only or Combined models.

For each pessimistic scenario, I specify that teachers improve from evaluation for accountability and growth at rates opposite those in the Optimistic scenario. Thus, for example, in the Growth-only accountability policy, I assume that evaluation improves teacher effectiveness by 0.03 SD when it focuses on support and growth and by 0.13 *SD* when it focuses on accountability. Further, I specify that these are upper-bounds of improvement that take 20 years to fully manifest following the concave function articulated in Table 1. I also model the possibility that a personnel management framework that has a high-degree of accountability and growth supports may inhibit teacher improvement. To do so, I specify that growth- and accountability-based evaluation are gross complements ($\sigma = 0.6$), indicating that changes in π produce changes in the marginal effectiveness of the other factor.

I also incorporate the possibility that some of these evaluation policy changes may alter the composition of latent skills in the supply of new teachers. Given evidence from Rothstein (2015) and Kraft et al. (2019) on the potential teacher shortages resulting from high-stakes evaluation, I specify that 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 also incorporate the possibility that some of these evaluation policy changes may negatively alter the composition of latent skills in the supply of new teachers. In both the Combined and Divided Growth and Accountability scenario, I assume that the average quality of the starting pool will decline to 0.15 *SD* worse than novice teachers at the start of the simulation. I assume this would not affect the Growth-only simulation. In all Pessimistic scenarios, I assume there is no skill-biased attrition.

Even using pessimistic assumptions, Growth-only and Divided evaluation policies improve overall teaching effectiveness over 40 years, but Combined approaches generate declines in average value-added. Rates of improvement in average teacher value-added for both the Growth-only and Divided scenarios plateau around 0.05 *SD* above starting values (Panel C). On the other hand, in the Combined evaluation framework average teacher true value-added declines over time to values worse than starting levels and 0.08 *SD* worse than the other two scenarios.

Due to weak correlations across outcome measures and yearly measurement error, the effects of the evaluation policies on teachers' value-added contributions to students' behavioral outcomes are substantially attenuated. Figure 3 replicates the parameters from Figure 2 with teachers' contributions to behavioral improvements in students that are correlated at 0.15 with test-score value added. Across all three evaluation frameworks in the most optimistic scenarios, impacts reach maximal values of 0.04 *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 (Jackson, 2018) are small in substantive magnitude.

Growth-only evaluation schemes lead to higher personnel costs due to the retention of more experienced staff. In Appendix Figure A1, I present the effect that each evaluation policy would have on the experience profile of teachers under baseline assumptions. I calculate the inter-quartile experience range and associated salary costs above an all-novice teaching force. To do so, I rely on Rothstein's (2015) estimate of the wage returns to experience of 0.015 per year. In the Growth-only evaluation framework (Panel A), teachers have more experience in the simulation due to lower rates of dismissal and attrition. The experience profiles resulting from each of the evaluation frameworks have significant budgetary implications. I plot the implied budgetary costs of the experience profile for each evaluation policy. The values on the second y-axis represent proportions of employee costs above a contract that employs only first-year teachers. The actual experience profile of U.S. teachers suggests a 19.1 percent added cost associated with experience given the current structure of teaching contracts and experience profile of teachers. The baseline Growth evaluation framework maintains this cost over time; whereas the Combined and Divided Growth and Accountability frameworks decline in cost as more experienced teachers are replaced with early-career ones such that costs stabilize around three percentage points lower than a Growth-only policy.

6 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. In this paper, I examine the assumptions underlying treating teacher evaluation for growth and accountability as substitutes or complements by developing a model of teacher effectiveness as a function of evaluation and presenting results of simple simulations that draw on the current causal evidence base. These simulations extend existing evidence by explicitly considering the conditions under which teachers improve through evaluation, rather than through dismissal alone. The results of these simulations provide evidence on the conditions required by modern evaluation systems to improve student outcomes.

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, meaningful differences in anticipated teacher effectiveness result from different types of evaluation policies, though these differences depend on particular assumptions. Under baseline scenarios, average teacher effectiveness in a growth-only evaluation policy results in 0.03-0.08 *SD* worse average teacher effectiveness than a policy that imposes some form of accountability pressures.

The effects of policies that impose accountability pressures and growth supports jointly on teachers depend in some part on the extent to which teachers experience growth and accountability as complements or substitutes. If teachers experience evaluation for growth and evaluation for accountability as substitutes (Figure 2, Panel B), joint-aim (Combined) evaluation policies would result in 0.05-0.07 *SD* worse teacher effectiveness than divided-aim policies. Ultimately, it is difficult to observe the effect of accountability on skill development independent from its effects on effort and from the effects of coaching activities. Thus, policy makers advance joint growth-and-accountability evaluation schemes under a condition of uncertainty.

Third, 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-only scenario, there is no ability-related attrition, 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 eliminate mid-career attrition (Years 20-34), teacher effectiveness in the Growth-only 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.

