Screening with Multitasking: Theory and Empirical Evidence from Teacher Tenure Reform*

Michael Dinerstein† Isaac M. Opper‡

January 2023

Abstract

What happens when employers screen their employees but only observe a subset of output? We specify a model with heterogeneous employees and show that their response to the screening affects output in both the probationary period and the post-probationary period. The post-probationary impact is due to their heterogeneous responses affecting which individuals are retained and hence the screening efficiency. We show that the impact of the endogenous response on both the unobserved outcome and screening efficiency depends on whether increased effort on one task increases or decreases the marginal cost of effort on the other task. If the response decreases unobserved output in the probationary period then it increases the screening efficiency, and vice versa. We then assess these predictions empirically by studying a change to teacher tenure policy in New York City, which increased the role that a single measure – test score value-added – played in tenure decisions. We show that in response to the policy teachers increased test score value-added and decreased output that did not enter the tenure decision. The increase in test score value-added was largest for the teachers with more ability to improve students' untargeted outcomes, increasing their likelihood of getting tenure. We estimate that the endogenous response to the policy announcement reduced the screening efficiency gap – defined as the reduction of screening efficiency stemming from the partial observability of output – by 28%, effectively shifting some of the cost of partial observability from the post-tenure period to the pre-tenure period.

*We thank Ran Abramitzky, Timothy Bresnahan, Raj Chetty, Liran Einav, Caroline Hoxby, Susanna Loeb, and Derek Neal for their early comments on the paper. We also thank conference participants at 2017 APPAM and 2017 AEFP and seminar participants at Brown University, the New York Federal Reserve, University of California - Irvine, and the University of Chicago Committee on Education for their helpful comments. We also thank Andrew McEachin, Christine Mulhern, and Lisa Abraham for their helpful feedback. We thank Terry Culpepper, Yiren Ding, Elena Istomina, Jora Li, and Jasper Snowden for research assistance. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A190148. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

† Kenneth C. Griffin Department of Economics, University of Chicago, NBER, and CESifo. Email: mdinerstein@uchicago.edu.

‡ RAND Corporation. Email: iopper@rand.org.

I Introduction

In many settings, the productivity of workers or institutions varies considerably.[1] Employers or policy-makers may therefore realize large gains from screening out low-performers after a probationary period, which has led to policy proposals for increased screening (Kraft et al., 2020). A key complication, though, is that worker or institutional output is multi-dimensional and rarely observed fully. Workers may then react to screening by distorting effort toward measured output, a concern in the education setting often referred to as "teaching to the test."

This concern is summarized by Campbell's Law, which states that "the more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (Campbell, 1979). While the literature usually worries about the distortion to current output, the distortion will also change the signal employers receive in evaluating employees (Neal, 2011). Understanding how this distortion affects the screening efficiency is crucial in determining the overall impact of screening policies.

We investigate the implications of screening on a single output in a multi-dimensional output environment and, in particular, how the endogenous response to the screening policy affects its efficiency. Theoretically, we specify a model where employees may distort their output in a probationary period to increase the probability they reach the post-probationary period. Holmstrom and Milgrom (1991) show that employees produce less of the untargeted output if and only if increased effort on the targeted task raises the effort cost on the untargeted task. We show that the same condition is necessary and sufficient for whether the endogenous response makes screening more efficient. Empirically, we study the teaching profession, where output is policy-relevant and most incentive-based policy is tied to a teacher's effect on test score value-added even though a recent literature has established that teachers affect a variety of student cognitive and behavioral outcomes (e.g., Jackson, 2018). We show that, relative to a policy in which the screening came as a surprise to the workers, the announcement of the screening policy leads to short-term losses in the untargeted measures due to multitasking but long-term gains due to improved screening. These gains have first-order policy effects: the endogenous response closes 28% of the reduction of screening efficiency stemming from the partial observability of output.

We start in Section II with a theoretical model that derives predictions about workers' behavioral responses to announced screening policies. A set of heterogeneous workers vary

[1] As examples, researchers have estimated large dispersion in quality across several types of agents: teachers (Chetty et al., 2014), doctors (Chan et al., 2022), and managers (Bertrand and Schoar, 2003). Estimated quality also varies across institutions like schools (Abdulkadiroğlu et al., 2020) and hospitals (Chandra et al., 2016). Quality may also differ across products, such as insurance plans (Abaluck et al., 2021).

in two dimensions – their returns to effort on two tasks. Workers are employed for up to two periods, and we consider a class of screening policies that make a worker's probability of reaching the second period an increasing function of first period output on the first task. An announced screening policy leads workers to exert more effort on the targeted task. Whether this leads to more or less output on the untargeted task depends on whether the increased effort on the first task decreases or increases the effort cost on the second task (Holmstrom and Milgrom, 1991). This response, though, is heterogeneous and depends on worker type. If workers who would produce higher output without incentives increase their targeted output the most with incentives, then the heterogeneous response makes screening more efficient. We show that this happens if and only if increased effort on the first task increases the effort cost on the second task. This condition is identical to the one that determines whether workers produce less untargeted output (multitask) in the probationary period. Our model thus predicts that multitasking and increased screening efficiency come as a package.

We then introduce our data in Section III, with a focus on how we measure teacher output in multiple dimensions. Here we build off of a recent literature that has explored teachers' effects on variables commonly collected by school districts but that capture different skills than test scores do (Gershenson, 2016; Jackson, 2018; Gilraine and Pope, 2020; Liu and Loeb, 2021). We extend beyond current test scores in two dimensions – timing and type of measure. For timing, we include measures of student outcomes in future years. For additional measures, we focus on student grades and attendance, which may capture non-cognitive or behavioral skills less related to what tests evaluate.[2] Consistent with the literature, we find that while teachers' effects are correlated across different output measures, these correlations are far from 1 and thus entail information not immediately revealed by test score value-added. We further anchor teachers' effects on different outputs to their effects on student graduation rates, which creates a one-dimensional measure of policy effectiveness.

In Section IV, we introduce the empirical context by describing teacher tenure policy in New York City (NYC) and how it changes over time. Prior to the 2009-2010 school year, nearly all teachers received tenure following their third year in NYC, and the decision had no explicit consideration of a teacher's test score value-added. Then in 2009-2010, the district changed the process to incorporate test score value-added, which coincided with a sharp 30 percentage point decrease in tenure rates. We translate these policy details into

[2]While the non-test score based measures are readily available to school districts, the future outcomes are not immediately available and thus district screening policy that requires immediate decisions on whether to give a teacher tenure does not necessarily have such measures at its disposal. As researchers analyzing data from prior years, we do not face that constraint and can thus characterize the policy's effect on multiple dimensions.

an empirical strategy to estimate the causal effect of the screening policy's incentives on teacher output.[3] As the policy's implementation is sudden and only affects teachers yet to receive tenure, we have a variety of useful comparison groups that allow us to control for confounders. Already-tenured teachers do not see a change to their incentives and thus they let us control for yearly shocks to students and teachers in NYC. Pre-tenure teachers in years prior to the policy change let us control for experience effects. And teachers for whom the policy change hit in the middle of their pre-tenure periods let us control for cohort or teacher effects that might reflect changes in entering cohorts' quality over time. We thus isolate causal effects by comparing how a cohort or teacher's outcomes change upon the introduction of the new policy, controlling for year and experience effects.

We use this empirical strategy in Section V to test how the policy change affected pre-tenure output. We focus on both targeted (test score value-added) output measures and untargeted output measures. We estimate that on the targeted measure the introduction of the screening policy increases output by 0.019-0.033 student standard deviations, depending on the estimator. This effect represents a meaningful fraction of the cross-sectional variation in teacher output in the unincentivized period. Based on the cross-sectional relationship between test score value-added and the other measures, we might expect the other measures to increase as well since the measures are positively correlated. Instead, we estimate that teacher output on the untargeted measures falls in response to the tenure policy. This negative effect on other outcomes is consistent with teachers substituting effort away from building student skills that persist (or are tested on future assessments) or non-tested skills and toward effort improving students' current test scores. Such substitution has immediate consequences. While the policy is focused on screening out low-performing teachers, in the pre-tenure period it has a positive effect on output in the targeted measure but negative effect on the untargeted measures.

Such changes could be temporary or persistent, depending on whether the change in effort alters teachers' future production functions. We estimate that a teacher's current test score value-added reverts to unincentivized levels (excluding the experience gradient) once she receives tenure and no longer faces the incentive. Thus, the policy had large, temporary incentive effects; the policy's total effect, however, also depends on how teachers are screened into and out of the district in the post-probationary period. How the policy affects screening in turn depends not only on the average incentive effect, but on how different teachers respond differently. Based on the substitution in the pre-tenure period, our model predicts that teachers with higher types (i.e., ability) will substitute more toward the targeted task in the presence of incentives. In Section VI, we use forecasts from a

[3]Loeb et al. (2015) study how this tenure policy change affected teacher quality in the district by replacing marginal teachers at the end of their probationary period with new teachers.

multi-dimensional value-added model to estimate a teacher's type. We then test whether a teacher's response varies with her type and find that teachers with higher (unincentivized) output on either dimension respond more strongly to the implied incentive on the targeted task. Therefore, the behavioral response to the incentive selects positively on the untargeted dimension.

We put these pieces together in Section VII and estimate the behavioral response's impact on the tenure policy's screening efficiency. Given that the policy is a threshold one (promote the top 67% based on test score value-added), the behavioral responses only matter for marginal teachers; unsurprisingly, we estimate that tenure outcomes change for just 4% of teachers. The teachers who only receive tenure due to the behavioral response, though, have considerably higher output on the untargeted dimension compared to the teachers who lose tenure (239% of the cross-sectional standard deviation). In contrast, these teachers have lower output on the targeted dimension, though these differences are minimal. The quality of the tenured teacher pool thus increases by 7.7% of a standard deviation on the untargeted dimension, while decreasing by 0.76% of a standard deviation on the targeted dimension. Together, these results suggest that the endogenous response to the policy reduced the screening efficiency gap – defined as the reduction of screening efficiency stemming from the partial observability of output – by 28%.

In sum, the behavioral response changes the types of teachers screened in to tenure, while simultaneously distorting output in the pre-tenure period. The behavioral response shifted some of the cost of partial observability of output from the post-tenure period to the pre-tenure period. Because the changes in the two dimensions differ in the pre-tenure period versus the post-tenure period – targeted output is higher pre-tenure and lower post-tenure while untargeted output is the reverse – whether the district benefits from the behavioral response depends on the length of the post-tenure period. We estimate that the district benefits from the behavioral response if teachers remain in the district between 3.5 and 45 years post-tenure.

I.A Related Literature

This paper brings together the literatures on screening, multi-dimensional output, and multitasking. From the screening perspective, our focus is on a monopsonist employer's on-the-job screening.[4] In many industries like education, the employer institutes an up-or-out policy where the employee partially reveals her type during a probationary period and then the employer decides whether to keep the employee, often under favorable contract terms

[4]The broader screening literature considers policies where employees select into or out of a job (e.g., Bénabou and Tirole, 2016; Brown and Andrabi, 2021) or policies where the employer screens job applicants (e.g., Spence, 1973).

like tenure. In the case where the employee's type is one-dimensional, the literature argues that these policies can improve screening efficiency (O'Flaherty and Siow, 1992; Demougin and Siow, 1994; Chen and Lee, 2009; Barlevy and Neal, 2019)[5] and reduce multitasking in the post-probationary period (Kou and Zhou, 2009). When an employee's type is multi-dimensional (Armstrong and Rochet, 1999), agents may manipulate the signals they send. Spence (1973) and Nichols and Zeckhauser (1982) offer models of multi-dimensional agents where a single-crossing condition means that the agent's signal manipulation reveals her type and thus improves screening efficiency. A recent literature has extended the framework to include flexible unobservable heterogeneity in manipulation costs. In Frankel and Kartik (2019), for example, this addition lowers screening efficiency, especially at high stakes.[6]

In the theoretical screening literature, manipulation usually matters to the extent it affects screening efficiency. We focus instead on a form of manipulation that affects the principal both directly, through distorted output in the probationary period, and indirectly, through a distorted signal. The distorted output ties into a literature on multitasking, which characterizes the optimal incentive contract when the principal cannot contract on all forms of output (Holmstrom and Milgrom, 1991; Baker, 1992). DeVaro and Gürtler (2016a), DeVaro and Gürtler (2016b), and DeVaro and Gürtler (2020) consider a related, though distinct, situation where agents strategically shirk in the first period not because some forms of output cannot be contracted on but because they seek to signal their suitability for a job with different tasks. Unlike many papers in these large literatures, we take the contract as given. Instead, we develop a new theoretical result that links output changes across the probationary and post-probationary periods. In Neal (2011)'s terminology, we characterize the relationship between Campbell's Law and "Campbell's Law turned on its side."

These topics have also been explored in the empirical literature, especially in the economics of education. Building on a large literature estimating wide dispersion in teacher test score value-added (e.g., Jacob and Lefgren, 2008; Staiger and Rockoff, 2010a; Kane et al., 2013; Chetty et al., 2014) that is difficult to predict at the time of hiring (Rockoff et al., 2011), Staiger and Rockoff (2010b) argue that policy should focus on aggressive screening of early-career teachers. A recent wave of teacher tenure reforms, including the one we study, have led researchers to estimate the effects of stringent dismissal policies on teacher (targeted) output before (Taylor, 2022), during (Jacob, 2013; Dee and Wyckoff, 2015; Taylor, 2022), and after the evaluation period (Ng, 2021; Taylor, 2022).

Meanwhile, an emerging literature has shown that teachers affect a range of student

[5]Barlevy and Neal (2019)'s model rationalizes the combination of up-or-out policies and heavy workloads during the probationary period.

[6]A large empirical literature has studied the efficiency of different screening devices, such as hassle costs (e.g., Deshpande and Li, 2019). Björkegren et al. (2020) develops an estimator of an optimal screening function that is robust to manipulation.

outcomes and that teacher effectiveness across these dimensions is only weakly correlated (Gershenson, 2016; Jackson, 2018; Kraft, 2019; Petek and Pope, 2016; Gilraine and Pope, 2020; Liu and Loeb, 2021). These correlations, though, are likely dependent on the incentive environment (Neal and Schanzenbach, 2010; Duflo et al., 2011; Neal, 2011). Given the multidimensionality of teacher output, education researchers have worried that teachers might multitask ("teach to the test") in response to incentive programs like performance pay (Neal, 2009). Several papers (Jacob, 2005; Glewwe et al., 2010; Corcoran et al., 2011; Fryer and Holden, 2012; Macartney et al., 2018) find treatment effects are stronger in targeted outcomes while Muralidharan and Sundararaman (2011) find similar effects on targeted and untargeted subjects. We argue that such multitasking may also arise from screening policies (Neal, 2010). Our empirical contribution is to show how multitasking affects screening efficiency.[7]

II Model

We develop a two-period model of heterogeneous agents (teachers) exerting effort on two tasks. A principal (school district) decides which agents to retain for the second period after observing only one of two outputs. Our goal is to develop predictions about how agents will respond to the principal's screening rule and, in particular, provide conditions under which the endogenous response improves (or worsens) screening efficiency. We present results in this section and provide proofs in Appendix A.

II.A Model Set-Up

Production of Student Outcomes and Teachers' Utility: We assume each teacher has some fixed type $\theta \in \Theta = [\underline{\theta}_1, \overline{\theta}_1] \times [\underline{\theta}_2, \overline{\theta}_2]$ and in each period chooses effort level $e \in \mathcal{E} = [\underline{e}_1, \overline{e}_1] \times [\underline{e}_2, \overline{e}_2]$. We assume that θ is continuously distributed over Θ and that all individuals produce some of each output; i.e., $\underline{\theta}_k > 0$ and $\underline{e}_k > 0$ for $k \in \{1, 2\}$. The type and effort level combine to affect student outcomes x in the following way: if the k^{th} dimension of the individual's type is θ_k and she exerts effort level e_k on task k, then she improves student outcome k by $x_k = e_k\theta_k$. We will refer to the first outcome as the teacher's "test score value-added."

We assume that teachers get value from increasing students' outcomes but incur a cost of effort, which limits the amount of x that each teacher produces. Specifically, we assume teachers' per-period utility is given by a function $b(x) - c(e)$ for some twice-differentiable concave benefit function b that is increasing in x and a twice-differentiable strictly convex

[7] De Philippis (2021) studies a university policy that simultaneously induced multitasking and compositional changes via selective attrition. We directly link the compositional change to the form of multitasking.

cost function c that is increasing in e. For simplicity, we will add the restriction that $b(x) = x_1 + x_2$, though we can relax this assumption at the cost of more notation.[8] With these assumptions, the interplay between the two tasks is completely determined by the cross-derivates of c, $\frac{\partial^2 c}{\partial e_1 \partial e_2}$. It is natural to assume that $\frac{\partial^2 c}{\partial e_1 \partial e_2} > 0$, which captures the idea that teachers face a tradeoff when deciding how to allocate their effort; one can also imagine contexts, however, where there are positive spillovers across tasks. We can model this positive spillover as one task getting easier when more effort is spent on the other task: $\frac{\partial^2 c}{\partial e_1 \partial e_2} < 0$. The teacher's per-period utility is $u(e, \theta) = b(\theta_1 e_1, \theta_2 e_2) - c(e_1, e_2)$, which we can equivalently express as utility over output: $u(x, \theta)$.

We denote a teacher's indirect utility – when she optimizes without dynamic considerations – as $v(\theta) = \max_{e \in \mathcal{E}} u(e, \theta)$. We denote the optimal choice of effort – again when she optimizes without dynamic considerations – as $e^*(\theta) = \arg\max_{e \in \mathcal{E}} u(e, \theta)$, with effort on task k denoted as $e_k^*(\theta)$. The effect on student outcome k under these optimal choices is $e_k^*(\theta) \cdot \theta_k$, which we denote as $x_k^*(\theta)$.

Defining the Screening Policy: We will assume that time can be split into discrete periods (e.g., school years) and that the pool of teachers is fixed within a period. We allow teachers to stay in the teaching profession for at most two periods: a pre-tenure period followed by a post-tenure period. Whether an individual teacher is granted tenure, i.e., allowed to stay in the profession for the second period, is determined by a function $p : X_1 \to [0, 1]$. That is, the probability that a teacher with test score value-added in the first period of x_1 is granted tenure is $p(x_1)$. We will assume the screening policy takes the form of a threshold policies in which $p(x_1) = \mathbf{1}\big(x_1 \geq \xi\big)$ for some threshold ξ. Such threshold policies will generally be the optimal policy and restricting to this class of policies will make comparing announced and unannounced policies straightforward. Note that though a teacher affects multiple outcomes, whether she gets tenure only depends on her output in the first dimension.

Dynamic Optimization: In our two-period model, teachers place a weight of one on the first period and a weight of $\lambda > 0$ on the second period. To simplify the analysis, we assume that all individuals place the same value on staying in their role. Specifically, if the outside option for an individual of type θ is $\tilde{v}(\theta)$, we assume that $v(\theta) - \tilde{v}(\theta) \equiv \Delta v$, where Δv does not depend on θ. This is a reasonable assumption if the outside option is a similar role as the one she is currently in, e.g., a teaching job in a different district, at a charter school, or at a private school; in that case, Δv corresponds to the cost of finding and switching jobs rather than the cost of taking a fundamentally different role, which is

[8] The crucial assumptions are twofold: one, that b is not too concave, from which we can conclude that $e_1^*(\theta)$ is increasing in θ_1; two, that the cross derivative of b is the opposite sign of the cross derivative of c, which means that the same forces on the substitution between tasks apply to b as they do to c.

plausibly independent of individual type. While this assumption is an important one, the analysis can be extended to allow for the relative value of staying in the role to vary based on θ with predictable results. We will assume that $\Delta v > 0$, so individuals have an incentive to stay in their current role.