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 up to 10 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 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 less-experienced teaching force. This has important budgetary implications. Divided- or Combined-Accountability and Growth evaluation frameworks would result in substantial human resource savings, up to three percentage points of total employee expenditures. These could be reinvested in teacher salaries to counteract potential negative effects on labor supply.

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 much, if at all, 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 0.1 *SD*, the differences between optimistic and pessimistic scenarios within evaluation frameworks are around 0.3 *SD*. This does not imply that the structure of evaluation schemes does not matter, but rather that even under optimistic assumptions, a Combined accountability- and growth-evaluation model does not perform substantially better than either Growth-only or Divided models. Given the existing evidence, risks in pessimistic scenarios, and potential political roadblocks, these simulations imply substantial risks and smaller benefits from the combined evaluation model than policy observers have heretofore considered.

In general, I interpret these findings as providing suggestively positive results for a teacher evaluation system that imposes a performance floor, under which teachers would be subject to accountability pressures, and above which teachers would be given clear signals that they were subject to no accountability but would intensive receive coaching and other instructional supports. 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

Acemoglu, D. (2002). Directed technical change. *Review of Economic Studies*, 69(4):781–809.

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

Ariely, D., Gneezy, U., Loewenstein, G., and Mazar, N. (2009). Large stakes and big mistakes. *Review of Economic Studies*, 76(2):451–469.

Armstrong, M. (2000). Performance management. In Dransfield, R., editor, *Human Resource Management*, pages 69–84. Heinemann, Oxford, UK.

Arrow, K. J., Chenery, H. B., Minhas, B. S., and Solow, R. M. (1961). Capital-Labor substitution and economic efficiency. *The Review of Economics and Statistics*, 43(3):225.

Burgess, S., Rawall, S., and Taylor, E. S. (2019). Teacher peer observation and student test scores: Evidence from a field experiment in English secondary schools. Technical report, Annenberg Institute EdWorkingPaper: 19-139, Providence, RI.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2593–2632.

Cullen, J. B., Koedel, C., and Parsons, E. (2019). The compositional effect of rigorous teacher evaluation on workforce quality. *Education Finance and Policy*, pages 1–85.

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

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

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

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

Ederer, F. P. (2010). Feedback and motivation in dynamic tournaments. *Journal of Economics & Management Strategy*, 19(3):733–769.

Figlio, D. N. (2006). Testing, crime and punishment. *Journal of Public Economics*, 90(4-5):837–851.

Garet, M. S., Wayne, A., Brown, S., Rickles, J., Song, M., and Manzeseke, D. (2017). The Impact of Providing Performance Feedback to Teachers and Principals (NCESS 2018-4001). Technical report, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Gershenson, S. (2016). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy*, 11(2):125–149.

Glewwe, P., Ilias, N., and Kremer, M. (2010). Teacher incentives. *American Economic Journal: Applied Economics*, 2(3):205–227.

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

Goodman, S. F. and 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.

Grissom, J. A. and 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.

Holmstrom, B. and Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization*, 7(special):24–52.

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.

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*, 51(1):1–36.

Kraft, M. A., Brunner, E. J., Dougherty, S. M., and Schwegman, D. J. (2019). Teacher accountability reforms and the supply and quality of new teachers. Technical report, Annenberg Institute EdWorkingPaper: 19-169.

Kraft, M. A. and Gilmour, A. F. (2017). Revisiting The Widget Effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, 46(5):234–249.

Kraft, M. A., Papay, J. P., and Chi, O. L. (2020). Teacher skill development: Evidence from performance ratings by principals. *Journal of Policy Analysis and Management*, 39(2):315–347.

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

Muralidharan, K. and Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy*, 119(1):39–77.

Murnane, R. and 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.

Murphy, R., Weinhardt, F., and Wyness, G. (2020). Who teaches the teachers? A RCT of peer-to-peer observation and feedback in 181 schools. Technical report, CESifo Working Paper Series: No. 8221.

Neal, D. and Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics*, 92(2):263–283.

OECD (2019). *Working and Learning Together: Rethinking Human Resource Policies for Schools*. OECD Reviews of School Resources. OECD Publishing, Inc, Paris.

Papay, J. and 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.

Papay, J., Taylor, E., Tyler, J., and Laski, M. (2020). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. *American Economic Journal: Economic Policy*, 12(1):359–388.

Phipps, A. R. (2018). Personnel contracts with production uncertainty: Theory and evidence from teacher performance incentives. Available from: https://aaronphippseconomics.files.wordpress.com/2017/11/teacherperformanceincentives_111517.pdf.

Phipps, A. R. and Wiseman, E. A. (2019). Enacting the rubric: Teacher improvements in windows of high-stakes observation. *Education Finance and Policy*, pages 1–51.

Pope, N. G. (2019). The effect of teacher ratings on teacher performance. *Journal of Public Economics*, 172:84–110.

Popham, W. (1988). The dysfunctional marriage of formative and summative teacher evaluation. *Journal of Personnel Evaluation in Education*, 1(3):269–273.

Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, 73(2):417–458.

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

Ross, E. and Walsh, K. (2019). State of the States 2019: Teacher and Principal Evaluation Policy. Technical report, National Council on Teacher Quality, Washington, DC.