It then follows that teachers choose effort in the first period to maximize: $u(e, \theta) + \lambda \cdot p(e_1\theta_1) \cdot \Delta v$. We denote the first period optimal effort choices and effects on student outcomes for an individual of type θ under some screening policy p as:

$$e^*(\theta|p) = \arg\max_{e \in \mathcal{E}} u(e, \theta) + \lambda \cdot p(e_1\theta_1) \cdot \Delta v \tag{1}$$

$$x_k^*(\theta|p) = e_k^*(\theta|p) \cdot \theta_k. \tag{2}$$

II.B Response to the Screening Policy

Even though the policy's main aim is to screen workers, it induces an incentive effect. In particular, the screening policy increases the incentive for teachers to have high test score value-added. This causes teachers to exert more effort on the first task, which in turn increases teachers' test score value-added in the pre-tenure period. How the incentives affect the other outcome depends crucially on how the increased effort on the first task affects the marginal cost of effort on the second task. If it makes effort on the other task more costly, then this extra effort spent increasing test score value-added leads to a decrease in other output. This is akin to the multitasking model of Holmstrom and Milgrom (1991) and reflects the concern that adding incentives to test score value-added measures would lead to more "teaching to the test." In contrast, if positive spillovers mean that the increased effort on the first task makes effort on the second task less costly, then the extra effort spent increasing test score value-added leads to an increase in the other output. Formally, we have the following theorem:

Theorem 1. *For any weakly increasing screening function, $x_1^*(\theta|p) \geq x_1^*(\theta)$ for every θ. Furthermore, for every θ:*

$$x_2^*(\theta|p) \leq x_2^*(\theta) \qquad \textit{if} \qquad \frac{\partial^2 c}{\partial e_1 \partial e_2} \geq 0$$

$$x_2^*(\theta|p) \geq x_2^*(\theta) \qquad \textit{if} \qquad \frac{\partial^2 c}{\partial e_1 \partial e_2} \leq 0.$$

This endogenous response clearly affects output in the first period. More subtly, the response may also distort output in the second period by changing which individuals are screened in and out. Whether this improves or harms second period output is the focus of the next subsection.

II.C Impact on Screening Efficiency

Our main concern is to characterize conditions under which we can conclude the endogenous response described in Section II.B makes the screening more (or less) efficient. To do so, we will set up a comparison with the (infeasible) policy that screens on $x_1^*(\theta)$, i.e., that screens on the equilibrium outcomes that occur absent the endogenous response. As a matter of terminology, we will call this an ex post (or surprise) screening policy.

To determine the efficiency of the screening, we take as given the principal's value function – denoted $\tilde{V}(x)$ – which determines how valuable it is for her to keep a teacher who will produce x in the post-tenure period. We can also define an equivalent value function of teacher type as: $V(\theta) = \tilde{V}\big(x^*(\theta)\big)$.[9] We assume that $\tilde{V}(x)$ is increasing in x and that $V(\theta)$ is increasing in θ.[10]

One challenge with understanding how the response changes screening efficiency is that the response not only changes who is screened in and out, but may also change how many individuals are screened in and out. We focus only on comparing policies that remove the same fraction of teachers. In particular, for any ex post screening policy p we call its comparison ex ante screening policy $\tilde{p}$ to be the one that retains the same fraction of individuals.[11] We can therefore define efficiency as follows:

Definition 1. *Consider an ex post screening policy p and its comparison screening policy $\tilde{p}$. We then say that the endogenous response makes screening* ***more efficient*** *if for every pair of threshold policies p and $\tilde{p}$ the average value of $V(\theta)$ is higher for those teachers who remain after policy $\tilde{p}$ than after policy p, i.e.:*

$$\mathbb{E}\Big[V(\theta)\cdot\tilde{p}(\theta)\Big] \geq \mathbb{E}\Big[V(\theta)\cdot p(\theta)\Big].$$

Conversely, we say that the response makes screening ***less efficient*** *if for every pair p and $\tilde{p}$ the average value of $V(\theta)$ is higher for those teachers who remain after policy p than after policy $\tilde{p}$.*

With this definition, we can introduce our main theorem:

Theorem 2. *There exists κ such that if either: a) $\lambda\cdot\Delta v \geq \kappa$ or b) $c(e)$ is quadratic then:*

1. *The endogenous response makes screening more efficient if $\frac{\partial^2 c}{\partial e_1 \partial e_2} > 0$;*

[9] Writing the equivalent functions this way implicitly assumes that teachers revert back to producing $x^*(\theta)$ when the incentive to remain disappears.

[10] Note that these two assumptions are not redundant. For example, if $\frac{\partial^2 c}{\partial e_1 \partial e_2} < 0$ and $\tilde{V}(x) = x_1 + \alpha x_2$, then for a very small α, $V(\theta)$ is decreasing in θ_2 even though $\tilde{V}(x)$ is increasing in both x_1 and x_2.

[11] In Appendix A, we show that for every $\rho \in [0,1]$ there exists a threshold $\tilde{\xi}$ such that the screening policy $\tilde{p}(x_1) = \mathbf{1}(x_1 \geq \tilde{\xi})$ retains exactly the fraction ρ of individuals. Thus, every ex post screening policy p has a comparison ex ante screening policy $\tilde{p}$.

2. The endogenous response makes screening less efficient if $\frac{\partial^2 c}{\partial e_1 \partial e_2} < 0$.

Proof Intuition. We first consider the case in which the incentive to stay is large, i.e., $\lambda \cdot \Delta v \geq \kappa$, and show in Figure 1 how it affects which teachers receive tenure.. We consider the extreme in which individuals' only concern is staying in the profession. In this case, the cost of deviating in the first period to produce a suboptimal amount of x_2 is trivial relative to the potential incentive of staying in the profession. The screening policy would then select individuals based purely on how costly it is for them to produce x_1; i.e., it screens individuals based solely on θ_1. (Showing that it is indeed possible to think of this extreme case as representative of what happens when $\lambda \cdot \Delta v$ gets large is one reason why the actual proof is more involved.)

From this logic, we can map in Θ-space the individuals who are retained and removed after the endogenous response by simply drawing a horizontal line at a specific θ_1; anyone above that line will be retained and anyone below that line will be removed. Hence, we refer to this line as the "endogenous response threshold." Similarly, we can map in X^*-space the individuals who are retained and removed *absent* the endogenous response by simply drawing a horizontal line for a specific $x_1^*(\theta)$; anyone above that line will be retained and anyone below that line will be removed. Hence, we refer to this line as the "ex post threshold." The question therefore is how Θ-space is partitioned absent the endogenous response and/or how X^*-space is partitioned after accounting for the endogenous response, which is where $\frac{\partial^2 c}{\partial e_1 \partial e_2}$ matters.

If $\frac{\partial^2 c}{\partial e_1 \partial e_2} > 0$, we get that $x_1^*(\theta)$ is decreasing in θ_2 which in turn implies that the ex post threshold in Θ-space is upward sloping. In contrast, we get that the endogenous response threshold in X^*-space is downward sloping. Thus, in Figure 1, the space A indicates individuals who are retained when there is an endogenous response and removed in the ex post screening case; the space B indicates the opposite. Whether individuals in A are preferred to individuals in B is ambiguous when viewing it in X^* space, but from the assumption that $V(\theta)$ is increasing θ it is clear in Θ-space that everyone in A is preferred to everyone in B. Thus, if $\frac{\partial^2 c}{\partial e_1 \partial e_2} > 0$ the endogenous response – at the extreme – makes screening more efficient.

If $\frac{\partial^2 c}{\partial e_1 \partial e_2} < 0$ the results are symmetric, with the "ex post threshold" being downward sloping in Θ-space and the "endogenous response threshold" being upward sloping in X^*-space. Here, whether individuals in A are preferred to individuals in B is ambiguous in Θ-space, but from the assumption that $\tilde{V}(x)$ is increasing in x it becomes clear in X^*-space that everyone in B (i.e., above the ex post threshold and below the endogenous response threshold) is preferred to everyone in A (i.e., below the ex post threshold and above the endogenous response threshold). Hence the endogenous response – again at the extreme – makes screening less efficient.

For smaller incentives, we can no longer ignore the cost of substituting away from x_2, and specifically the fact that individuals vary in how costly this substitution is, when considering which individuals are retained. To illustrate the logic for smaller incentives, we will again consider the extreme case; here, the extreme case is one in which individuals place virtually no value on remaining within the profession. Individuals' response to the policy will thus be almost non-existent and therefore the only individuals who might have their retention outcome changed based on their endogenous response to the announcement are ones whose unincentivized production of x_1 is essentially identical to the announced threshold. In other words, we only need focus on how individuals on the "ex post threshold" line respond to the incentive. A key advantage of just focusing on individuals on the "ex post threshold" line is that we are, by definition, comparing individuals who all chose the same x_1 output absent any additional incentive. (Again this is a loose argument, with the formal proofs relegated to Appendix A.)

How one individual who produces some amount of x_1^* responds differently from another who produces the same x_1^* depends on the relative costs of the two individuals to deviate away from this same initial optimum. This in turn depends on how convex their respective costs function are, i.e., on the relative sizes of the second derivatives of their cost function. This will in general depend on the third derivatives of the cost function, since they choose different effort levels to produce the same x_1. But if the third derivatives are all zero, i.e., $c(e)$ is quadratic, and $\frac{\partial^2 c}{\partial e_1 \partial e_2} > 0$, we show in Appendix A that higher θ individuals will find it less costly to deviate from their optimum and hence respond more to the policy; since the higher θ individuals respond more, the response makes the screening more efficient. In contrast, the opposite is true if $c(e)$ is quadratic and $\frac{\partial^2 c}{\partial e_1 \partial e_2} < 0$.

□

II.D Model Implications

Our model has several implications for policy and empirical analysis. First, we show that the multitasking response may be heterogeneous by teacher type, and that the heterogeneity can make screening better or worse depending on the effort cost function. This warrants empirical analysis that examines heterogeneity in how teachers respond to incentives.

Second, the model implies that whether teachers substitute away from unincentivized tasks and whether that response improves screening efficiency are determined by the same condition on the effort cost function. While not all incentive schemes may induce such substitution, the model implies when it happens the substitution will be concentrated among the agents with higher abilities on that task. This induces a negative correlation between total pre-tenure output (on the unincentivized task) and post-tenure output, which smooths out any welfare losses from multitasking. We will test in our empirical setting whether this

joint prediction holds.

Finally, our theoretical predictions concern the directions of effects. We will quantify the magnitude of these channels with empirical analysis.

III Data and Outcomes

III.A Data and Measuring Output

The model suggests that teachers respond to the incentives along multiple dimensions and that responses are likely to be heterogeneous across teachers. An ideal data set for empirical analysis thus has several features. First, it must include both measures of output that are targeted by incentives and measures of output that are untargeted and which capture skills distinct from the targeted measures. Relatedly, having a long-term outcome – and a long enough panel to observe the outcome – helps characterize the trade-off between gains in different short-run output measures based on how well they predict the long-term outcome. Second, the data set should track teachers over time and observe them with different levels of incentives. Such panel data, with within-teacher incentive variation, would allow the study of how different types of teachers respond differentially to incentives.

We meet these needs by turning to yearly administrative data from the New York City Department of Education (NYCDoE). This data includes all NYC public schools from 2006-07 through 2014-15, which is a long enough panel to follow teachers over time and to observe longer-term outcomes for early cohorts. We infer a teacher's output from the outcomes her students achieve. For each student-year observation, we observe the student's school and grade and end-of-year (externally graded) test scores, the outcome that personnel policy will target. We further observe other student outcomes – attendance rates and grades in tested and other subjects (for middle school students) – that we will use as untargeted outcomes, as we describe below, and students' high school graduation status, a longer-run outcome. We also observe student demographic information and thus can construct a broad set of control variables that will help us isolate a teacher's impact. Importantly, we observe a mapping between the student and the teacher she had in each subject-year. This mapping lets us link student outcomes to individual teachers and thus construct measures of teacher output.

On the teacher side, we can follow teachers over time and across schools within NYC. Our policy variation will depend on whether a teacher is tenured, and how the timing of her pre-tenure period lines up with policy changes. Our data include teachers' experience each year and whether they are tenured in NYC.

We provide summary statistics for the student and teacher data in Table 1. Nearly 80% of our student sample is eligible for free or reduced price lunch (high-poverty), and just

over 10% is an English language learner. We normalize the test scores to have mean 0 and standard deviation 1 for each grade-year. Students' attendance rates are relatively high, averaging 94% with a standard deviation of 6%, while grades vary between 10 and 100 with means near 80% and standard deviations of about 10%. The mean teacher in our data has been in the district for nearly 9 years and at the same school for about 6 years. Just over one-fifth of the teacher-years correspond to the pre-tenure (probationary) period, which will be the focus of policy variation. NYC's size leaves us with large sample sizes that enable us to have powerful empirical tests. We have nearly 29,000 teachers in the sample, and they contribute almost 100,000 teacher-years. At the subject-year level, we have an average of 32 students per teacher, such that the mean of a teacher's students' outcomes carries some signal of the teacher's output.

III.B Constructing Teacher Outcomes

We seek to estimate changes in teacher output along targeted and untargeted dimensions. In choosing student outcomes that may capture skill development separate from contemporaneous test score measures, we turn to a recent literature that estimates teacher effects on several student outcomes. Jackson (2018) shows that teachers' effects on attendance and grades – student outcomes regularly collected by districts – capture information beyond a teacher's effect on student test scores, while Petek and Pope (2016) and Kraft (2019) extend the analysis using detailed behavioral and psychological measurements. We follow this literature in choosing attendance (Gershenson, 2016; Liu and Loeb, 2021) and grades as behavioral outcomes that tenure policy does not directly target.

In addition to heterogeneity in contemporaneous outcomes, the persistence of a teacher's effect on a specific outcome may capture teacher heterogeneity relating to different modes of teaching. For instance, if a teacher focuses instruction on test-based memorization, the students may perform well on contemporaneous tests but not have the skills to build upon for future grades. We thus follow Carrell and West (2010) and Gilraine and Pope (2020) in measuring a teacher's impact on the future realizations of test scores, grades, and attendance.

We will use these student outcomes for two purposes: measuring how teacher output changes in response to incentives and classifying teachers into heterogeneous types. As this introduces several identification and estimation challenges, we specify a statistical model of student outcomes. We provide an overview here, with more detail in Appendix B. Let i index students, j index teachers, c index classrooms, t index years, and k index outcomes. Student i has a vector of outcomes y_{it} where the k^{th} element of the vector is $y_{i,t+\tau(k)}$. $\tau()$ is a function that describes when an outcome is realized. For contemporaneous outcomes, $\tau(k) = 0$, while for outcomes realized in the future, like next year's test scores, $\tau(k) > 0$. Without

loss of generality, we place contemporaneous test scores as the first outcome ($k = 1$). We standardize all outcomes to have mean 0 and standard deviation 1 for each grade-year. We model student outcomes as:

$$y_{i,t+\tau} = \Lambda X'_{it} + \sum_{e'} \rho_{e'} \mathbb{1}\{e_{jt} = e'\} + \mu_{jt} + \nu_{ct} + \phi_{c',t+1}\mathbb{1}(\tau \geq 1) + \phi_{c'',t+2}\mathbb{1}(\tau = 2) + \epsilon_{it} \quad (3)$$

where e_{jt} is a teacher's experience level, X'_{it} is a vector of P student covariates that may include lagged outcomes, and Λ is a $K \times P$ matrix of coefficients. ρ is a $K \times 1$ vector of experience effects. μ_{jt}, ν_{ct}, and ϵ_{it} are $K \times 1$ vectors of teacher-year effects, classroom effects, and student idiosyncratic variation, respectively. $\phi_{c,t+\tau}$ are combined $K \times 1$ classroom-teacher effects for students' assignments in years after t. We start with an independence assumption:

Assumption 1 (Independence of teacher, classroom, and student effects)**.**

$$\begin{aligned}
\mu_{jkt} &\perp (\nu_{clt}, \epsilon_{ilt}, \phi_{c',l,t+\tau})|X_{it}, e_{jt} \quad \forall k, l, \tau \\
\nu_{ckt} &\perp (\mu_{jlt}, \epsilon_{ilt}, \phi_{c',l,t+\tau})|X_{it}, e_{jt} \quad \forall k, l, c, c'\tau \\
\epsilon_{ikt} &\perp (\mu_{jlt}, \nu_{clt}, \phi_{c',l,t+\tau})|X_{it}, e_{jt} \quad \forall k, l, \tau \\
\phi_{c',k,t+\tau} &\perp (\mu_{jlt}, \nu_{clt}, \epsilon_{ilt})|X_{it}, e_{jt} \quad \forall k, l, \tau
\end{aligned}$$

This assumption corresponds to selection of students to teachers based on observables (X_{it}), but extended to a setting with multi-dimensional output. While we impose independence across effects that occur at different levels (e.g., teacher versus classroom), we allow a given effect to be correlated across outcomes. Prior work has validated this assumption for contemporaneous test scores using experimental variation (Kane and Staiger, 2008) or mover designs (Chetty et al., 2014; Jackson, 2018; Delgado, 2021; Gilraine and Pope, 2020). For our baseline analysis, we control for cubic functions of the $t-1$ outcome, with the exception of the subject-specific grade outcomes because often the lagged values are missing. In that case, we control for a cubic function of the $t-1$ test score.[12]

We estimate teacher j's realized causal effect on outcome k in year t, $\hat{\mu}_{jkt}$, by estimating Equation 3 with OLS for outcome k, constructing student-level residuals ($y_{it} - \hat{\Lambda} X'_{it}$), and taking the teacher-year-subject mean over the residuals. This procedure yields our (noisy) measure of a teacher's annual realized output on each dimension. Teacher effects may be

[12] Assumption 1 is sufficient for all of our analysis. In fact, for our estimation of the causal effect of the change in incentives on teacher output, we can relax the assumption provided that any sorting of students (or classroom effects) to teachers is orthogonal to the policy variation. In this case, we do not need to control for X_{it} for identification purposes, but we may still do so to increase our estimates' precision. In Section V we show no evidence of sorting changes in response to the policy and show that our main results are robust to excluding controls.

correlated across outcomes. A teacher who is effective at raising students' contemporaneous test scores may also be effective at raising students' future test scores. $\hat{\mu}_{jkt}$ will let us meet our first goal of measuring how a teacher's output changes in response to incentives.

Our second goal is to classify teachers into heterogeneous types based on their unincentivized output. This introduces a few challenges. Teachers' output is not completely in their control, as classroom shocks or idiosyncratic student shocks mean that some years teachers may have higher or lower output than would be predicted by their type and effort. Thus, we want to develop a forecast of a teacher's mean output, where the forecast uses observed outcomes but adjusts for the presence of shocks. The precision of this forecast will be very limited if we only use data from a single year of a teacher's career. Hence, unlike the estimate of teacher's realized output, which was year-specific, we now want to pool data across a teacher's career to develop a forecast of her mean output in the unincentivized state. Pooling, though, imposes structure across years that might seem inconsistent with a setting where teacher output is a product of a teacher's type *and* the environment. We thus pool data only for years in the unincentivized state under the assumption that the constant environment lets us assume a teacher's type is fixed in expectation across years.

We now add an assumption about the structure of drift:

Assumption 2 (Joint stationarity of teacher effects without incentives)**.** *The non-experience part of teacher value-added for each outcome follows a joint stationary process if there are no incentives. The covariances between the teacher's value-added across years depend only on the number of years elapsed:*

$$\mathbb{E}\left[\mu_{jkt}|t\right] = \mathbb{E}\left[\mu_{jks}|t\right] = 0 \tag{4}$$

$$Var\left(\mu_{jkt}\right) = \sigma^2_{\mu_k},\ Cov\left(\mu_{jkt}, \mu_{jk't}\right) = \sigma_{\mu_k\mu_{k'}} \tag{5}$$

$$Cov\left(\mu_{jkt}, \mu_{jk,t+s}\right) = \sigma_{\mu_k s} \tag{6}$$

$$Cov\left(\mu_{jkt}, \mu_{jk',t+s}\right) = \sigma_{\mu_k\mu_{k's}} \tag{7}$$

for all k, k', t and s.

We follow Mulhern and Opper (2021) in estimating the multi-year and multidimensional teacher value-added model; as Mulhern and Opper (2021) discuss, if we further assume that the error terms are jointly normal, these estimates are also the empirical Bayes' value-added measures and so we will generally refer to them as such. We estimate the model jointly across outcomes and using only data from unincentivized teacher-years. For each year, we construct a forecast of a teacher's value-added using data from other (unincentivized) periods only. We construct forecasts even for incentivized years; for these periods, the forecasts provide an estimate of the how the teacher would have affected her students'

outcomes had incentives not changed. The forecast comes from a multidimensional empirical Bayes procedure (Mulhern and Opper, 2021). We label the forecast $\tilde{\mu}_{jt}$.

These forecasts have several properties usually associated with current test score value-added. Within-teacher, the value-added on a given outcome is autocorrelated, which implies that the measures have some predictive power for adjacent years (Appendix Table A1). Second, for all measures, we see that teachers in their careers have rapid growth as they accrue experience (Appendix Figure A3). This experience profile is important because our empirical strategy will compare teachers at different experience points. We will thus need to control for changes in output we would expect from changes in experience, even in the absence of the policy. Further, the flattening out of the experience profile will allow us to pool teachers at high experience levels. Third, graduation outcomes are positively correlated with all of the measures, in a univariate sense (Appendix Table A2). We also see that value-added for each outcome is positively correlated with current test score value-added, though correlations are well below 1 (Appendix Table A1). This positive correlation means that policies that screen on one dimension are likely to positively select teachers on other dimensions. But as our model highlights, the correlation could reflect properties of the distribution of ability or the cost function and thus does not imply the form of teachers' behavioral responses.

III.C Combining Outcomes into an Index

Having multiple outcomes will allow us to estimate the effect of incentive changes on each outcome. We now combine untargeted outcomes into an index that assigns weights based on how differently the outcomes predict a student's long-term outcome. For example, an outcome that captures skills distinct from the other outcomes would receive a higher weight in the index. As the main distinction in the paper is whether an outcome is targeted by the tenure policy, we will keep current test score value-added as its own index and assess how the remaining outcomes co-vary.[13]

We create the untargeted index by anchoring the measures to their relative predictiveness of whether a student graduates from high school.[14] Let $Grad_{ijt}$ be whether student i graduated from high school, which we match to her teacher j in year t. We then regress

[13]Throughout, we will divide measures by whether they are ever targeted by the tenure policy ("targeted" vs. "untargeted") and periods by whether they come with the stronger tenure incentives ("incentivized" vs. "unincentivized").

[14]We also conduct a Principal Component Analysis. The first principal component loads heavily on grade measures while the second loads on measures of future outcomes (Appendix Table A4). We will also show that our main results are robust to summarizing untargeted measures using the first principal component.

graduation on measures of teacher j's value-added in year t:[15]

$$Grad_{ijt} = \omega' \tilde{\mu}_{jt} + \upsilon_{ij}, \tag{8}$$

where ω is a vector of anchoring weights.[16] We estimate using data from the unincentivized period and note that j's value-added in year t is estimated leaving out data from year t, so we avoid having the same classroom effects or student idiosyncratic shocks on both sides of the regression. We present the estimated weights in Appendix Table A3, which shows that most outcomes continue to predict graduation positively, with the largest coefficients on current attendance and future grades in untested subjects.[17] We use the estimated weights to construct two measures: targeted output ($\tilde{\mu}^T_{jt} = \tilde{\mu}_{j1t}$) and an index of untargeted output ($\tilde{\mu}^U_{jt} = \frac{1}{\hat{\omega}_1} \sum_{k=2}^{K} \hat{\omega}_k \tilde{\mu}_{jkt}$).[18] Because we divide the untargeted outcomes by $\hat{\omega}_1$, we measure both the targeted and index of untargeted output in units of (current) test score student standard deviations, which allows for comparisons in equivalent units that are common in the literature. We see that the index maintains the properties of the individual measures: high autocorrelation (Appendix Table A1), steep experience profile (Appendix Figure A3), and strong univariate predictor of graduation (Appendix Table A2).

IV Context and Empirical Strategy

IV.A Tenure Policy

In NYC, teachers become eligible for tenure after accumulating three years of teaching experience within the district. Once a teacher is eligible for tenure, the district may grant tenure, deny tenure, or extend the probationary (pre-tenure) period for further evaluation. Tenure denial makes the teacher ineligible to teach in the district while teachers who receive

[15] Instead of an indicator for whether a student graduated, we use a graduation residual that residualizes the indicator with a cubic function of i's test scores in year $t-1$ and other student observable characteristics to improve precision.

[16] We face a missing data challenge where some teachers only have effects on a subset of outcomes. For instance, because we use grade 3-8 test scores, we do not have future test score value-added measures for 8th-grade teachers. We follow Mulhern and Opper (2021) in forecasting missing measures based on the covariances between the missing and non-missing measures, though we will also show our main results are robust if we restrict the sample to teachers with no missing measures.

[17] This anchoring regression uses cross-sectional teacher variation. Teachers, of course, may differ in other ways such that the cross-sectional relationship does not predict treatment effects when individual outcomes change. For our main results, we also present estimates that do not aggregate outcomes using the estimated anchoring weights. These measure-specific results also highlight that the results are not driven by the negative estimated weights on subject grades, and we get similar results with other weighting schemes that ensure all measures get positive weights.

[18] We also apply these weights to the unshrunken measures for further indices, $\hat{\mu}^T_{jt} = \hat{\mu}_{jkt}$ and $\hat{\mu}^U_{jt} = \frac{1}{\hat{\omega}_1} \sum_{k=2}^{K} \hat{\omega}_k \hat{\mu}_{jkt}$.

tenure are provided extra employment protections for the rest of their careers.

Before the 2009-2010 school year, nearly all eligible teachers in NYC received tenure. For example, during the 2007 and 2008 school years,[19] 94% of all eligible teachers received tenure (Loeb et al., 2015). The tenure process, however, changed dramatically starting in 2009. In November, 2009, Mayor Michael Bloomberg announced at a panel discussion at the Center for American Progress that not using student achievement scores to evaluate teachers up for tenure was "like saying to hospitals, 'You can evaluate heart surgeons on any criteria you want - just not patient survival rates.'" He was therefore directing "our school Chancellor Joel Klein to ensure that principals actually use student achievement data to help evaluate teachers who are up for tenure this year."[20,21] From that point forward, NYCDoE would begin to consider how effective teachers are at increasing their students' test scores when deciding whether to give them tenure. This change, which mirrors many similar policies in other urban districts, proved controversial, as teachers argued decisions would incorporate unreliable measures and would affect the workplace environment negatively (McGuinn, 2012; Murphy et al., 2013; Bleiberg et al., 2021).

Perhaps due to the rapid implementation schedule and the contentious reaction among teachers, the details in how test score value-added affected tenure changed over time. For the 2010 school year, the district automatically coded a teacher as having "tenure in doubt" ("tenure likely") if the 95% confidence interval of her value-added scores over the previous two years fell below (above) the median. Teachers whose confidence intervals included the median received no recommendation based on value-added.[22] Although this coding only informed a final decision, the burden shifted onto the principal to argue why she was making a recommendation contrary to what the coding suggested.

In subsequent years, a low value-added score no longer automatically coded a teacher as having her tenure in doubt. Yet teacher value-added scores continued to be a major focus of the tenure process. In 2011, teachers with low value-added scores were flagged as having an "Area of Concern" and those with high value-added scores were flagged as having "Notable Performance." In 2013 and on, value-added scores remained part of the tenure evaluation process, but the district provided no explicit guidance on their use.

[19]To simplify notation, we refer to each school year as the calendar year it ended; e.g., the 2009-2010 school year will be called the 2010 school year.

[20]Bloomberg's full remarks from this talk are available at: https://www.c-span.org/video/?290247-1/white-house-education-agenda-state-us-schools

[21]Bloomberg made this announcement despite the fact that the New York State Legislature had banned the use of value-added in tenure decisions the year before, because in his words, "after a very close reading our lawyers tell use that the current law... only applies to teachers hired after July 1, 2008." In addition, because the law expired by the time teachers hired after July 1, 2008 were up for tenure, it never had any affect on the tenure-granting process in NYCDoE.

[22]This policy relies on the cardinality of test score scales. See Barlevy and Neal (2012) for a discussion of incentive provision and screening with ordinal scales.

Mayor Bloomberg's announcement signaled a major shift in the teacher tenure policy in NYC. These changes had two first-order effects: they lowered tenure rates and they made tenure decisions increasingly dependent on value-added scores. Indeed, we observe a significant decrease in the probability that a teacher received tenure following the reform. We plot the fraction of newly tenure-eligible (fourth year) teachers receiving tenure over time in Figure 2. We see that tenure rates fell precipitously from 97% in 2010 to 64% in 2012.[23] When teachers did not receive tenure, they could either have tenure denied or have their probationary (pre-tenure) period extended. Loeb et al. (2015), who have access to the specific tenure decisions, show that most non-tenured cases led to extensions of the probationary period. But these extensions were not merely delays leading to the same outcome, as most of the extended teachers left their schools or even the district.[24] Thus, while the policy did not lead to a large increase in tenure denial rates, the policy still dramatically decreased the fraction of teachers continuing in the district with tenure.

Tenure rates remained fairly flat through 2010, the first year following Bloomberg's announcement, while the substantial decrease in tenure rates came a year later in 2011. This introduces an important question of how to measure the policy's timing. While the prior analysis shows how tenure outcomes changed over time, the relevant policy timing impact is when it first affected teachers' *incentives* to achieve higher value-added scores. The policy's announcement – at the start of the 2010 school year – marked the point when teachers became aware that value-added scores would matter for future tenure decisions. We further provide evidence of the increased focus on tenure during the 2010 school year by examining the mentions of "tenure" on the teachers' union (UFT) website; in Figure 3 we show a spike in mentions in 2010, the school year of the announcement. Thus, for our analysis we will treat the 2010 school year as the first under the new policy regime.

Beyond affecting aggregate tenure rates, the policy also tied tenure more closely to test score value-added measures. Using data and value-added measurement methods described in Section III, we assess the relationship between tenure and test score value-added before and after the policy change. In Figure 4, we bin teachers into ten deciles according to their value-added scores during their third year of experience. The y-axis is the fraction of teachers in each bin who have achieved tenure by the end of their fourth year. We plot the relationships separately for the periods before and after the change in the tenure process. Prior to the policy change, we see very little relationship between a teacher's value-added score and her probability of receiving tenure. This is not surprising as the high tenure rates left little variation to explain. After the policy change, however, value-added scores become

[23] We do not have access to specific tenure decisions in our data; instead, for each year we observe whether the teacher has tenure.

[24] In Appendix Figure A1, we show that after the policy, teachers who would be entering their fourth year were much less likely to continue teaching in NYC.

strong predictors of tenure outcomes.[25]

We further explore the tenure rules and how they vary with teachers' test score value-added and other measures, by estimating linear tenure rules separately for each policy regime. We estimate the following screening functions:

$$Tenure_{jt} = \pi_0^0 + \pi_1^0 \tilde{\mu}_{jt}^T + \pi_2^0 \tilde{\mu}_{jt}^U + \upsilon_{jt}^0 \tag{9}$$

$$Tenure_{jt} = \pi_0^1 + \pi_1^1 \tilde{\mu}_{jt}^T + \pi_2^1 \tilde{\mu}_{jt}^U + \upsilon_{jt}^1, \tag{10}$$

where superscript 0 is for pre-reform (2008-2009) tenure decisions and superscript 1 indicates post-reform (2011-2012) tenure decisions. $Tenure_{jt}$ is an indicator for whether the teacher receives tenure in the first year she is eligible to. We present the screening function coefficients in Table 2. In the pre-reform period, the mean tenure rate is high (97%), and neither test score value-added nor the index of other outcomes enters statistically significantly. The coefficients on these measures are also quite small, indicating precise zero estimates. In the second column, we show the estimated screening function in the post-reform period. As expected, the mean tenure rate falls (to 67%) while test score value-added is now a strong, and statistically significant, predictor of receiving on-time tenure. Interestingly, output on the untargeted measures does not enter the screening function significantly. This confirms that teachers' tenure-based incentives around non-test score value-added output did not change due to the policy.

IV.B Concurrent Events

Although the change in tenure policy provides nice within-teacher variation in their incentives, the policy did not occur in a vacuum. For example, about two years before the change in tenure policy (in Fall 2007), student test score growth began to affect the grade each school received (Rockoff and Turner, 2010). Around the same time, NYC established a pilot program that distributed information about teacher value-added to 112 principals (Rockoff et al., 2012). Within the next year, in Fall of 2008, these value-added measures were distributed to every teacher in NYC, although they were told that "they won't be used in tenure determinations or the annual rating process."[26] As discussed above, this decision was reversed in the following year by Mayor Bloomberg. It was not until 2012, however, that the teacher data reports were made public.

[25]While the policy rules might imply a nonlinear relationship, they apply to confidence intervals that account for sampling error differently than value-added scores do.

[26]This quote comes from the New York Times article on the policy titled, "Teachers to Be Measured Based on Students' Standardized Test Scores," which was written on October 1, 2008 and cites the quote as coming from a memo written by Chancellor Klein.

At the same time as the policy was being announced and implemented, NYC was feeling the beginnings of the Great Recession, which likely affected the types of teachers who were entering the teaching market (Nagler et al., 2015). Furthermore, the Great Recession caused a large financial shortfall for the NYC government, leading the Chancellor to institute a hiring freeze of new teachers. While there were some exceptions, the fraction of first-year teachers in NYC plunged from around 10% of total teachers to a mere 2% (Appendix Figure A2).

IV.C Empirical Strategy

Our first goal is to estimate the behavioral response of pre-tenured teachers to the added incentive. While posing potential threats to identification, the concurrent events described above also suggest the types of comparisons that might isolate a causal effect. Unlike the school report card and value-added information dissemination policies, the tenure reform only affected a portion of the teachers: the untenured teachers. In short, the change in the tenure process caused a sudden increase in untenured teachers' incentives to increase their value-added scores relative to their tenured colleagues, which suggests the use of a difference-in-difference estimator. Consider teacher j with outcome y_{jt} in school year t. A simple difference-in-difference specification would then be:

$$y_{jt} = \tau UntenuredPost_{jt} + \nu Untenured_{jt} + \eta_t + \epsilon_{jt}, \tag{11}$$

where $UntenuredPost_{jt}$ is an indicator for j being untenured in year t and the tenure policy change being in place. This specification thus compares how untenured and tenured teachers' outcomes differentially changed once the tenure policy was announced.

While straightforward, this specification is not sufficient since the composition of these two groups is changing over time for several reasons. First, reduced hiring of new teachers during the Great Recession may have led recent new teacher cohorts to differ from prior cohorts. If more selective hiring led to more productive incoming teachers, then we might infer that the tenure policy improved untenured teachers' output when it was simply selection. Second, the Great Recession or the tenure incentives themselves may have affected which NYC public school teachers choose to remain teaching in the district. Third, since the tenure reform directly affects teacher tenure rates, tenure status is endogenous with respect to the policy. In the extreme, if low-output teachers have delayed tenure decisions but keep teaching in NYC, then the composition of the group of untenured teachers would decline in quality over time.

We deal with these challenges by making three adjustments to the base difference-in-difference specification. First, we address the endogeneous selection into receiving tenure

by making a sample restriction that will apply to all of our specifications. Namely, we exclude teachers once they have become tenure eligible – that is, they have accumulated at least 3 prior years of experience – under the new tenure policy. These excluded observations correspond to periods when some of these teachers may have tenure while others may still be in their probationary period. If we had left these teachers in the analysis, then comparisons across (current) tenure status would likely involve a large degree of selection.[27] We will therefore only use tenured teachers who received tenure prior to the policy change, when tenure rates were nearly 100%, such that both the untenured and tenured teachers in our analysis will have been subject to minimal involuntary attrition.

Second, we assign each teacher j to a cohort $m(j)$ based on the year j started teaching in NYC and include cohort fixed effects in the empirical specification. If the Great Recession changes selection among new teachers based on productivity levels, then the cohort fixed effects should control for any cross-cohort differences upon entry. Thus, instead of looking across-teachers, we will use the fact that a few cohorts of teachers were teaching in NYC (in the probationary period) when the policy was announced. They therefore spent their initial year(s) teaching without test score value-added incentives, then were suddenly told that their tenure would depend on their test score value-added. We will thus explore how they responded to this sudden change, on both the targeted and untargeted tasks.

Adding cohort fixed effects, though, complicates the untenured versus tenured comparison, as within-cohort changes may differ between these groups for reasons unrelated to the policy itself. In particular, a large literature on the shape of the teaching experience profile has documented the possibility that early-career teachers have a steeper profile (Rockoff, 2004; Rice, 2013) and it is precisely the early-career teachers who are untenured.[28] Hence, as we add cohort fixed effects, we also introduce a vector of fixed effects for each level of prior experience in NYC; we combine teachers with six or more years of prior experience. This second adjustment implies that we will identify the policy effects based on how output growth rates change differentially between untenured and tenured teachers based on the policy implementation, beyond what we would expect from generic experience effects.

We summarize how these research design choices affect identification in Table 3. We show teacher cohorts - based on when they entered NYC – in the rows and the academic

[27] Alternatively, we could have left these observations in the analysis and instrumented for tenure incentive with whether the teacher was in the *standard* probationary period – i.e., fewer than 3 years of prior experience. We choose to restrict the sample because we worry about monotonicity of the proposed instrument. When a cohort of teachers advances from 2 to 3 years of prior experience and the instrument switches values, those teachers who did receive tenure see a large decrease in their incentives while those teachers whose probationary periods were extended have even higher incentives than before.

[28] Wiswall (2013) and Papay and Kraft (2015) show that using different identifying restrictions can generate a linear experience profile. Our framework could accommodate imposing a linear experience profile for later-career teachers.

years in the columns. In the intersection of a row and column, we show the cohort's incentive status, which is whether the cohort's teachers are in the probationary period *and* NYC has already incorporated value-added into the tenure decision process. Because we include cohort fixed effects, our identification will come from cohorts (rows) for whom the incentive status varies across columns (i.e., the 2008 and 2009 cohorts).