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

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

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

Sojourner, A. J., Mykerezi, E., and West, K. L. (2014). Teacher pay reform and productivity: Panel data evidence from adoptions of Q-Comp in Minnesota. *Journal of Human Resources*, 49(4):945–981.

Speroni, C., Wellington, A., Burkander, P., Chiang, H., Herrmann, M., and Hallgren, K. (2019). Do educator performance incentives help students? Evidence from the Teacher Incentive Fund National Evaluation. *Journal of Labor Economics*.

Stecher, B., Holtzman, D., Garet, M., Hamilton, L., Engberg, J., Steiner, E., Robyn, A., Baird, M., Gutierrez, I., Peet, E., Brodziak de los Reyes, I., Fronberg, K., Weinberger, G., Hunter, G., and Chambers, J. (2018). Improving Teaching Effectiveness: Final Report: The Intensive Partnerships for Effective Teaching Through 2015-2016. Technical report, RAND Corporation, Santa Monica.

Steinberg, M. P. and 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.

Steinberg, M. P. and 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.

Taylor, E. S. and Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, 102(7):3628–3651.

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).

Winters, M. A. and 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.

Parameters	Function	Details
SD of teacher test VAM[a]		0.15
r behavior VAM-test VAM[b]		0.15
SD of behavior-index VAM[c]		0.10
SD of noise in observed VAM[d]		0.183
r observed VAM-eval. score[e]		0.3
Experience effect on VAM[f]	r(exp)	$\begin{cases} \text{-0.07 if } yrs\_exp = 1 \\ \text{-0.04 if } yrs\_exp = 2 \\ \text{-0.02 if } yrs\_exp = 3 \\ 0.003(t) \text{ if } 3 < yrs\_exp \leq 15 \end{cases}$
Evaluation-based improvements from accountability[g] & growth[h]	$VAM = (\pi G_{it}^{\rho} + (1-\pi)A_{it}^{\rho})^{1/\rho}$ A_{it} and $G_{it} = \alpha(1 - e^{\beta yrs\_eval})$ if $yrs\_eval > 1$	Baseline: ρ=1, A_{it} and G_{it}=0.11 if yrs_eval=1; α=0.1, β =0.2 if $yrs\_eval > 1$
General attrition[i]	1. P($Attrite \mid yrs\_exp) = 1 - \alpha e^{-be^{-c(yrs\_exp)}}$ if $yrs\_exp < 20$ 2. Random attrition if cohort > 1.5% of teaching force when $20 \leq yrs\_exp < 35$ 3. Attrite if $yrs\_exp \geq 35$	Baseline parameters: α=1, b=0.18, c=0.4
Skill-biased attrition[j]	Small: P($Attrite \mid pctile\_eval \leq 10) = 1/6$ Large: P($Attrite \mid pctile\_eval \leq 10) = 1/2$	Small: attriters perform 0.05 *SD* worse Large: attriters perform 0.08 *SD* worse
Labor market changes[k]	$VAM = \alpha(1 - e^{-\beta(t^{\gamma})})$	$\alpha = \pm 0.05$ (small) or ± 0.15 (large), β=1/100, γ=2

[a]Rothstein (2015), Rivkin et al. (2005), Rothstein (2010), Chetty et al. (2014); [b]Gershenson (2016), Jackson (2018), Kraft (2019); [c]Jackson (2018); [d]Rothstein (2015), Sass (2008); [e]Kraft et al. (2020), Grissom and Loeb (2017), Rockoff et al. (2012); [f]Papay and Kraft (2015); [g]Phipps (2018), Pope (2019); [h]Burgess et al. (2019), Papay et al. (2020), Phipps (2018), Steinberg and Sartain (2015), Taylor and Tyler (2012); [i]Adnot et al. (2017); [j]National Teacher and Principal Survey (2017); [k]Rothstein (2015), Kraft et al. (2019).

Table 1: Functions and common parameters for simulations

	Baseline	Optimistic	Pessimistic
A. Growth			
Years 2-5	0.130	0.138	0.013
Years 6-10	0.167	0.185	0.023
Years 11-15	0.179	0.206	0.031
Years 16-20	0.187	0.224	0.040
Years 21+	0.192	0.247	0.050
B. Combined Growth & Accountability			
Years 2-5	0.129	0.089	0.025
Years 6-10	0.181	0.146	0.037
Years 11-15	0.213	0.199	0.025
Years 16-20	0.241	0.241	0.018
Years 21+	0.264	0.293	-0.019
C. Divided Growth & Accountability			
Years 2-5	0.130	0.139	0.017
Years 6-10	0.181	0.208	0.031
Years 11-15	0.213	0.265	0.039
Years 16-20	0.239	0.312	0.049
Years 21+	0.263	0.364	0.050

Notes: Average test score value-added estimates derived from simulation described in text.

Table 2: Comparison of average teacher test-score value-added in year bins under growth- and accountability-oriented evaluation policies

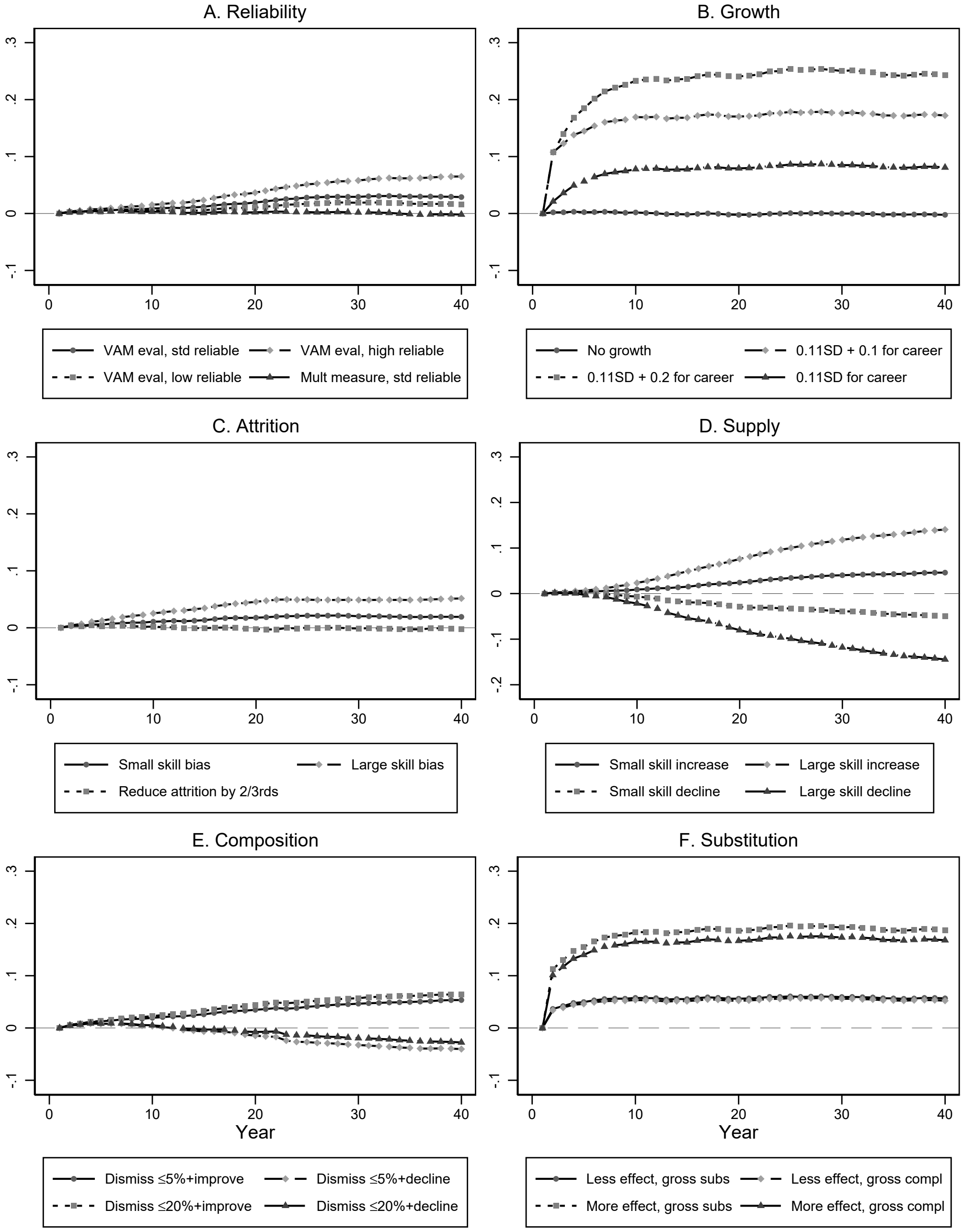

Figure 1: Average teacher test-score value-added profile over 40 simulated years, by varying functions and parameters of conditions related to teacher improvement from evaluation

Notes: Simulation descriptions in text. Parameters and functions listed in Table 1 and Appendix Table A1.