We translate this variation into the following empirical specification:

$$y_{jt} = \tau Incentive_{jt} + \lambda_e + \nu_m + \eta_t + \epsilon_{jt}, \tag{12}$$

where $e = e(j,t)$ is j's level of prior experience in year t and $Incentive_{jt}$ is an indicator for j being in the first 3 years of teaching and the tenure policy change being in place. Our parameter of interest is τ, which is the coefficient of our covariate that indicates whether teacher j had incentives in time t. We cluster our standard errors by teacher and find similar standard errors when we cluster by school. We will also show permutation tests where we estimate a distribution of placebo effects by permuting the exposed cohorts (Abadie et al., 2010; Idoux, 2021).

This strategy does not deal with the other compositional concerns though. If the tenure policy or Great Recession changes attrition patterns, our comparisons will still be confounded. We will therefore run analyses that include teacher fixed effects instead of cohort fixed effects:

$$y_{jt} = \tau Incentive_{jt} + \lambda_e + \nu_j + \eta_t + \epsilon_{jt}. \tag{13}$$

Any selective within-cohort attrition will then be controlled for, provided it is based on time-invariant differences across teachers (rather than, say, growth rates).

To summarize, our empirical strategy relies on the assumption that after accounting for yearly shocks and selection into and out of teaching based on time-invariant unobservables, the returns to experience for teachers in the probationary period would not have changed post-2009 in the absence of the policy change. We will attribute any systematic deviation from the "unincentivized" experience profile to the policy change.

IV.D Selective Attrition

Before we turn to the effects on output, we return to the possibility of selective voluntary attrition. Indeed, attrition may be a response to the policy itself, as teachers unlikely to get tenure may attrit early, once the policy is announced. While the specification with teacher fixed-effects controls for some forms of attrition, we here explore attrition and its relationship with our instrument.

In Figure 5, we plot survival curves for cohorts grouped by their exposure to the treatment. The solid blue line, for instance, shows the survival curve for teachers in never

exposed cohorts – i.e., those who had already completed the standard probationary period before the policy change. At the other extreme, the dashed orange line shows cohorts always exposed – i.e., those who entered the district after the policy's announcement. We show unincentivized years with dots and incentivized years with diamonds. The survival curves largely lie on top of each other, indicating no differential attrition based on policy exposure.

Even with similar cohort-level attrition rates, we could still have selective attrition if the composition of the attriters changes in response to the policy. We again split cohorts based on policy exposure and plot the difference in mean test score value-added between stayers (teachers who stay in the district) and leavers (Figure 6). Again, we see no discernible relationship. For instance, in the third year of teaching, the blue (never exposed) and green (exposed for two years) lines lie on top of each other, despite the green line representing a cohort that had the incentive treatment in its third year of teaching. This lack of a systematic pattern of attrition thus makes us less worried that selective attrition explains our results.

V Estimated Effects on Teacher Output

V.A Targeted Measure

We start by showing the effect of the tenure policy on the targeted measure: current test score value-added.

We show the empirical strategy visually in Figure 7, where we group teaching cohorts based on their policy exposure (i.e., never exposed, exposed for one year, etc.). We then plot each composite cohort's mean test score value-added (after netting out year and experience effects) during the first three years of their probationary period. We see that in cohorts' first years in NYC, they differ significantly in terms of their output.[29] We focus on how output evolves over time, *within-cohort*. We see that for one year of prior teaching experience, the green line (which represents the cohorts exposed for 2 years) transitions from not being treated with tenure incentives to being treated. This is also the cohort that has the largest gain between years 0 and 1. Then for output after two years of prior teaching experience, we see that the cohort newly treated in this year (the red line) jumps up to "join" the other treated lines. The fact that the deviations from the unincentivized pattern occur in the specific years the cohorts receive the incentive treatment increases our confidence that our empirical strategy is picking up response to the policy.

[29] Our goal is not a full policy evaluation that considers every impact of the tenure policy. But the different intercepts hint at possible effects of the policy on selection into the district. The always exposed cohorts, represented by the orange line, have higher initial test score value-added than the other composite cohorts, which could indicate that the policy induced better selection of teachers into the district. But as highlighted above, the differential selection may also be driven by other factors like the Great Recession.

We complement the graphical evidence with estimation of Equations 12 and 13, where we use current test score value-added as the outcome. We present the estimates in Table 4. In column (1), we show that that – consistent with Figure 7 – the specification that includes cohort fixed effects suggests that teachers responded in meaningful ways to the change in incentive. Specifically, it suggests that the tenure incentive increased value-added by 0.033 student standard deviations (σ) and is statistically significant at the 1% level. While this specification includes cohort fixed effects to control for changes in selection into teaching over time, the estimated effect may still reflect some compositional changes from attrition. For example, if inexperienced teachers who are unsure whether they would like to make a career out of teaching are more likely to attrit during the Great Recession, and these teachers tend to have lower value-added, then we might incorrectly attribute the output gains to response to the policy's incentive changes. The visual results of Section IV.D suggest this may not be a first-order issue, but we test this by replacing cohort fixed effects with teacher fixed effects in our preferred specification. As shown in column (2) of Table 4, our estimate of the policy effect is slightly smaller but still large: 0.019σ. It remains statistically significant at the 5% level and we cannot reject that the estimate is the same as the estimate with cohort fixed effects (p-value = 0.19).[30] In Appendix Figure A4, we conduct a permutation test and show our estimates are more extreme than all but 1 placebo estimate (column 1) or than all placebo estimates (column 2).[31]

These estimates are economically significant. The response to incentives ($0.019-0.033\sigma$) is equivalent to 13-23% of the cross-sectional standard deviation of forecasted teacher value-added (0.142).

We explore heterogeneity in treatment effects based on subject tested and grade level in Appendix Table A5. We find that the effect is slightly larger in elementary grades and similar across subjects. But we lack precision to either reject equal effects by level or subject or to rule out meaningful differences.

Thus, the policy had large incentive effects, which increased teachers' immediate output in the targeted measure. But because teachers produce output in multiple dimensions, multitasking may have led teachers to substitute out of effort developing untargeted skills.

V.B Untargeted Measures

We test for the policy's effects on untargeted measures with the same specifications (Equations 12 and 13), but varying the type of teacher output. We start by using our index of

[30] These estimates are similar to the 0.023σ effect Taylor (2022) finds in response to a Tennessee tenure reform.

[31] We also conduct randomization inference. Because randomization tests require an approximate symmetry assumption (Canay et al., 2017), we drop two post-policy years from our data. The test yields a t-statistic of 2.61.

untargeted measures, which we specified in Section III, and present the results in the last two columns of Table 4. Unlike output in the targeted measure, the index of untargeted measures does not increase in response to the tenure policy change. Instead, as shown in columns (3) and (4), we find statistically significant decreases. In our preferred specification with teacher fixed effects, we estimate that output on the untargeted index falls by the equivalent of 0.064 student test score standard deviations. In Appendix Figure A4, we conduct a permutation test and show our estimates are more extreme than all placebo estimates.[32]

The negative effect is surprising if one naively extrapolates from the positive cross-sectional relationship between measures and the positive effect on test score value-added to predict the policy response on the other measures. But as we developed in our model, this substitution away from untargeted measures is consistent with teachers multitasking and distorting effort toward tasks that affect targeted measures. Because we normalize the untargeted measure index to be in test score value-added units (based on their respective predictiveness of graduation rates), we can compare the magnitude of the estimates across the targeted and untargeted measures. We estimate that the decrease in the untargeted measures exceeds the increase in the targeted measures such that students' graduation rates may fall. This findings reflects the concern among some policy-makers that "teaching to the test" may lower the quality of instruction students receive.[33] Our theoretical contribution, however, implies a second conclusion: given the substitution out of the untargeted tasks, the heterogeneous response will improve screening efficiency. We test this prediction in Section VII.

We further decompose this effect to show how each of our untargeted measures changes in response to the policy. We present the teacher fixed effects results in Table 5.[34] We examine the effects on future test scores (in $t+1$ and $t+2$), current and future attendance, current and future grades, and current and future grades in untested subjects. Consistent with the effect on the index, we find statistically significant decreases in $t+1$ test scores and $t+1$ grades in both tested and untested subjects (and a marginally significant decrease in $t+1$ attendance). These effects are reasonably large. In standard deviation units, these estimates dominate the effect on current test scores. We find a statistically significant increase in current grades in the tested subject, arguably the measure most closely related to the targeted one. For the other outcomes, we fail to reject no change.

Though our untargeted measures do not necessarily map neatly into lower-dimensional representations of skills, we might expect that future test scores are more likely to capture

[32]Randomization inference yields an absolute t-statistic of 1.76.

[33]This result demonstrates the downside of pay for performance on one dimension, as all policy effects occur during the period with the incentive.

[34]Appendix Table A6 shows the estimates from a specification with cohort fixed effects.

development of cognitive skills while attendance and grades are more likely to capture development of behavioral or non-cognitive skills (Jackson, 2018). We see similar effects across measures in these two groups. Instead, we see a clearer division between the effects on current versus future measures. Thus, teachers may be substituting toward tasks with short-run payoffs but that fail to build (cognitive or non-cognitive) skills that persist.[35]

One worry about the untargeted outcomes is that, unlike the targeted outcome, lagged outcomes may be unavailable or poor predictors of current or future outcomes. If lagged outcomes are necessary controls for estimating causal effects, then we could be picking up effects that are not causal. As we mentioned in Section IV, our identification strategy (for this section) does not require causal estimates of teacher effects but rather that students do not change sorting to teachers systematically in response to the policy. We test this directly by replacing the outcome in Equation 13 with a predetermined student characteristic. We show the estimates in Table 6. We find precise zero effects of the incentive on student sorting to teachers based on several observable characteristics. Assuming no change in student sorting, we can estimate the effects of the incentive on the (unresidualized) untargeted outcomes. We present the results in Appendix Table A7 and find very similar patterns to our results using residualized outcomes.

A second worry is that future outcomes depend on the following teacher's actions. For our analysis, we might incorrectly attribute changes in, say, value-added on $t+1$ test scores, as reflecting the current teacher's response to incentives rather than the subsequent teacher's response. In Appendix Tables A8 and A9, we show that controlling for the treatment status of the subsequent teachers hardly changes our estimates. In Appendix Table A10, we show that our results are robust to controlling for the identity of the subsequent teachers.

We also explore robustness of our results to different ways of constructing the index of untargeted measures. In Appendix Table A11, we run our analysis using the first principal component of the untargeted measures rather than an index anchored to graduation rates. In the first two columns we find a negative effect of the incentive on the first principal component. The other columns show robustness for the results from Sections V.C and VI. In Appendix Table A12 we restrict our sample to teachers with no missing measures in the index. Standard errors increase due to the smaller sample, but the point estimates are similar to our analysis using the baseline index.

[35] We find the strongest effects for future grades. The Principal Component Analysis shows these measures entering both the first and second components with large weights, which suggests substitution is not along the lines that principal components capture.

V.C Persistence

In Section VII, we will look at the effects of these behavioral responses on predicted output in the post-probationary period. Such effects could operate through two channels: changes to screening and changes to teacher output (even once tenure incentives no longer matter). We will focus on the former in Section VII, but now we examine the latter. Changes to teacher output could persist if teachers respond to the incentives by making investments that change their future production functions. For instance, a teacher might develop new lesson plans and continue to use them after receiving tenure. We test whether the responses to incentives persist, even once the incentives disappear, with the following specification (for teacher fixed effects):

$$y_{jt} = \tau Incentive_{jt} + \phi PostIncentive_{jt} + \lambda_e + \nu_j + \eta_t + \epsilon_{jt}. \tag{14}$$

$PostIncentive_{jt}$ is an indicator for whether the teacher has tenure but faced the new tenure policy at some point during her probationary period. In this analysis, we drop our sample restriction to allow teachers to be in the analysis for both the (incentivized) probationary period and the post-probationary period. We present the estimates in Table 7. For the targeted measure, we estimate that teachers revert back to their pre-incentive levels (excluding the experience profile). For the untargeted measures we estimate a coefficient in the same direction as the incentive effect, though we fail to reject complete reversion. Thus, we do not find strong evidence that teaches' responses to incentives had persistent effects by changing their future teaching output.[36] Given the larger standard error on the post-incentive untargeted output, we will explore how our screening results change if teachers do not revert to their unincentivized output.

VI Heterogeneous Responses

We now extend our empirical model to allow for heterogeneous responses across teachers based on differences in their output in the unincentivized regime. This is a relevant source of heterogeneity for two reasons. First, because tenure is absorbing, once teachers have been screened for tenure, their output incentives disappear. As we saw in Section V, teachers' post-tenure output reverts to their unincentivized output (once we have removed the experience gradient). Thus, the teachers' unincentivized output arguably provides the best measure of how valuable a teacher will be to the district after she receives tenure. If the teachers with the highest unincentivized output are the ones most likely to respond to the

[36]This result matches Ng (2021)'s finding that after receiving tenure, in a policy regime where teacher value-added matters for tenure decisions, teachers' math value-added falls.

incentive, then the behavioral response increases screening efficiency.

Second, such heterogeneity maps neatly into our theory model developed in Section II. The model predicts differential responses based on the teacher's unincentivized output x^*.[37] Our results in Section V imply a cost function with a positive cross-partial derivative, which (in the large incentive case) further predicts the form of the differential response: conditional on x_1^*, teachers with a higher x_2^* will increase their targeted output in the incentivized state more than those with lower x_2^* (see Figure 1 for the visual comparison). Hence, we can test the model prediction by classifying teachers based on their unincentivized output and estimating heterogeneous responses to the policy.

To classify teachers, we use the fact that the multi-year value-added model from Section III yields forecasts for each teacher-year-measure. Specifically, $\tilde{\mu}_{jt}^T$ is the forecast for teacher j's test score value-added (absent the common experience profile) in year t and $\tilde{\mu}_{jt}^U$ is the forecast for an index of the untargeted measures. We start by examining heterogeneous responses based on $\tilde{\mu}_{jt}^T$ or $\tilde{\mu}_{jt}^U$ separately before classifying heterogeneous responses based on their joint distribution.

As in Section V, we first summarize the data graphically, in Figure 8. Here, we fix a cohort and year and regress a teacher's unshrunken test score value-added (i.e, $\hat{\mu}_{jt}^T$) on $\tilde{\mu}_{jt}^U$ and plot the coefficients. In the first year of teaching, this coefficient is close to 1.2 for all three cohorts. Note that for all of the cohorts plotted, the first year teaching comes prior to the tenure reform, so this coefficient captures the positive cross-sectional relationship between a teacher's effect on current test scores and other output that exists where there are no additional incentives. We then show how the cross-sectional relationship changes as experience accrues and as incentives change due to the tenure reform. Focusing on the cohort never exposed to the reform, we see the cross-sectional relationship holds steady. For the other cohorts, in contrast, we see immediate (and persistent) increases in the regression coefficient once the reform is implemented. The sudden change in the joint distribution indicates that the response to the reform's incentives was stronger among teachers with higher output on the untargeted measures (in the unincentivized period).

To show the results in regression form, we modify our main estimating equations (Equations 12-13) to allow for heterogeneous impacts (in the incentivized period) based on output

[37] Our model does not have a stochastic component to output whereas observed output does. The predictions therefore pertain to forecasted output – i.e., the predictable part of output that is attributable to the teacher.

forecasts from the unincentivized period. The modified teacher fixed effects specification is:

$$\begin{aligned}\hat{\mu}_{jt}^{T} = \tau Incentive_{jt} + \xi_1 \tilde{\mu}_{jt}^{T} Incentive_{jt} + \xi_2 \tilde{\mu}_{jt}^{U} Incentive_{jt} \\ + \sum_{e'} \pi_1^{e'} \mathbb{1}\{e_{jt} = e'\} \tilde{\mu}_{jt}^{T} + \sum_{e'} \pi_2^{e'} \mathbb{1}\{e_{jt} = e'\} \tilde{\mu}_{jt}^{U} \\ + \lambda_e + \nu_j + \eta_t + \epsilon_{jt}.\end{aligned} \tag{15}$$

We control for the predictiveness of $\tilde{\mu}_{jt}^{T}$ and $\tilde{\mu}_{jt}^{U}$, without incentives, flexibly by letting it vary by experience level. Because we are interested in how teachers respond to the incentive by shifting into the targeted measure, our outcome is the teacher's mean (current) test score residual ($\hat{\mu}_{jt}^{T}$).

We present the estimates in Table 8. In columns (1) and (2), which differ based on the level of fixed effects, we focus only on heterogeneity based on $\tilde{\mu}_{jt}^{T}$ (i.e., imposing $\xi_2 = 0$ and $\pi_2^{e'} = 0 \ \forall e'$). We see some evidence that teachers with higher forecasted value-added in the targeted measure respond more strongly to the tenure policy incentives, though we lack statistical precision to make confident statements. In columns (3) and (4), we compare teachers' responses based on $\tilde{\mu}_{jt}^{U}$ and find that higher forecasted value-added in the index of untargeted measures predicts a larger increase in current test scores. This is consistent with Figure 8.

Columns (5) and (6) include both forms of heterogeneity jointly. As in the one-dimensional heterogeneity analysis, we see that teachers with higher forecasts of current test score value-added may respond more than teachers with lower forecasts, but we cannot statistically reject no differential response. We also see that the response to the incentive is stronger for teachers with higher forecasts of the untargeted measures ($\tilde{\mu}_{jt}^{U}$), and we can reject equal response at the 5% level. In Appendix Figure A5, we conduct a permutation test and show this estimate is more extreme than all placebo estimates. While the estimated coefficient on $\tilde{\mu}_{jt}^{U}$ is smaller than the estimated coefficient on $\tilde{\mu}_{jt}^{T}$, cross-sectional differences in $\tilde{\mu}_{jt}^{U}$ are over 3 times as large as cross-sectional differences in $\tilde{\mu}_{jt}^{T}$. Thus, a one standard deviation difference in $\tilde{\mu}_{jt}^{T}$ translates to a similar estimated response as a one standard deviation difference in $\tilde{\mu}_{jt}^{U}$.

This result is important for policy and consistent with our model. The behavioral response to the screening-induced incentives leads teachers to substitute into the targeted measure. We see higher degrees of substitution among teachers who are better (without incentives) on the untargeted measure. The screening policy thus leads to a multitasking problem that lowers output on untargeted measures in the pre-tenure period but increases output on untargeted measures in the post-tenure period by selecting different teachers.

VII Estimated Effects on Screening Efficiency

We now quantify the improvement in screening efficiency. For each teacher j, we require 4 objects for our calculation: (a) forecasted targeted output in the probationary period, with incentives, x_{j1}^{w}; (b) forecasted targeted output in the probationary period, without incentives, x_{j1}^{wo}; (c) forecasted output in the post-probationary period, x_{j1}^{p} and x_{j2}^{p}. We then define a tenure screening policy that keeps the top $p\%$ of teachers according to their output on the targeted dimension. Letting $r()$ be a function that converts teacher's output on the targeted measure to a percentile ranking, we calculate the expected post-tenure output (for output dimension d) with behavioral responses as:

$$\mathbb{E}(x_d|w) = \mathbb{E}(x_{jd}^{p}|r(x_{j1}^{w}) > p) \tag{16}$$

and the expected post-tenure output without behavioral responses as:

$$\mathbb{E}(x_d|wo) = \mathbb{E}(x_{jd}^{p}|r(x_{j1}^{wo}) > p). \tag{17}$$

The difference in these expectations is the impact of the behavioral response on the composition of tenured teachers. Note that the incentives only matter through the selection margin because if teachers are granted tenure, the amount they produce does not depend on how much they responded to the probationary period incentives.