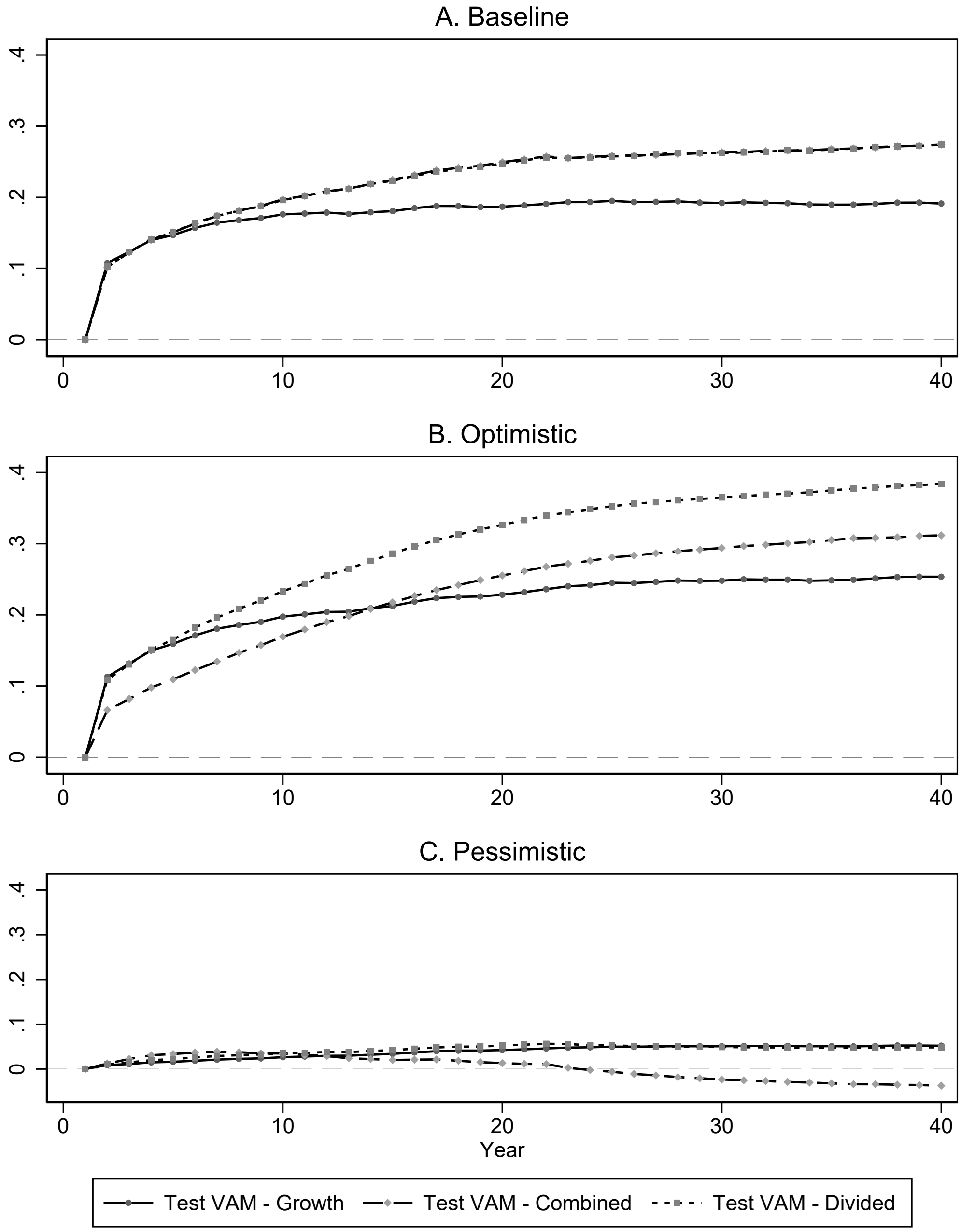

Figure 2: Average teacher test-score value-added over 40 simulated years under (A) Growth-only, (B) Combined Growth and Accountability, and (C) Divided Growth and Accountability evaluation policies

Notes: Simulation descriptions in text. Parameters and functions listed in Table 1 and Appendix Table A2.

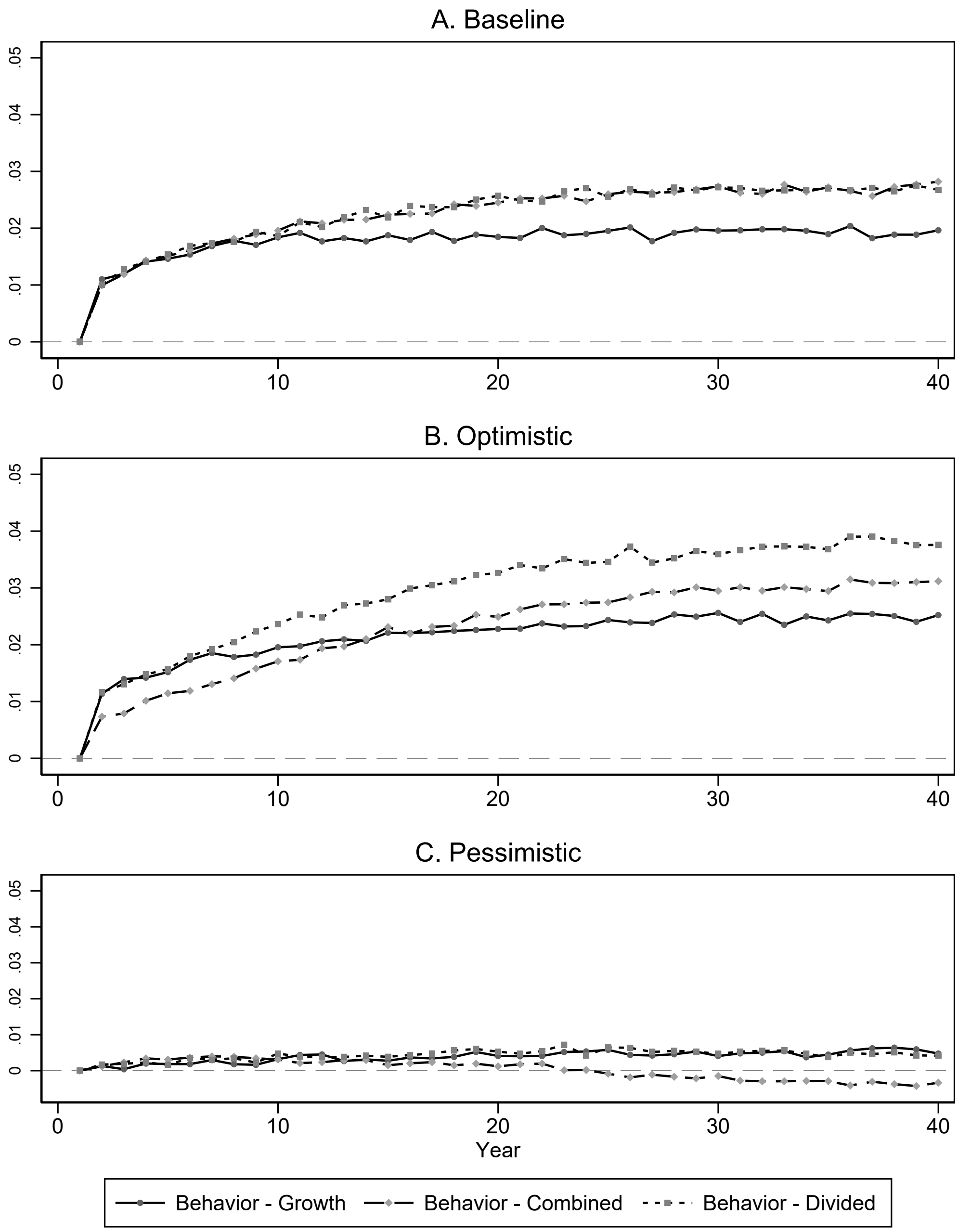

Figure 3: Average teacher value-added on index of behavioral outcomes (suspensions, absences and GPA) over 40 simulated years under (A) Growth-only, (B) Combined Growth and Accountability, and (C) Divided Growth and Accountability evaluation policies

Notes: Simulation descriptions in text. Parameters and functions listed in Table 1 and Appendix Table A2.