We use our prior analysis to estimate the 4 necessary objects. We use our multi-year value-added model from Section III to forecast x_{j1}^{wo}:

$$x_{j1}^{wo} = \tilde{\mu}_{jt}^{T}. \tag{18}$$

This forecast relies on data only from the unincentivized periods. We further impose that post-probationary output also matches this forecast (for all dimensions):

$$x_{j1}^{p} = \tilde{\mu}_{jt}^{T} \qquad \text{and} \qquad x_{j2}^{p} = \tilde{\mu}_{jt}^{U}. \tag{19}$$

This assumption is motivated by our analysis showing that once they receive tenure, teachers revert to their pre-incentives level of output, though we will also report results if teachers do not revert.[38]

Finally, we use the estimated coefficients from Equation 15, to relate targeted output

[38]For the analysis, we consider the screening for a fixed cohort, based on forecasted output in a specific year, t.

with and without a behavioral response:

$$x_{j1}^{w} = x_{j1}^{wo} + \hat{\tau} + \hat{\xi_1}\tilde{\mu}_{jt}^{T} + \hat{\xi_2}\tilde{\mu}_{jt}^{U} \tag{20}$$

The key assumption is that we have summarized the heterogeneity in treatment effects with our specification.[39] The behavioral response will only change the composition of tenured teachers if $\hat{\xi_2} \neq 0$. Otherwise, any behavioral response might increase short-run output but does not change the ordering of teachers.[40]

We focus our analysis on the cohort of teachers that entered NYC in 2007 (and were first eligible for tenure in 2010) and apply a 67% tenure rate, which is similar to the tenure rate in NYC post-tenure reform.[41] In Figure 9, we show the joint distribution of x_{j1}^{wo} and x_{j1}^{w}. In Section VI we showed that conditional on $\tilde{\mu}_{jt}^{U}$ we saw a limited differential response along the targeted dimension. But because $\tilde{\mu}_{jt}^{U}$ covaries with $\tilde{\mu}_{jt}^{T}$, the joint distribution of targeted output with and without the behavioral response can rotate. We see indeed see a slight rotation of the relationship between x_{j1}^{wo} and x_{j1}^{w}. The distribution of test score value-added becomes more dispersed without much change in teachers' relative positions.

To more directly understand how the behavioral response affects screening along the two dimensions, in Figure 10 we order teachers based on their forecasted two-dimensional output in the post-probationary period along the x- and y-axes. This figure is the empirical counterpart to Figure 1. We divide each teacher into one of four categories and label the teacher's point on the figures accordingly. The categories indicate both whether the teacher receives tenure when there is no behavioral response and whether the teacher receives tenure when there is a behavioral response. The solid and dashed lines show the tenure cutoffs in the two regimes, under a hypothetical policy in which 67% of teachers receive tenure. In the top panel, we zoom in on teachers with test score value-added between -0.1σ and 0.1σ – i.e., those close to the tenure cutoffs – while in the bottom panel we zoom out and show all teachers.

When there is no behavioral response, tenure depends solely on one's unincentivized test score value-added, so we see a flat line at the cutoff. When there is a behavioral response,

[39] Our economic object of interest is how the treatment effect heterogeneity relates to a teacher's (estimated) type. While for some analyses, any other forms of treatment effect heterogeneity that are orthogonal to a teacher's (estimated) type would not affect the conclusions, here we plug these predictions into a non-linear screening function.

[40] The requirement that $\hat{\xi_2} \neq 0$ to change selection is specific to the tenure policy that keeps a fixed fraction of teachers. Compositional changes would still occur under counterfactual policies that apply an absolute threshold for receiving tenure, even with $\hat{\xi_2} = 0$. Furthermore, we focus on a threshold policy with a deterministic rule. If, instead, the policy had some stochastic component, then $\hat{\xi_1} > 0$ would imply increased dispersion in test score value-added, which in turn would increase the signal districts have in its tenure decisions. This could be an additional screening benefit to explore in future work.

[41] Appendix Table A14 shows the results for the 2006 cohort.

in contrast, we see that the tenure cutoff line rotates clockwise due to the fact that those with higher unincentivized value-added on the untargeted measures (the x-axis) respond more to the policy. Thus, as we predicted in the model (Figure 1), teachers with high levels of unincentivized value-added on the untargeted measures may still receive tenure despite low levels of unincentivized test score value-added. The triangles show the teachers who receive tenure only when there is no behavioral response and the circles show the teachers who receive tenure only when there is a behavioral response. These groups are relatively small, comprising of only 4.3% of all teachers. We only see a limited number of changed tenure decisions because the tenure policy employs a threshold rule such that many behavioral responses are among teachers who are ex post inframarginal. But despite a very concentrated change in the composition of tenured teachers, these teachers vary dramatically in their $\tilde{\mu}^U_{jt}$. The *mean* difference in $\tilde{\mu}^U_{jt}$ across teachers shifted into tenure versus out of tenure is 1.73σ, or 239% of the cross-sectional standard deviation in $\tilde{\mu}^U_{jt}$. The differences in mean $\tilde{\mu}^T_{jt}$ are much smaller: 0.024σ, or 18% of the cross-sectional standard deviation.

We quantify the changes in screening from the behavioral response ($\mathbb{E}(x_d|w) - \mathbb{E}(x_d|wo)$) and present the results in Table 9. We see that mean x^p_{j1} falls by just 0.001σ (0.76% of the cross-sectional standard deviation) while mean x^p_{j2} increases by 0.056σ (7.7% of the cross-sectional standard deviation).

Teachers' behavioral responses to the policy thus change the composition of the tenured employees. Does the district prefer the new composition? This depends on how the district values output in the targeted measure versus output in the untargeted measures. The conditions on the district's value function in our theoretical model implied that the district prefers the new composition. But without imposing the theoretical assumptions, we can make quantitative statements based on our estimates. In particular, because we anchored the measures to their predictiveness of graduation rates, they are measured in comparable units. Hence, we see that the district's screening has become more efficient in graduation units, as the improved selection on the untargeted measures exceeds the worsened selection on the targeted measures. If the district has reasons to value the outputs beyond their predictiveness of graduation rates, we estimate that the district prefers the new composition unless it values output in the targeted measure to the untargeted measure at a rate more than 56 times their predictiveness of graduation.

If we return to the graduation units, we can add the two measures up as "mean total output," which we show for the different policies in the last column of Table 9. We see that the behavioral response causes the tenured teachers' predicted total output to increase by 0.055σ. We can also compare our results to the gains the district would achieve if it could (infeasibly) observe all forms of output and assign tenure according to the sum. We show

the associated output of the tenured teachers under this policy in the second-to-last row of Table 9. We find, predictably, that total output is highest in this infeasible policy. Thus, the feasible policies do not achieve the first-best. If we define the "screening efficiency gap" as the difference in mean total output between an infeasible policy that screens on both dimensions and the ex post screening policy that screens on test score value-added without the behavioral response, we estimate that this screening efficiency gap is 0.198σ. We find that the behavioral response to the policy closes 28% of the screening efficiency gap.[42]

This improvement in the screening efficiency comes at the cost of distorting probationary period output. A final cost-benefit analysis thus compares the endogenous response's effects during the probationary period with the effects on screening efficiency post-tenure. On the targeted measure there is a large increase in probationary period output and relatively small decrease in post-tenure output, while on the untargeted measures there is a large increase in the post-tenure output and a relatively small decrease in the probationary period output. On both measures the benefits are larger than the costs, which suggests that the endogenous response improved overall efficiency. We can quantify this by noting that teachers would need to stay in the profession for 45 years for the (negative) post-tenure effects to overtake the (positive) 3-year probationary period effects on the targeted output, but the (positive) post-tenure effect overtakes the (negative) probationary period effect after 3.5 years for the untargeted output. Thus, provided tenured teachers remain in the district for between 3.5 and 45 years, the endogenous response leads to gains in both targeted and untargeted output.

VIII Conclusion

It is a longstanding concern that evaluating individuals or institutions on a single measure will cause that measure to lose meaning, a worry social sciences often refer to as Campbell's Law (Campbell, 1979).[43] Nowhere has this been more discussed than in the education setting and, in particular, using test scores to evaluate students, teachers, or schools. In developing Campbell's Law, for example, Donald Campbell used test scores as an example, writing that: "Achievement tests may well be valuable indicators of general school achievement under conditions of normal teaching... but when test scores become the goal of the

[42]Even in the pessimistic scenario where teachers continue producing less untargeted output in the post-probationary period ($x^p_{j2} = x^{wo}_{j1} + \hat{\pi}$, where $\hat{\pi}$ is the estimate of the incentive's effect on untargeted output in the probationary period), we still find that the behavioral response reduces the screening efficiency gap by 9.4%.

[43]A similar idea is referred to as Goodhart's Law, which states that "when a measure becomes a target, it ceases to be a good measure." Goodhart's Law, like the Lucas Critique, initially referred to macroeconomic models and monetary policy, but have since been applied to a number of different contexts.

teaching process they lose their value of indicators of educational status."[44] In this paper, we illustrate both theoretically and empirically that the opposite can also be true – that evaluating individuals or institutions on a single measures may instead make the measure more informative, rather than less, of the individual or institution's underlying ability.

[44]Holmstrom and Milgrom (1991) also used test scores were also used as the motivating example of multitasking.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller, "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 2010, *105* (490), 493–505.

Abaluck, Jason, Mauricio Caceres Bravo, Peter Hull, and Amanda Starc, "Mortality Effects and Choice across Private Health Insurance Plans," *The Quarterly Journal of Economics*, 2021, *136* (3), 1557–1610.

Abdulkadiroğlu, Atila, Parag A Pathak, Jonathan Schellenberg, and Christopher R Walters, "Do Parents Value School Effectiveness?," *American Economic Review*, 2020, *110* (5), 1502–39.

Armstrong, Mark and Jean-Charles Rochet, "Multi-Dimensional Screening:: A User's Guide," *European Economic Review*, 1999, *43* (4-6), 959–979.

Baker, George P, "Incentive Contracts and Performance Measurement," *Journal of Political Economy*, 1992, *100* (3), 598–614.

Barlevy, Gadi and Derek Neal, "Pay for percentile," *American Economic Review*, 2012, *102* (5), 1805–31.

_ **and** _ , "Allocating Effort and Talent in Professional Labor Markets," *Journal of Labor Economics*, 2019, *37* (1), 187–246.

Bénabou, Roland and Jean Tirole, "Bonus Culture: Competitive Pay, Screening, and Multitasking," *Journal of Political Economy*, 2016, *124* (2), 305–370.

Bertrand, Marianne and Antoinette Schoar, "Managing with Style: The Effect of Managers on Firm Policies," *The Quarterly journal of economics*, 2003, *118* (4), 1169–1208.

Björkegren, Daniel, Joshua E Blumenstock, and Samsun Knight, "Manipulation-proof machine learning," *arXiv preprint arXiv:2004.03865*, 2020.

Bleiberg, Joshua, Eric Brunner, Erica Harbatkin, Matthew A. Kraft, and Matthew G. Springer, "The Effect of Teacher Evaluation on Achievement and Attainment: Evidence from Statewide Reforms," Technical Report, Working Paper 2021.

Brown, Christina and Tahir Andrabi, "Inducing Positive Sorting through Performance Pay: Experimental Evidence from Pakistani Schools," *University of California at Berkeley Working Paper*, 2021.

Campbell, Donald T, "Assessing the impact of planned social change," *Evaluation and program planning*, 1979, *2* (1), 67–90.

Canay, Ivan A, Joseph P Romano, and Azeem M Shaikh, "Randomization tests under an approximate symmetry assumption," *Econometrica*, 2017, *85* (3), 1013–1030.

Carrell, Scott E and James E West, "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors," *Journal of Political Economy*, 2010, *118* (3), 409–432.

Chan, David C, Matthew Gentzkow, and Chuan Yu, "Selection with Variation in Diagnostic Skill: Evidence from Radiologists," *The Quarterly Journal of Economics*, 2022, *137* (2), 729–783.

Chandra, Amitabh, Amy Finkelstein, Adam Sacarny, and Chad Syverson, "Health Care Exceptionalism? Performance and Allocation in the US Health Care Sector," *American Economic Review*, 2016, *106* (8), 2110–44.

Chen, Zhao and Sang-Ho Lee, "Incentives in Academic Tenure under Asymmetric Information," *Economic Modelling*, 2009, *26* (2), 300–308.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 2014, *104* (9), 2593–2632.

Corcoran, Sean P, Jennifer L Jennings, and Andrew A Beveridge, "Teacher Effectiveness on High-and Low-Stakes Tests.," *Society for Research on Educational Effectiveness*, 2011.

Dee, Thomas and James Wyckoff, "Incentives, Selection, and Teacher Performance: Evidence from IMPACT," *Journal of Policy Analysis and Management*, Spring 2015, *34* (2), 267–297.

Delgado, William, "Heterogeneous Teacher Effects, Comparative Advantage, and Match Quality," 2021.

Demougin, Dominique and Aloysius Siow, "Careers in Ongoing Hierarchies," *The American Economic Review*, 1994, pp. 1261–1277.

Deshpande, Manasi and Yue Li, "Who Is Screened out? Application Costs and the Targeting of Disability Programs," *American Economic Journal: Economic Policy*, 2019, *11* (4), 213–48.

DeVaro, Jed and Oliver Gürtler, "Strategic shirking: a theoretical analysis of multitasking and specialization," *International Economic Review*, 2016, *57* (2), 507–532.

_ **and** _ , "Strategic shirking in promotion tournaments," *The Journal of Law, Economics, and Organization*, 2016, *32* (3), 620–651.

_ **and** _ , "Strategic shirking in competitive labor markets: A general model of multi-task promotion tournaments with employer learning," *Journal of Economics & Management Strategy*, 2020, *29* (2), 335–376.

Duflo, Esther, Pascaline Dupas, and Michael Kremer, "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya," *American economic review*, 2011, *101* (5), 1739–74.

Frankel, Alex and Navin Kartik, "Muddled Information," *Journal of Political Economy*, 2019, *127* (4), 1739–1776.

Fryer, Roland G and Richard T Holden, "Multitasking, Learning, and Incentives: A Cautionary Tale," 2012.

Gershenson, Seth, "Linking Teacher Quality, Student Attendance, and Student Achievement," *Education Finance and Policy*, 2016.

Gilraine, Michael and Nolan G. Pope, "Making Teaching Last: Long- and Short-Run Value-Added," *Working Paper*, 2020.

Glewwe, Paul, Nauman Ilias, and Michael Kremer, "Teacher Incentives," *American Economic Journal: Applied Economics*, 2010, *2* (3), 205–27.

Holmstrom, Bengt and Paul Milgrom, "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership and Job Design," *Journal of Law, Economics and Organization*, 1991, *7*, 24–52.

Idoux, Clemence, "Integrating New York City Schools: The Role of Admission Criteria and Family Preferences," 2021.

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

Jacob, Brian A., "Acountability, Incentives and behavior: The Impact of High-Stakes Testing in the Chicago Public Schools," *Journal of Public Economics*, 2005, *89*, 761–796.

Jacob, Brian A, "The effect of employment protection on teacher effort," *Journal of Labor Economics*, 2013, *31* (4), 727–761.

Jacob, Brian A. and Lars Lefgren, "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluations in Education.," *Journal of Labor Economics*, 2008, *26* (1), 101–136.

Kane, Thomas J. and Douglas O. Staiger, "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," *NBER*, 2008, (14607).

_ , Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger, *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment*, Seattle, WA: Bill and Melinda Gates Foundation, 2013.

Kou, Zonglai and Min Zhou, "Multi-tasking vs. Screening: A Model of Academic Tenure," *CCES, Fudan University Working Paper*, 2009.

Kraft, Matthew A, "Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies," *Journal of Human Resources*, 2019, *54* (1), 1–36.

_ , Eric J Brunner, Shaun M Dougherty, and David J Schwegman, "Teacher Accountability Reforms and the Supply and Quality of New Teachers," *Journal of Public Economics*, 2020, *188*, 104212.

Liu, Jing and Susanna Loeb, "Engaging Teachers Measuring the Impact of Teachers on Student Attendance in Secondary School," *Journal of Human Resources*, 2021, *56* (2), 343–379.

Loeb, Susanna, Luke C. Miller, and James Wyckoff, "Performance Screens for School Improvement: The Case of Teacher Tenure Reform in New York City," *Educational Researcher*, 2015, *44* (4).

Macartney, Hugh, Robert McMillan, and Uros Petronijevic, "Teacher Value-Added and Economic Agency," Technical Report, National Bureau of Economic Research 2018.

McGuinn, Patrick, "Stimulating Reform: Race to the Top, Competitive Grants and the Obama Education Agenda," *Educational Policy*, 2012, *26* (1), 136–159.

Mulhern, Christine and Isaac Opper, "Measuring and Summarizing the Multiple Dimensions of Teacher Effectiveness," 2021.

Muralidharan, Karthik and Venkatesh Sundararaman, "Teacher Performance Pay: Experimental Evidence from India," *Journal of Political Economy*, 2011, *119* (1), 39–77.

Murphy, Joseph, Philip Hallinger, and Ronald H Heck, "Leading via Teacher Evaluation: The Case of the Missing Clothes?," *Educational Researcher*, 2013, *42* (6), 349–354.

Nagler, Markus, Marc Piopiunik, and Martin R. West, "Weak Markets, Strong Teachers: Recession at Career Start and Teacher Effectiveness," *NBER Working Paper*, 2015.

Neal, Derek, "Designing Incentive Systems for Educators" forthcoming in Performance Incentives: Their Growing Impact on American K-12 Education, edited by Matthew Springer, Brookings," 2009.

_ , "Aiming for efficiency rather than proficiency," *Journal of Economic Perspectives*, 2010, *24* (3), 119–32.

_ , "The Design of Performance Pay in Education," in "Handbook of the Economics of Education," Vol. 4, Elsevier, 2011, pp. 495–550.

_ **and Diane Whitmore Schanzenbach**, "Left behind by design: Proficiency counts and test-based accountability," *The Review of Economics and Statistics*, 2010, *92* (2), 263–283.

Ng, Kevin, "The Effects of Teacher Tenure on Productivity and Selection," 2021.

Nichols, Albert L and Richard J Zeckhauser, "Targeting Transfers through Restrictions on Recipients," *The American Economic Review*, 1982, *72* (2), 372–377.

O'Flaherty, Brendan and Aloysius Siow, "On the Job Screening, Up or Out Rules, and Firm Growth," *Canadian Journal of Economics*, 1992, pp. 346–368.

Papay, John P and Matthew A Kraft, "Productivity Returns to Experience in the Teacher Labor Market: Methodological Challenges and New Evidence on Long-Term Career Improvement," *Journal of Public Economics*, 2015, *130*, 105–119.