A Appendix Tables and Figures

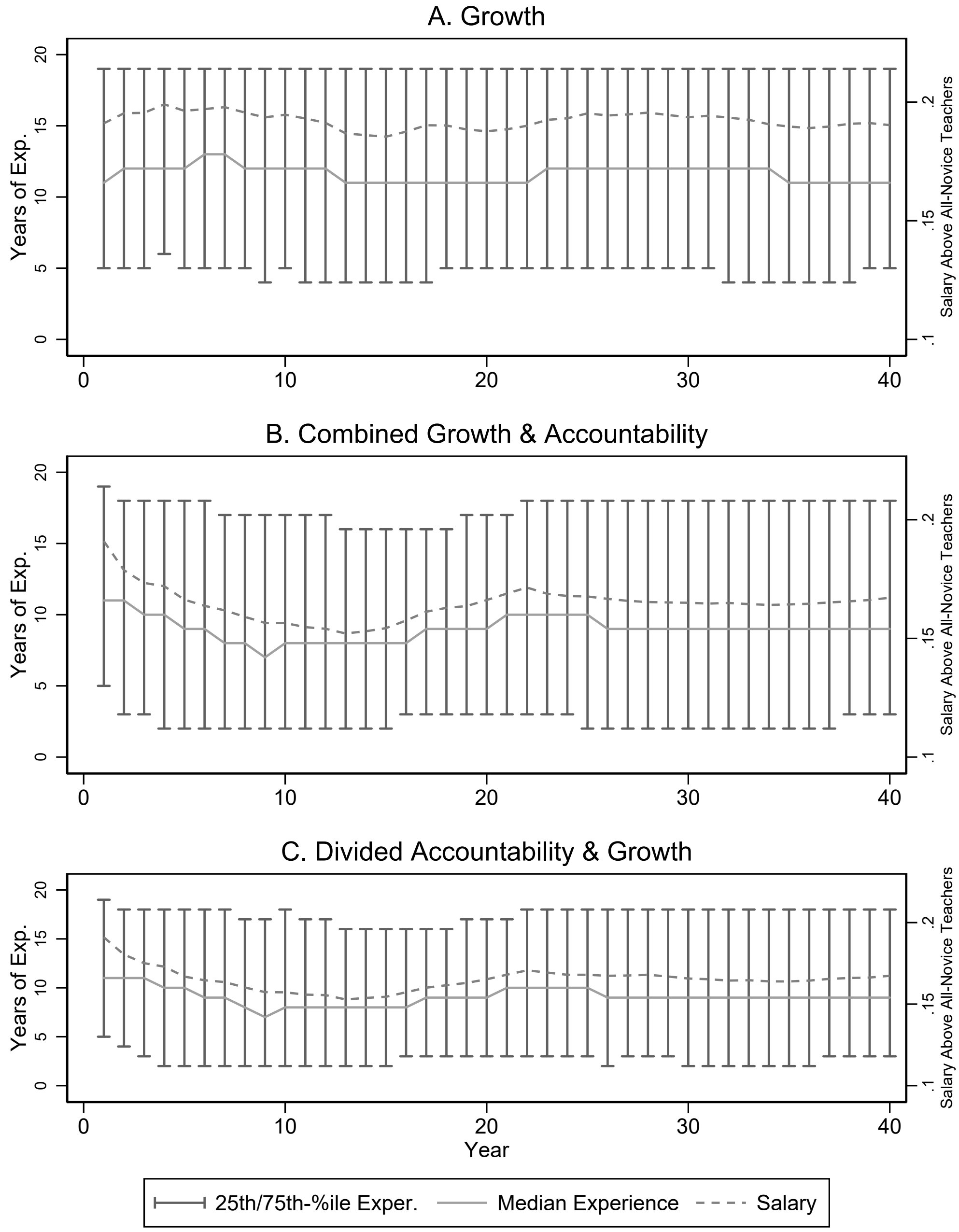

Figure A1: Teacher experience and experience-based compensation profiles over 40 years under growth- and accountability-oriented evaluation policies.

Notes: Simulation descriptions in text. Parameters and functions listed in Table 1 and Appendix Table A2.

Condition tested	Common	Scenario-specific
A. Reliability	General attrition Dismiss if percentile $\leq$20 after 3 years Standard r(*exp*) Growth from evaluation: none Labor supply: constant	a. *SD* noise in observed VAM = 0.183; VAM-only evaluation b. *SD* noise in observed VAM = 0.05; VAM-only evaluation c. *SD* noise in observed VAM = 0.30; VAM-only evaluation d. *SD* noise in observed VAM = 0.183; multiple-measure evaluation scheme
B. Improvement from evaluation	General attrition Dismissals: none Standard r(*exp*) Labor supply: constant	a. No growth b. +0.11 if *yrs_eval*=1, α=0.1, β=0.2 if *yrs_eval* > 1 c. +0.11 if *yrs_eval*=1, α=0.2, β=0.2 if *yrs_eval* > 1 d. α=0.11, β=0.2
C. Attrition	General attrition Dismissals: none Standard r(*exp*) Growth from evaluation: none Labor supply: constant	a. 1 in 6 of *pct_eval* $\leq 10 \rightarrow$ attriters 0.05 worse than mean performers w/in same *yrs_exp* cohort b. 1 in 2 of *pct_eval* $\leq 10 \rightarrow$ attriters 0.08 worse c. No skill bias, b=0.06 to reduce overall attrition by 2/3rds
D. Supply	General attrition Dismissals: none Standard r(*exp*) Growth from evaluation: none	a. Labor supply $\alpha = +0.05$ b. Labor supply $\alpha = +0.15$ c. Labor supply $\alpha = -0.05$ d. Labor supply $\alpha = -0.15$
E. Composition of entrants	General attrition Standard r(*exp*) Growth from evaluation: none	a. Dismiss if *pct_eval* $\leq$ 5 & *yrs_exp*=3; $\alpha = +0.05$ b. Dismiss if *pct_eval* $\leq$ 20 & *yrs_exp*=3; $\alpha = +0.05$ c. Dismiss if *pct_eval* $\leq$ 5 & *yrs_exp*=3; $\alpha = -0.05$ d. Dismiss if *pct_eval* $\leq$ 20 & *yrs_exp*=3; $\alpha = -0.05$
F. Substitution	General attrition Dismissals: none Standard r(*exp*) Labor supply: constant	G_{it}=0.13, A_{it}=0.03; π=0.9; σ=1.4 G_{it}=0.13, A_{it}=0.03; π=0.9; σ=0.6 G_{it}=0.13, A_{it}=0.03; π=0.1; σ=1.4 G_{it}=0.13, A_{it}=0.03; π=0.1; σ=0.6

Notes: Values correspond to simulations presented in Figure 1

Table A1: Contributions of evaluation conditions to teachers' value-added measures

A. Growth

Parameter	Baseline	Optimistic	Pessimistic
Learning from evaluation	$\pi = 0.9$, $\sigma = \infty$ $A_{it} = G_{it} = 0.11$ if $yrs\_eval = 1$ $\alpha = 0.1$ if $yrs\_eval > 1$	$\pi = 0.9$, $\sigma = 1.4$ $G_{it} = 0.13$, $A_{it} = 0.03$ if $yrs\_eval = 1$ $\alpha_G = 0.13$, $\alpha_A = 0.03$ if $yrs\_eval > 1$	$\pi = 0.9$, $\sigma = 0.6$ $\alpha_G = 0.13$, $\alpha_A = 0.03$ $\beta = 0.2$, $\gamma = 1$
Dismissal criteria	$pct\_eval \leq 1$	$pct\_eval \leq 1$	$pct\_eval \leq 1$
Skill-biased attrition	NA	NA	NA
Labor market supply changes	NA	$\alpha = 0.05$	NA