Petek, Nathan and Nolan G. Pope, "The Multidimensional Impact of Teachers on Students," *Working Paper*, 2016.

Philippis, Marta De, "Multi-task agents and incentives: The case of teaching and research for university professors," *The Economic Journal*, 2021, *131* (636), 1643–1681.

Rice, Jennifer King, "Learning from Experience? Evidence on the Impact and Distribution of Teacher Experience and the Implications for Teacher Policy," *Education Finance and Policy*, 2013, *8* (3), 332–348.

Rockoff, Jonah and Lesley J Turner, "Short-run Impacts of Accountability on School Quality," *American Economic Journal: Economic Policy*, 2010, *2* (4), 119–47.

Rockoff, Jonah E, "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *The American Economic Review*, 2004, *94* (2), 247–252.

_ , Brian A Jacob, Thomas J Kane, and Douglas O Staiger, "Can you recognize an effective teacher when you recruit one?," *Education finance and Policy*, 2011, *6* (1), 43–74.

Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane, and Eric S. Taylor, "Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools," *American Economic Review*, 2012, *102* (7), 3184–3213.

Spence, Michael, "Job Market Signaling," *The Quarterly Journal of Economics*, 1973, *87* (3), 355–374.

Staiger, Douglas O. and Jonah E. Rockoff, "Searching for Effective Teachers with Imperfect Information," *Journal of Economic Perspectives*, Summer 2010, *24* (3), 97–118.

Staiger, Douglas O and Jonah E Rockoff, "Searching for effective teachers with imperfect information," *Journal of Economic perspectives*, 2010, *24* (3), 97–118.

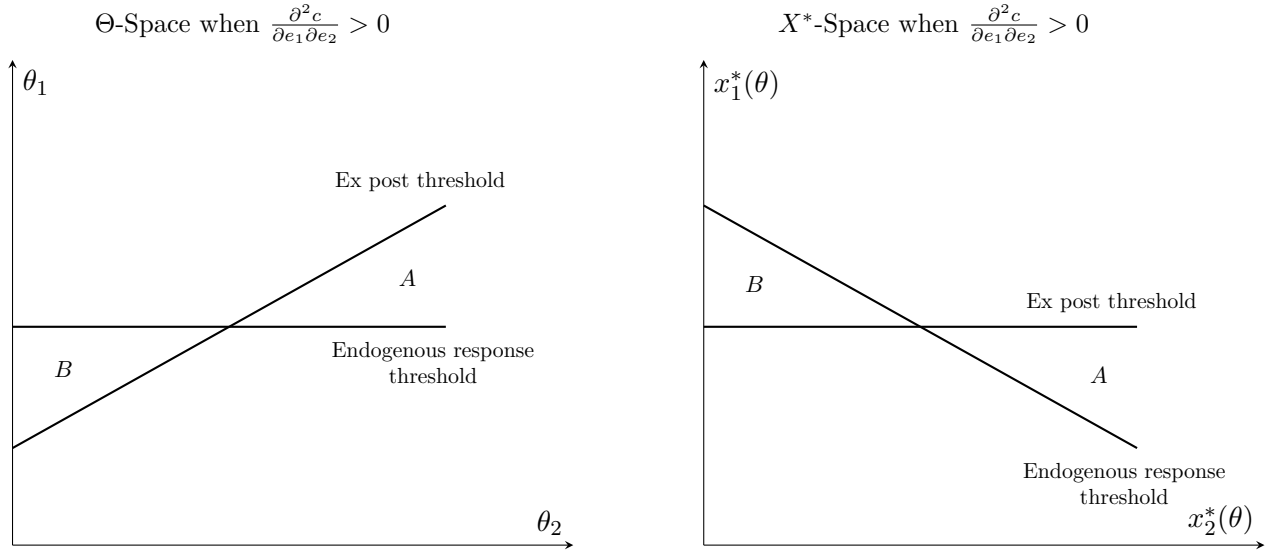
Taylor, Eric S, "Employee evaluation and skill investments: Evidence from public school teachers," Technical Report, National Bureau of Economic Research 2022.

Wiswall, Matthew, "The Dynamics of Teacher Quality," *Journal of Public Economics*, 2013, *100*, 61–78.

IX Tables and Figures

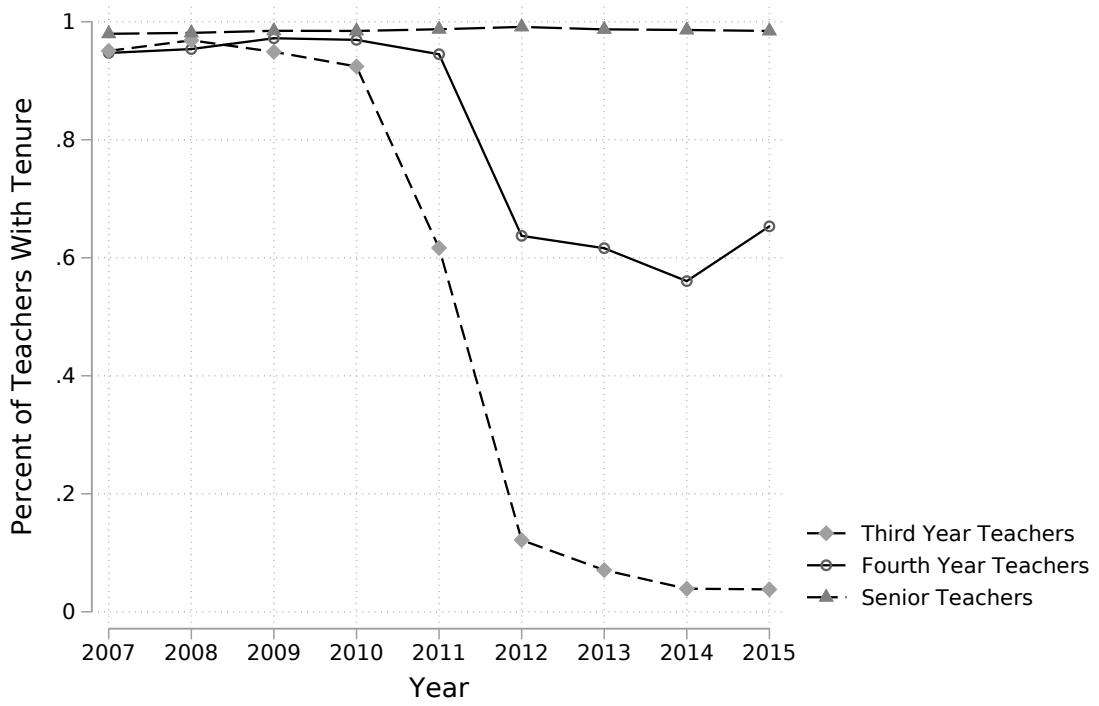
IX.A Figures

Figure 1: Screening in Type and Output Space



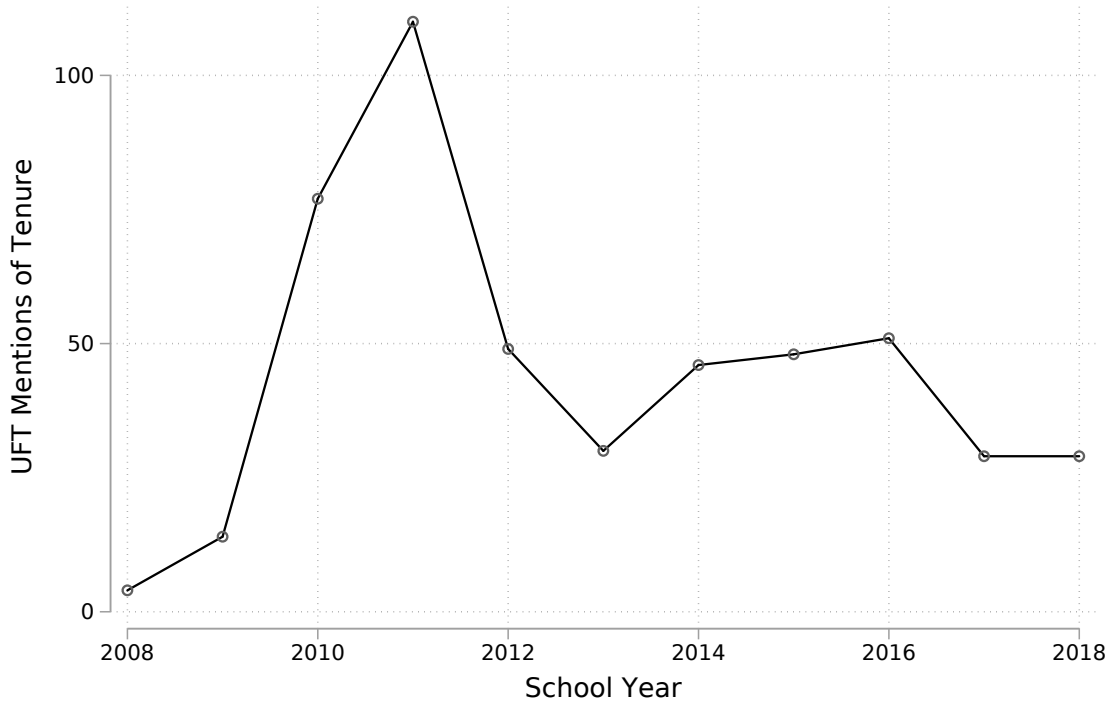
Note: The figure shows our model predictions about which teachers receive tenure in the case where the effort on the first task increases the effort cost on the second task. Teachers are classified in type (left) or unincentivized output (right) space. Teachers above the "Ex post threshold" receive tenure with a surprise policy that they do not endogenously respond to. Teachers above the "Endogenous response threshold" receive tenure with an announced policy that they do endogenously respond to. Teachers in region "A" only receive tenure with the endogenous response while teachers in region "B" only receive tenure without the endogenous response. The endogenous response's effect on screening efficiency then depends on whether the district prefers the teachers in region "A" to the teachers in region "B."

Figure 2: Changes in Tenure Rates



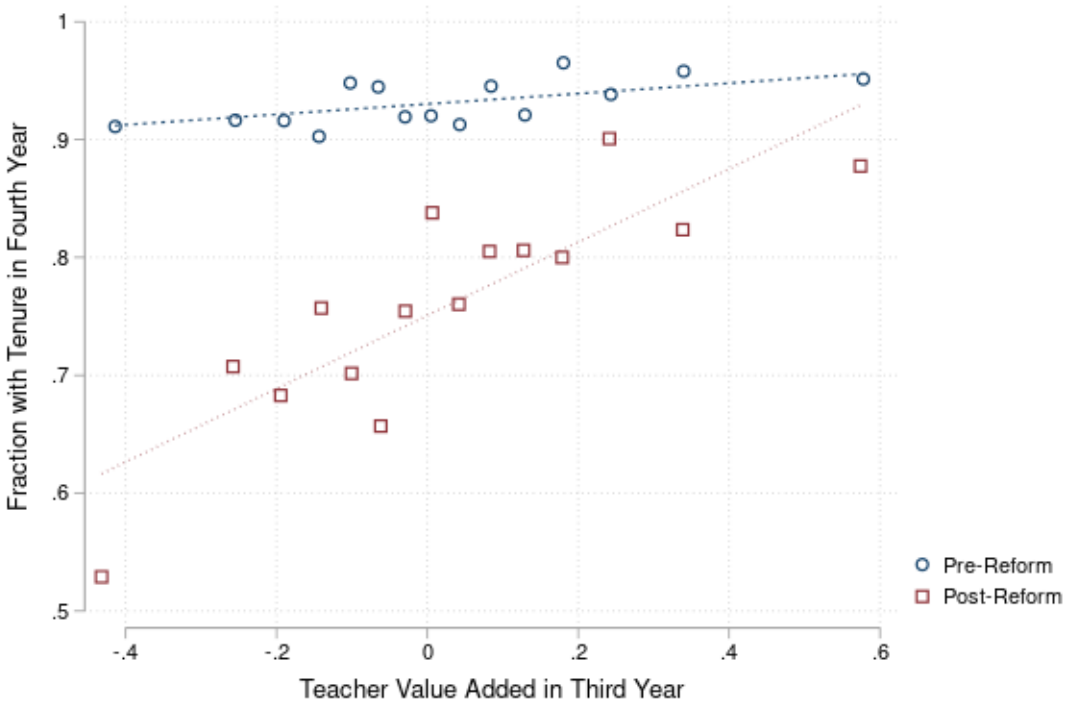
Note: The figure shows the fraction of teachers with tenure over time, split by years of experience. Teachers with fewer than three prior years of experience are typically not yet tenure eligible while teachers in their fourth year or later are tenure eligible. "Senior Teachers" have six or more prior years of experience.

Figure 3: Union Website Mentions of Tenure



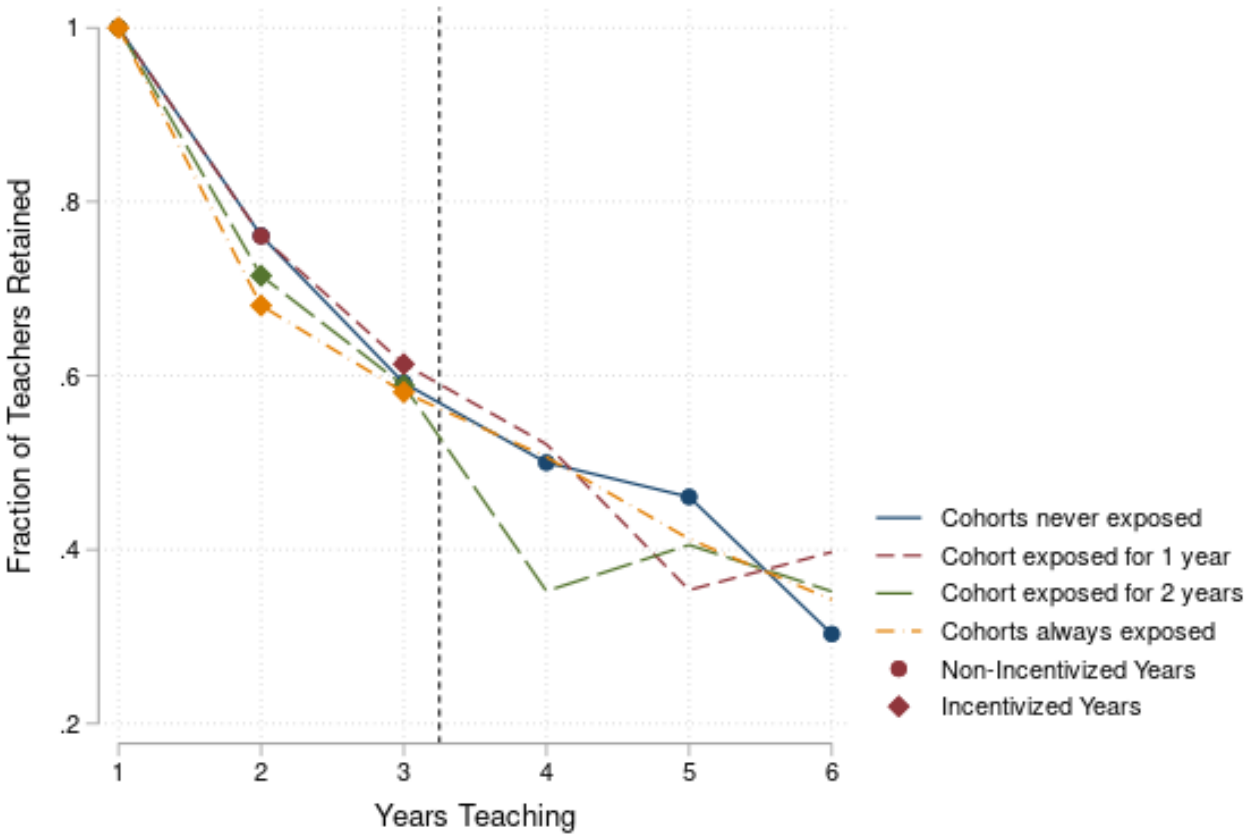
Note: The figure shows the number of mentions of "tenure" on the teachers union (UFT) website, over time. The policy was announced in the 2010 school year.

Figure 4: Tenure Probability by Value-Added



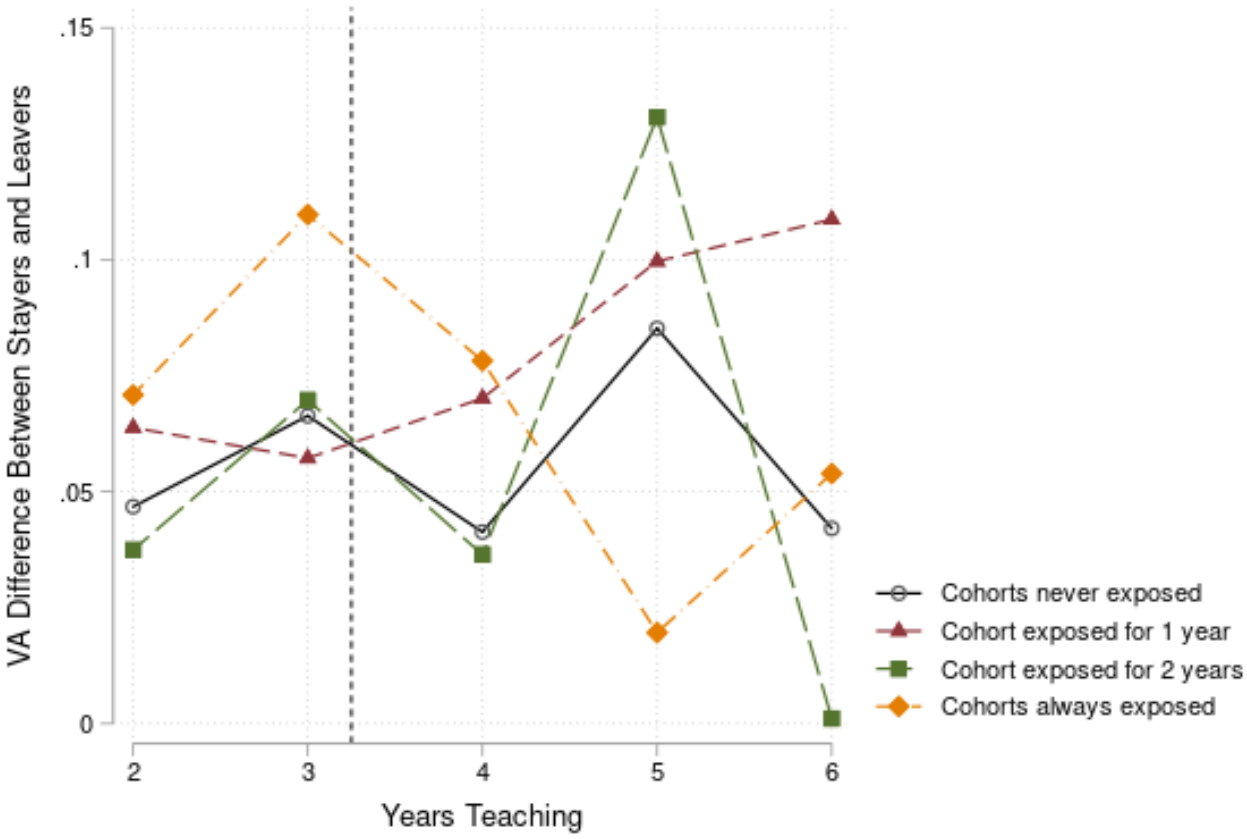
Note: This figure is a binscatter that groups teachers into twenty ventiles according to their (unshunken) test score value-added scores during their third year of experience. The y-axis is the fraction of teachers in each bin who have received tenure by the end of that year (entering their fourth year). We plot the relationships separately for the periods before and after the changes in the tenure process.

Figure 5: Survival Curves



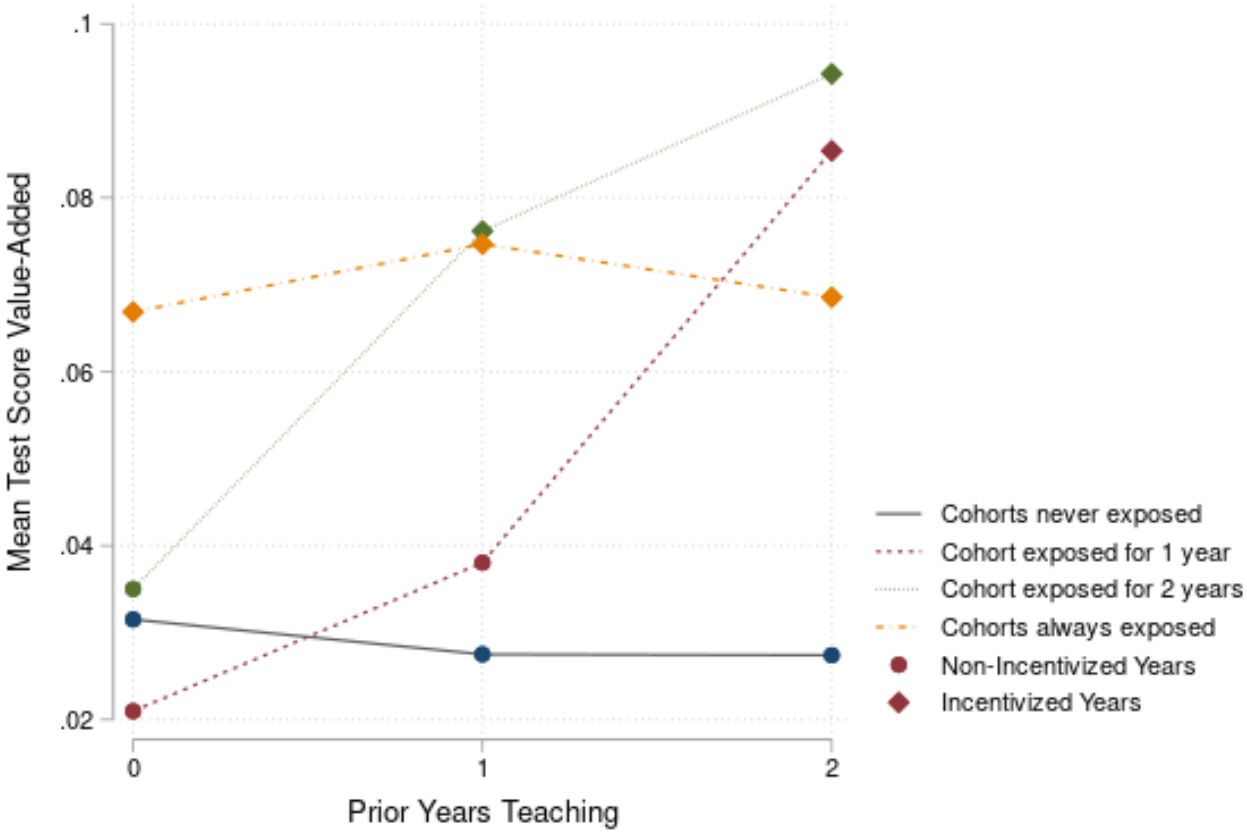
Note: This figure shows the fraction of teachers who stay teaching in the district, by years of experience. For instance, about 60% of teachers reach at least 3 years of experience. We plot separate lines for teacher cohorts based on the number of years they were exposed to the new tenure policy. Diamonds designate years in the probationary period under the new tenure policy.

Figure 6: Attrition Related to Test Score Value-Added



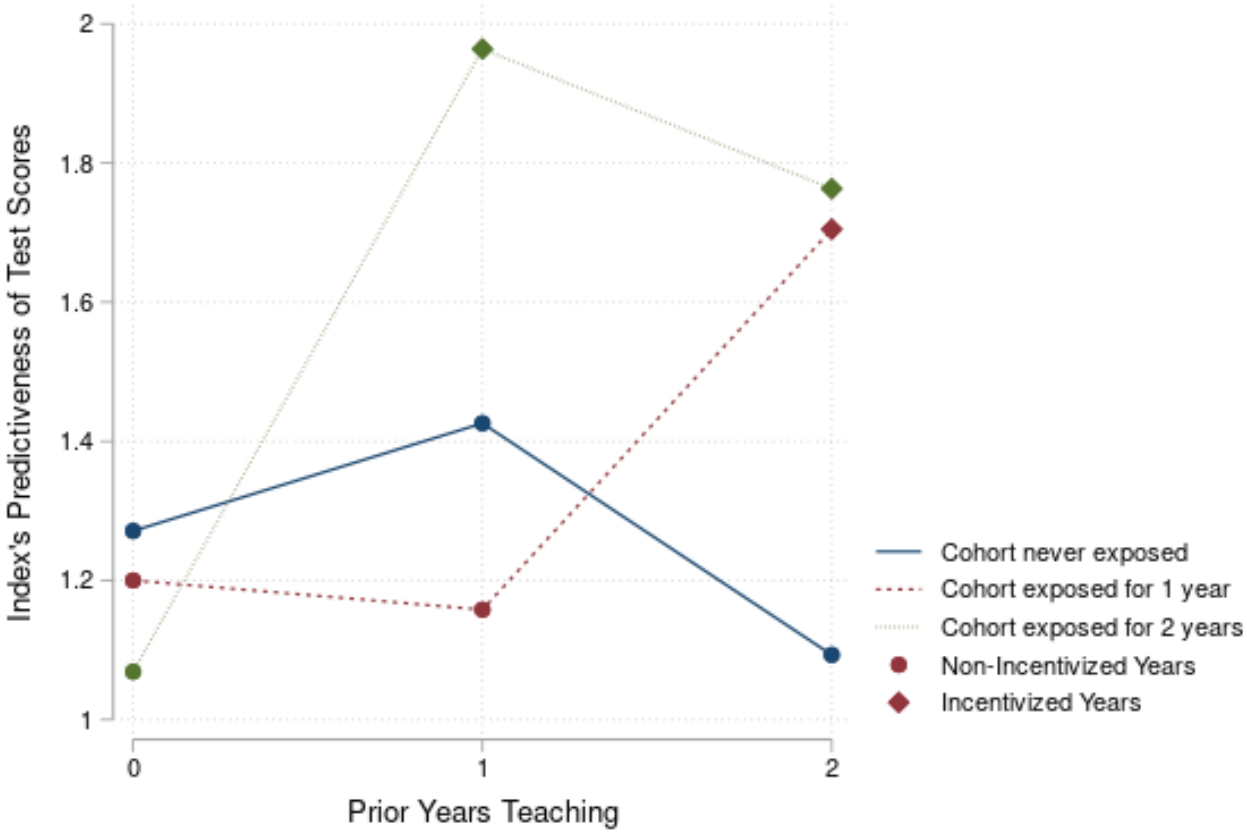
Note: This figure shows composition differences between teachers who stay for the following year and teachers who leave the district, split by years of experience. Each point represents the mean test score value-added difference between stayers and leavers at a specific experience level. The vertical line shows the end of the standard probationary period. We plot separate lines for teacher cohorts based on the number of years they were exposed to the new tenure policy. Diamonds designate years in the probationary period under the new tenure policy.

Figure 7: Change in Test Score Value-Added



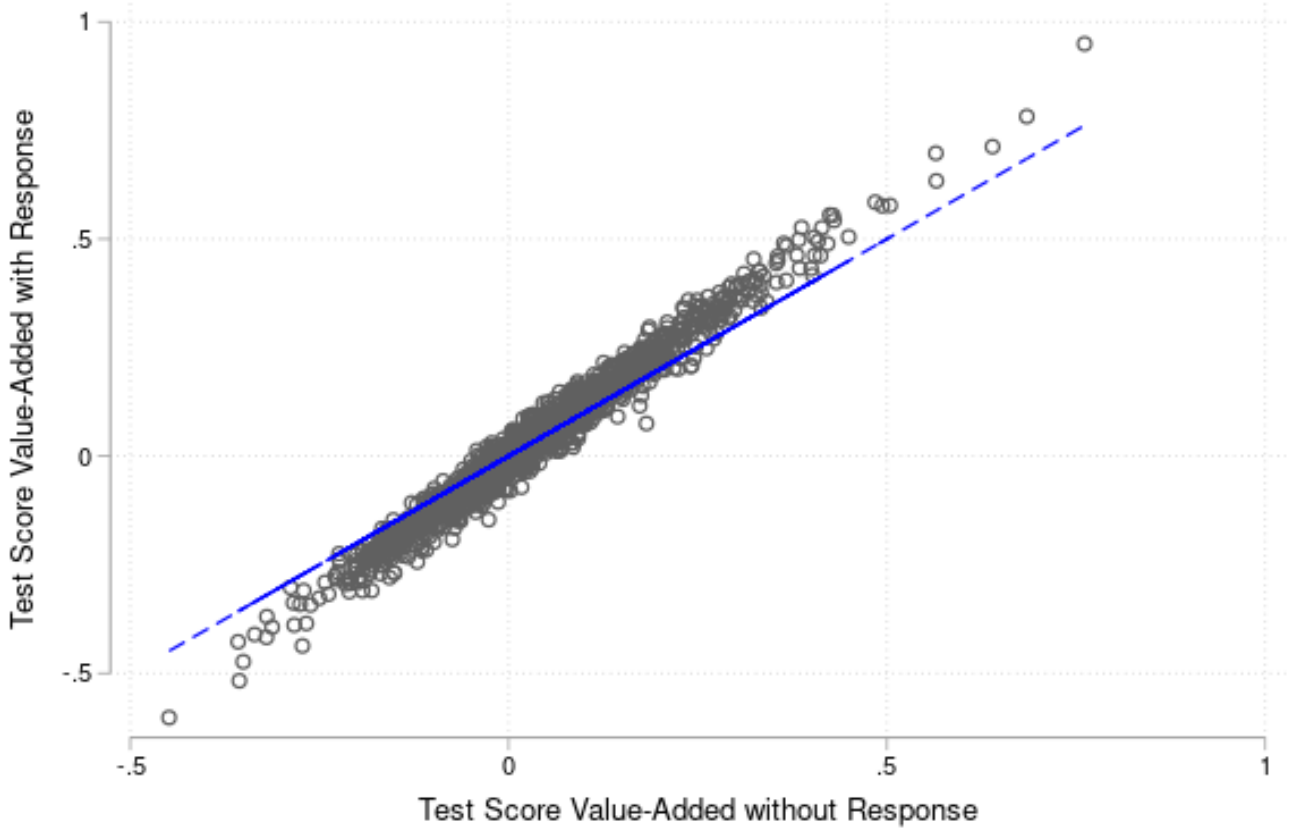
Note: This figure plots mean test score value-added for each cohort of teachers in the standard probationary period and at each level of prior experience. We plot separate lines for teacher cohorts based on the number of years they were exposed to the new tenure policy. Diamonds designate years in the probationary period under the new tenure policy.

Figure 8: Heterogeneous Response



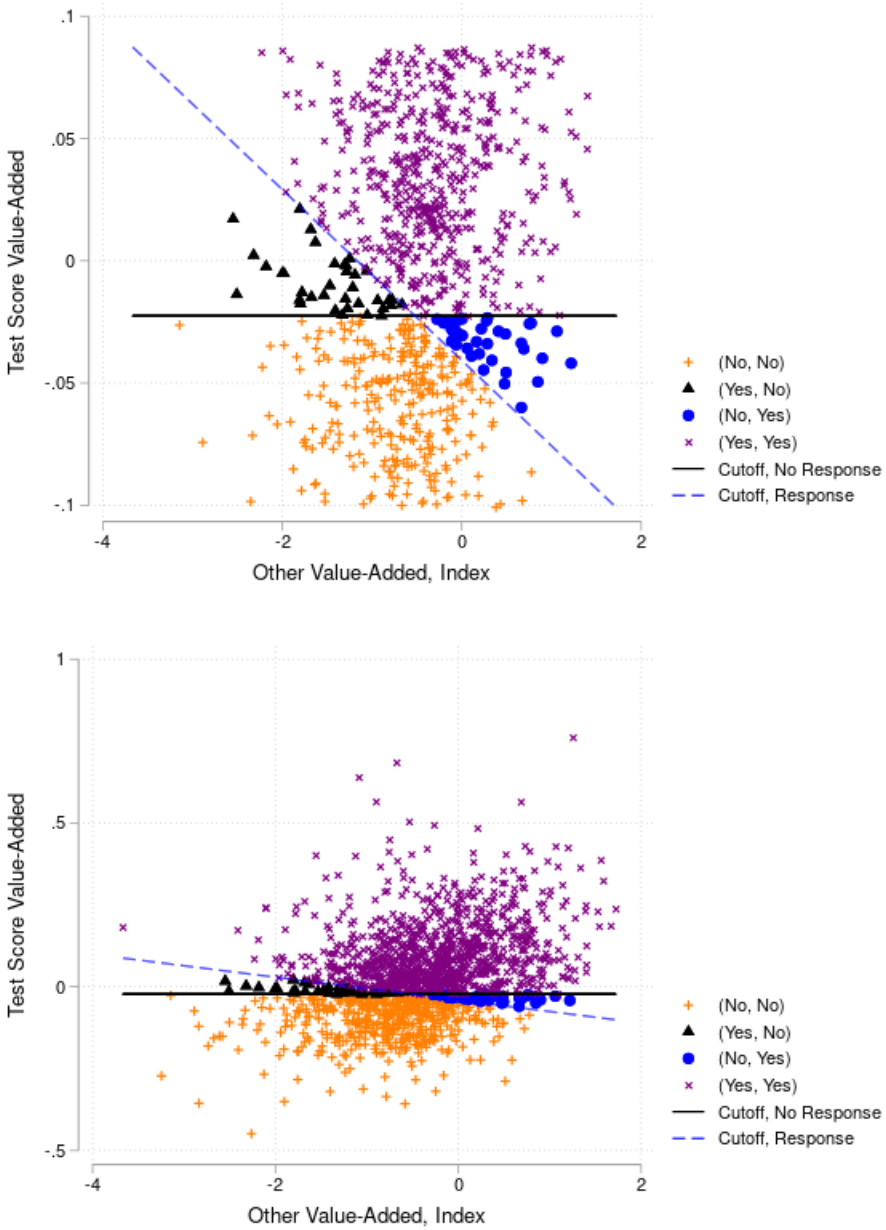
Note: This figure plots the predictiveness of the untargeted index forecast on test scores. We estimate this predictiveness by running a cross-sectional regression of a teacher's mean test score residuals on her forecasted value-added in the index of untargeted measures. We run a separate regression for each year and plot the coefficient on the forecasted value-added in the index. We plot separate lines for teacher cohorts based on the number of years they were exposed to the new tenure policy. Diamonds designate years in the probationary period under the new tenure policy. The "always exposed" cohorts are not included because this figure relies on forecasted test score value-added, as estimated in unincentivized periods. Because they are "always exposed," these cohorts do not have data to form such a forecast.

Figure 9: Response in Test Score Value Added



Note: This figure shows the joint distribution of teachers' test score value-added with and without the behavioral response to the policy. We construct the response based on our estimates. The blue line is the 45-degree line.

Figure 10: Tenure Predictions, with and without Behavioral Responses



Note: This figure shows the joint distribution of teacher forecasted value-added in the untargeted measure index (x-axis) and in test scores (y-axis). Each teacher in the 2007 cohort is plotted based on her forecasted value-added in unincentivized periods. The different symbols and colors reflect whether, according to our estimates, the teacher would receive tenure if (a) no teachers' behavior responded to the policy change and (b) if all teachers' behavior responded. For instance, "(No, Yes)" designates teachers who would only receive tenure when there are behavioral responses. The lines show the tenure cutoffs without behavioral responses (solid black) and with behavioral responses (dashed blue). The top panel zooms in on the test score value-added range between -0.1σ and 0.1σ while the bottom panel zooms out and shows all teachers.

IX.B Tables

Table 1: Summary Statistics

	Obs.	Mean	Std. Dev.	Min	Max
Student Data					
Male	4,602,185	0.49	0.50	0.00	1.00
Asian	4,602,185	0.17	0.38	0.00	1.00
Black	4,602,185	0.27	0.45	0.00	1.00
Hispanic	4,602,185	0.39	0.49	0.00	1.00
White	4,602,185	0.16	0.36	0.00	1.00
High-Poverty	4,602,185	0.80	0.40	0.00	1.00
English Language Learner	4,602,185	0.11	0.31	0.00	1.00
Middle Schooler	4,602,185	0.57	0.49	0.00	1.00
Math Score	2,325,544	0.00	1.00	-7.29	4.13
ELA Score	2,276,641	-0.00	1.00	-12.59	8.46
Attendance Rate	4,595,853	0.94	0.06	0.00	1.00
Grade in Math	1,032,550	80.36	11.99	10.00	100.00
Grade in ELA	815,248	79.19	11.30	10.00	100.00
Grade in Untested Subjects	2,573,954	81.81	9.35	10.00	100.00
Teacher Data					
Years Teaching at Current School	97,687	6.12	5.59	0.00	49.08
Years Teaching in District	97,687	8.34	6.66	0.00	49.17
In Probationary Period	98,520	0.21	0.41	0.00	1.00
Counts					
Number of Students	899,291				
Number of Student-Years	2,478,028				
Number of Student-Year-Subjects	4,602,185				
Number of Teachers	28,946				
Number of Teacher-Years	98,520				
Number of Teacher-Year-Subjects	145,021				

This table shows summary statistics for our student and teacher estimation samples. "High-Poverty" indicates eligibility for free or reduced price lunch. Math and ELA scores are normalized to have mean 0 and standard deviation 1. The probationary period is the pre-tenure period.

Table 2: Relationship between Tenure Decisions and Output

	On-Time Tenure	On-Time Tenure
Targeted Output	-0.00974	0.260***
	(0.0130)	(0.0582)
Untargeted Output	-0.000118	0.0127
	(0.00152)	(0.0130)
Constant	0.973***	0.658***
	(0.00356)	(0.0177)
Sample	2008-2009	2011-2012
Mean DV	0.972	0.670
N	2137	700

This table shows the relationship between teacher output and whether she receives tenure by the beginning of her fourth year (on-time). Targeted output is the (unshrunken) mean test score residual. Untargeted output is the (unshrunken) index of other measures. An observation is a teacher in her fourth year of experience The columns cover samples before and after the reform, respectively.

Table 3: Summary of Empirical Strategy

Entry Cohort	2007	2008	2009	2010	2011	2012
2007	0	0	0	-	-	-
2008		0	0	1	-	-
2009			0	1	1	-
2010				1	1	1

This tables summarizes the empirical strategy by showing how we use within-cohort variation. Rows are the year the teacher entered the district and the columns are (spring) academic years. A blank entry means the teacher is not yet in the district. A dash means that the teacher is out of the standard probationary period. A blue zero indicates the teacher is in the probationary period, but *without* test score value-added incentives. A red one indicates the teacher is in the probationary period, *with* test score value-added incentives.