B. Combined Accountability & Growth

Parameter	Baseline	Optimistic	Pessimistic
Learning from evaluation	$\pi = 0.5$, $\sigma = \infty$ $A_{it} = G_{it} = 0.11$ if $yrs\_eval = 1$ $\alpha = 0.1$ if $yrs\_eval > 1$	$\pi = 0.5$, $\sigma = 1.4$ $G_{it} = 0.03$, $A_{it} = 0.13$ if $yrs\_eval = 1$ $\alpha_G = 0.03$, $\alpha_A = 0.13$ if $yrs\_eval > 1$	$\pi = 0.5$, $\sigma = 0.6$ $\alpha_G = 0.13$, $\alpha_A = 0.03$
Dismissal criteria	$pct\_eval \leq 5$ if $yrs\_eval < 3$ $pct\_eval \leq 10$ if $yrs\_eval = 3$ $pct\_eval \leq 5$ if $yrs\_eval > 3$	$pct\_eval \leq 5$ if $yrs\_eval < 3$ $pct\_eval \leq 10$ if $yrs\_eval = 3$ $pct\_eval \leq 5$ if $yrs\_eval > 3$	$pct\_eval \leq 5$ if $yrs\_eval < 3$ $pct\_eval \leq 10$ if $yrs\_eval = 3$ $pct\_eval \leq 5$ if $yrs\_eval > 3$
Skill-biased attrition	$\mathrm{P}(Attrite \mid pctile\_eval \leq 10 \cap$ $3 < yrs\_eval \leq 2) = 1/6$; $\mathrm{P}(Attrite \mid pctile\_eval \leq 15 \cap$ $yrs\_eval = 3) = 1/6$	$\mathrm{P}(Attrite \mid pctile\_eval \leq 10 \cap$ $3 < yrs\_eval \leq 2) = 1/6$; $\mathrm{P}(Attrite \mid pctile\_eval \leq 15 \cap$ $yrs\_eval = 3) = 1/2$	NA
Labor market supply changes	$\alpha = 0.05$	$\alpha = 0.15$	$\alpha = -0.15$

	C. Divided Accountability & Growth		
Parameter	Baseline	Optimistic	Pessimistic
Learning from evaluation	$\pi = 0.9$ if $pct\_eval > 5$ of dismissal threshold; $\pi = 0.1$ if $pct\_eval \leq 5$ of dismissal, $\sigma = \infty$; $A_{it} = G_{it} = 0.11$ if $yrs\_eval = 1$ $\alpha = 0.1$ if $yrs\_eval > 1$	If $pct\_eval > 5$ of dismissal: $\pi = 0.9$, $G_{it} = 0.13$, $A_{it} = 0.03$; If $pct\_eval \leq 5$ of dismissal: $\pi = 0.1$, $G_{it} = 0.03$, $A_{it} = 0.13$, $\sigma = 1.4$; $\alpha_G = 0.03$, $\alpha_A = 0.13$ if $yrs\_eval > 1$	$\pi = 0.9$ if $pct\_eval > 5$ of dismissal threshold; $\pi = 0.9$; if $pct\_eval \leq 5$ of dismissal, $\pi = 0.1$, $\sigma = 0.6$ $\alpha_G = 0.03$, $\alpha_A = 0.13$
Dismissal criteria	$pct\_eval \leq 5$ if $yrs\_eval < 3$ $pct\_eval \leq 10$ if $yrs\_eval = 3$ $pct\_eval \leq 5$ if $yrs\_eval > 3$	$pct\_eval \leq 5$ if $yrs\_eval < 3$ $pct\_eval \leq 10$ if $yrs\_eval = 3$ $pct\_eval \leq 5$ if $yrs\_eval > 3$	$pct\_eval \leq 5$ if $yrs\_eval < 3$ $pct\_eval \leq 10$ if $yrs\_eval = 3$ $pct\_eval \leq 5$ if $yrs\_eval > 3$
Skill-biased attrition	$\mathrm{P}(Attrite \mid pctile\_eval \leq 10 \cap 3 < yrs\_eval \leq 2) = 1/6$; $\mathrm{P}(Attrite \mid pctile\_eval \leq 15 \cap yrs\_eval = 3) = 1/6$	$\mathrm{P}(Attrite \mid pctile\_eval \leq 10 \cap 3 < yrs\_eval \leq 2) = 1/6$; $\mathrm{P}(Attrite \mid pctile\_eval \leq 15 \cap yrs\_eval = 3) = 1/2$	NA
Labor market supply changes	$\alpha = 0.05$	$\alpha = 0.15$	$\alpha = -0.15$

Notes: Values correspond to simulations presented in Figure 2, Figure 3 and Table 2.

Table A2: Values of parameters in complete evaluation framework simulations