Table 4: Effect of Policy Change on Probationary Period Output

	Test Score	Test Score	Untargeted Index	Untargeted Index
Incentive	0.0302***	0.0187**	-0.0576*	-0.0644**
	(0.00881)	(0.00869)	(0.0303)	(0.0307)
Fixed Effects	Cohort	Teacher	Cohort	Teacher
N Teachers	16724	16724	16724	16724
Mean DV	0.102	0.102	0.0244	0.0244
N	100405	100405	100405	100405

This table shows the causal effect of the tenure policy change on targeted and untargeted output in the probationary period. The columns switch between cohort and teacher fixed effects. An observation is a teacher-subject-year. Standard errors are clustered by teacher. The sample covers years 2006 to 2014. The teachers with more than 3 years of experience are only included if they finished the standard probationary period before the tenure policy change. We include teachers with targeted and untargeted measures. All outcome units are test score student standard deviations.

Table 5: Effect of Policy Change on Specific Untargeted Outcomes

	Score 1	Score 2	Attend	Attend 1	Grades	Other Grades	Grades 1	Other Grades 1
Incentive	-0.0263**	-0.00648	0.00282	-0.0187*	0.0372**	-0.00310	-0.0630***	-0.0405**
	(0.0115)	(0.0143)	(0.00245)	(0.0113)	(0.0179)	(0.0169)	(0.0203)	(0.0197)
Teacher FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean DV	0.0482	0.0433	0.00386	0.0277	0.0246	-0.00840	0.0362	0.0380
N	105861	84359	127444	106376	32984	43418	56362	70194

This table shows the causal effect of the tenure policy change on individual untargeted (residualized) outcomes in the probationary period. The "1" or "2" in the column headers indicate the measure's number of years into the future. "Grades" are in the tested subject while "Other" grades are in untested subjects. All variables are standardized at the grade-year level to have mean 0 and standard deviation 1 in the full population. All regressions include teacher fixed effects. An observation is a teacher-subject-year. Standard errors are clustered by teacher. The sample covers years 2006 to 2014. The teachers with more than 3 years of experience are only included if they finished the standard probationary period before the tenure policy change.

Table 6: Effect of Policy Change on Predetermined Student Characteristics

	Lag Score	Lag Other Score	Male	Black	Hispanic	Asian	ELL	Poverty
Incentive	0.000390	-0.00311	0.00121	0.00180	0.00111	-0.000643	-0.0000865	-0.00301
	(0.0141)	(0.0142)	(0.00407)	(0.00467)	(0.00498)	(0.00345)	(0.00506)	(0.00671)
Teacher FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean DV	-0.0175	-0.0144	0.488	0.290	0.397	0.158	0.119	0.813
N	127678	127678	127678	127678	127678	127678	127678	127678

This table shows the causal effect of the tenure policy change on predetermined student characteristics in the probationary period. "Score" is the test score in the same subject while "Other Score" is in the opposite subject. "ELL" are English language learners. "Poverty" indicates students eligible for free or reduced price lunch. All regressions include teacher fixed effects. An observation is a teacher-subject-year. Standard errors are clustered by teacher. The sample covers years 2006 to 2014. The teachers with more than 3 years of experience are only included if they finished the standard probationary period before the tenure policy change.

Table 7: Effect of Policy Change on Post-Probationary Period Output

	Test Score	Untargeted Index
Incentive	0.0185***	-0.0421*
	(0.00588)	(0.0223)
Post-Incentive	-0.0114	-0.0366
	(0.00862)	(0.0439)
Fixed Effects	Teacher	Teacher
N Teachers	22710	17126
Mean DV	0.0972	0.0286
N	135361	107261

This table shows the causal effect of the tenure policy change on targeted and untargeted output. We estimate separate coefficients for the standard probationary period (first 3 years) and for after the teacher has tenure ("Post"). Each of these indicators is interacted with whether the new tenure policy is in place. All regressions include teacher fixed effects. An observation is a teacher-subject-year. Standard errors are clustered by teacher. The sample covers years 2006 to 2014. All outcome units are test score student standard deviations.

Table 8: Heterogeneous Responses to the Incentive

	Score	Score	Score	Score	Score	Score
Incentive	0.0142*	0.00125	0.0241***	0.0185*	0.0148*	0.0122
	(0.00756)	(0.00988)	(0.00897)	(0.0104)	(0.00818)	(0.0108)
Incentive * Targeted VA	0.0703	0.217***			0.0546	0.164**
	(0.0517)	(0.0725)			(0.0524)	(0.0734)
Incentive * Untargeted VA			0.651*	1.128***	0.0158	0.0407**
			(0.338)	(0.423)	(0.0110)	(0.0165)
FEs	Cohort	Teacher	Cohort	Teacher	Cohort	Teacher
SD VA1	.142	.142			.142	.142
SD VA2			.482	.482	.482	.482
Mean DV	0.00594	0.00594	0.00594	0.00594	0.00594	0.00594
N	84399	84399	84399	84399	84399	84399

This table shows the causal effect of the tenure policy change on targeted output, split by a teacher's forecasted targeted and untargeted value-added. The forecasts are based on a value-added model estimated with data from the unincentivized period. Regressions include either cohort or teacher fixed effects. Each value-added measure is interacted with experience indicators. "SD VA1" and "SD VA2" are the cross-sectional standard deviations of targeted and untargeted forecasted value-added. An observation is a teacher-subject-year. Standard errors are clustered by teacher. The sample covers years 2006 to 2014. The teachers with more than 3 years of experience are only included if they finished the standard probationary period before the tenure policy change. Outcome units are test score student standard deviations.

Table 9: Output under Different Screening Regimes

	Obs.	Mean Targeted Output	Mean Untargeted Output	Mean Total Output
Teachers' Tenure under Different Responses				
Never Tenured	510	-0.104	-0.786	-0.890
Only Tenured w/o Behavioral Response	36	-0.009	-1.444	-1.453
Only Tenured w/ Behavioral Response	36	-0.033	0.286	0.253
Always Tenured	1,077	0.100	-0.237	-0.136
Tenured Teachers under Different Policies				
Screening w/o Behavioral Response	1,113	0.097	-0.276	-0.179
Screening w/ Behavioral Response	1,113	0.096	-0.220	-0.124
(Infeasible) Screening on Both Dimensions	1,113	0.064	-0.046	0.019
Gains Relative to Infeasible First-Best				
Gains (Fraction)				0.279

This table shows the mean output for different groups of teachers and under different policies. The sample is teachers who started in the district in 2007. The top panel splits teachers into four groups based on whether they would receive tenure in a regime without a behavioral response and whether they would receive tenure in a regime with a behavioral response. The middle panel shows the mean output associated with the set of teachers receiving tenure under different policies. The first two policies are screening on the targeted measure, without and with a behavioral response. The last policy is an infeasible policy screening on the sum of output across both dimensions. The final panel shows the fraction of gains the behavioral response achieves relative to the distance between the screening without behavioral response regime and the infeasible policy screening on the sum of output. "Mean Targeted" output is the mean forecasted test score value-added. "Mean Untargeted Output" is the mean forecasted value-added on the untargeted index. "Mean Total Output" is the sum of the mean forecasted value-added across the targeted and untargeted measures. All outcome units are test score student standard deviations.

A Proofs

In this Section, we present the proofs of the two theorems in the main body of the paper. We start in Subsection A.1 with the proof of Theorem 1. For clarity, we then break Theorem 2 into its two components: Theorem A.1, which deals with the case in which incentives are large, and Theorem A.2, which deals with the case in which costs are quadratic. In both cases, we provide the main structure of the proofs in Subsection A.1, while leaving the main "economic" components of the proofs to Subsection A.2 and Subsection A.3. We leave other supporting Lemmas – ones that were omitted from the proofs to improve clarify – to Subsection A.4.

A.1 Main Proofs

Theorem 1. *For any weakly increasing screening function, $x_1^*(\theta|p) \geq x_1^*(\theta)$ for every θ. Furthermore, for every θ:*

$$x_2^*(\theta|p) \leq x_2^*(\theta) \qquad \textit{if} \qquad \frac{\partial^2 c}{\partial e_1 \partial e_2} \geq 0$$
$$x_2^*(\theta|p) \geq x_2^*(\theta) \qquad \textit{if} \qquad \frac{\partial^2 c}{\partial e_1 \partial e_2} \leq 0.$$

Proof. We begin by defining:

$$f(e, \theta, a) = u(e, \theta) + a \cdot \lambda \Delta v \cdot p(e_1 \theta_1). \tag{21}$$

Note that under the announced screening policy, when $a = 1$ the function is equivalent to the optimization problem for an individual that defines $x^*(\theta|p)$ and when $a = 0$ the function is equivalent to the optimization problem for an individual that defines $x^*(\theta)$. Therefore, if we can show that the function is supermodular in $(e_1, -e_2, a)$ we can appeal to Topkis' Theorem to show that $e_1^*(\theta|p) \geq e_1^*(\theta)$ and $e_2^*(\theta|p) \leq e_2^*(\theta)$, and hence $x_1^*(\theta|p) \geq x_1^*(\theta)$ and $x_2^*(\theta|p) \leq x_2^*(\theta)$. Conversely, if we can show that the function is supermodular in (e_1, e_2, a) we can appeal to Topkis' Theorem to show that $e_1^*(\theta|p) \geq e_1^*(\theta)$ and $e_2^*(\theta|p) \geq e_2^*(\theta)$, and hence $x_1^*(\theta|p) \geq x_1^*(\theta)$ and $x_2^*(\theta|p) \geq x_2^*(\theta)$.

To look at whether f supermodular in (e_1, e_2, a) or $(e_1, -e_2, a)$, we first note that:

$$f(e, \theta, a = 1) - f(e, \theta, a = 0) = \lambda \Delta v \cdot p(e_1 \theta_1). \tag{22}$$

Since $p(e_1\theta_1)$ is increasing in e_1 it follows that f is supermodular in (e_1, a). The fact that e_2 does not appear in the equation also implies that f is both supermodular $(-e_2, a)$ and (e_2, a). If $\frac{\partial^2 c}{\partial e_1 \partial e_2} \geq 0$, then f is also supermodular in $(e_1, -e_2)$ and so f is pairwise

supermodular – and therefore supermodular – in $(e_1, -e_2, a)$. If $\frac{\partial^2 c}{\partial e_1 \partial e_2} \leq 0$ then f is also supermodular in (e_1, e_2) and so f is pairwise supermodular in (e_1, e_2, a) □

Theorem 2. *There exists κ such that if either: a) $\lambda \cdot \Delta v \geq \kappa$ or b) $c(e)$ is quadratic then:*

1. *The endogenous response makes screening more efficient if $\frac{\partial^2 c}{\partial e_1 \partial e_2} > 0$;*
2. *The endogenous response makes screening less efficient if $\frac{\partial^2 c}{\partial e_1 \partial e_2} < 0$.*

Proof. This theorem has two conditions – a) that the incentive is large, i.e., $\lambda \cdot \Delta v \geq \kappa$, and b) the $c(e)$ is quadratic – with the conclusions holding if either one is true. To clarify the exposition, we separate the main theorem into two different theorems each of which focus on a single one of the two conditions. These two theorems and proofs are listed below and directly imply this theorem. □

Thm A.1. *There exists κ such that if $\lambda \cdot \Delta v \geq \kappa$ then:*

1. *The endogenous response makes screening more efficient if $\frac{\partial^2 c}{\partial e_1 \partial e_2} > 0$;*
2. *The endogenous response makes screening less efficient if $\frac{\partial^2 c}{\partial e_1 \partial e_2} < 0$.*

Proof. Consider any ex post screening policy that retains a fraction ρ of individuals and define the threshold for this policy as $\overline{\xi}$, i.e., the policy is $p_{\overline{\xi}}(x_1) = \mathbf{1}(x_1 \geq \overline{\xi})$. Then for any $\lambda\Delta v$, define $\xi(\lambda\Delta v)$ as the threshold that ensures the announced screening policy with a threshold screening function $\tilde{p}(x_1) = \mathbf{1}(x_1 \geq \xi(\lambda\Delta v))$ also retains a fraction ρ of individuals. We show in Lemma A.5 that such a $\xi(\lambda\Delta v)$ exists for all values of ρ and $\lambda\Delta v$. Finally, define e_1^{max} to be the maximum feasible effort level, i.e., $e_1^{max} = sup\{e_1 | c(e) < \infty\}$, and ξ^{max} to be the threshold where only a fraction of ρ individuals can produce ξ^{max} of the first good when exerting effort level e_1^{max}, i.e., the ξ^{max} for which where $\mathbb{E}_\Theta[\mathbf{1}(e_1^{max}\theta_1 \geq \xi^{max})] = \rho$.[45]

We next consider a sequence $\lambda\Delta v_n$ such that $\lambda\Delta v_n \to \infty$. Based on this sequence, we can define a corresponding sequence of thresholds ξ_n such that $\xi_n = \xi(\lambda\Delta v_n)$. We show in Lemma A.6 that $\xi_n \leq \xi^{max}$ and that $\xi_n \to \xi^{max}$. Similarly, we can consider a sequence of functions $p_n : \Theta \to \{0, 1\}$ as follows:

$$p_n(\theta) = \mathbf{1}\Big(\tilde{u}(\xi_n, \theta) + \lambda\Delta v_n - v(\theta) \geq 0\Big) \tag{23}$$

where $\tilde{u}(x_1, \theta)$ is the indirect utility conditional on producing at least x_1 of the first good, i.e.,

$$\tilde{u}(x_1, \theta) \equiv \begin{cases} \max_{e \in \mathcal{E}} u(e, \theta) \text{ s.t. } e_1\theta_1 \geq x_1 & \text{if } e_1^{max}\theta_1 \geq x_1 \\ -\infty & \text{if } e_1^{max}\theta_1 < x_1. \end{cases} \tag{24}$$

[45] In our model, $\mathcal{E}$ is a compact set and so we are guaranteed that e_1^{max} exists and is finite.

Note that the utility for an individual with ability θ of producing enough x_1 output to remain under the screening policy with threshold ξ_n and value of remaining of $\lambda\Delta v_n$ is $\tilde{u}(\xi_n, \theta) + \lambda\Delta v_n$ and the value of not producing enough is $v(\theta)$. Thus, $p_n(\theta)$ is a mapping in Θ-space of the individuals retained under that screening policy. Note also that by the definition of ξ_n, the fraction retained will be exactly ρ.

We then let $F_{\theta_1}^{-1}$ be the generalized inverse CDF of θ_1 and define $p_\infty(\theta) = \mathbf{1}\Big(\theta_1 \geq F_{\theta_1}^{-1}(1-\rho)\Big)$.[46] We refer to this function as p_∞ because $p_n \to p_\infty$ pointwise at almost every θ, as we show in Lemma A.7.

Loosely speaking, this means that for large enough n the functions p_n and p_∞ are nearly identical. Crucially, this means when assessing how the endogenous response affects screening efficiency under a large incentive, instead of comparing the functions p_n to the ex post screening function, which we had denoted as $p_{\bar{\xi}}$, we can instead compare p_∞ to $p_{\bar{\xi}}$ and from there infer that the comparison also holds for all functions p_n for sufficiently large n.

More formally, we start with the following equation:

$$\mathbb{E}_\Theta\Big[V(\theta)\cdot p_n(\theta)\Big] - \mathbb{E}\Big[V(\theta)\cdot p_{\bar{\xi}}\big(x_1^*(\theta)\big)\Big] = \mathbb{E}_\Theta\Big[V(\theta)\cdot\big(p_n(\theta) - p_\infty(\theta)\big)\Big] + \mathbb{E}_\Theta\Big[V(\theta)\cdot\big(p_\infty(\theta) - p_{\bar{\xi}}\big(x_1^*(\theta)\big)\big)\Big] \tag{25}$$

Now suppose that $\mathbb{E}_\Theta\Big[V(\theta)\cdot\big(p_\infty(\theta) - p_{\bar{\xi}}\big(x_1^*(\theta)\big)\big)\Big] > 0$, which implies that $\mathbb{E}_\Theta\Big[V(\theta)\cdot\big(p_\infty(\theta) - p_{\bar{\xi}}\big(x_1^*(\theta)\big)\big)\Big] = 2\epsilon$ for some $\epsilon > 0$. From the bounded convergence theorem and the fact that $p_n \to p_\infty$ pointwise at almost every θ, we can then choose an $N > 0$ such that $\forall n > N$ we have $\Big|\mathbb{E}_\Theta\Big[V(\theta)\cdot\big(p_n(\theta) - p_\infty(\theta)\big)\Big]]\Big| < \epsilon$. Plugging this into Equation 25, we get that:

$$\mathbb{E}_\Theta\Big[V(\theta)\cdot p_n(\theta)\Big] - \mathbb{E}\Big[V(\theta)\cdot p_{\bar{\xi}}\big(x_1^*(\theta)\big)\Big] > -\epsilon + 2\epsilon = \epsilon > 0.$$

In a similar fashion, if $\mathbb{E}_\Theta\Big[V(\theta)\cdot\big(p_\infty(\theta) - p_{\bar{\xi}}\big(x_1^*(\theta)\big)\big)\Big] < 0$ we can choose a large enough N such that $\forall n > N$ we have that $\mathbb{E}_\Theta\Big[V(\theta)\cdot p_n(\theta)\Big] - \mathbb{E}\Big[V(\theta)\cdot p_{\bar{\xi}}\big(x_1^*(\theta)\big)\Big] < 0$.

The question of how the endogenous response affects screening efficiency (under large incentives) thus boils down to a question of the sign of $\mathbb{E}_\Theta\Big[V(\theta)\cdot\big(p_\infty(\theta) - p_{\bar{\xi}}\big(x_1^*(\theta)\big)\big)\Big]$. As we show in Lemma A.1, this in turn depends on the sign of $\frac{\partial^2 c}{\partial e_1 \partial e_2}$. Namely, if $\frac{\partial^2 c}{\partial e_1 \partial e_2} < 0$, we get that $\mathbb{E}_\Theta\Big[V(\theta)\cdot\big(p_\infty(\theta) - p_{\bar{\xi}}\big(x_1^*(\theta)\big)\big)\Big] < 0$; in contrast, if $\frac{\partial^2 c}{\partial e_1 \partial e_2} > 0$, we get that $\mathbb{E}_\Theta\Big[V(\theta)\cdot\big(p_\infty(\theta) - p_{\bar{\xi}}\big(x_1^*(\theta)\big)\big)\Big] > 0$.

□

Thm A.2. *If $c(e)$ is quadratic then:*

[46] Formally, this is defined as $F_{\theta_1}^{-1}(1-\rho) = \inf\{\theta_1 | F_{\theta_1}(\theta_1) = 1-\rho\}$, where F_{θ_1} is the CDF of θ_1. Note that if a fraction of $1-\rho$ individiuals have $\theta_1 < F_{\theta_1}^{-1}(1-\rho)$ then a fraction ρ have $\theta_1 > F_{\theta_1}^{-1}(1-\rho)$.

1. *The endogenous response makes screening more efficient if* $\frac{\partial^2 c}{\partial e_1 \partial e_2} > 0$*;*

2. *The endogenous response makes screening less efficient if* $\frac{\partial^2 c}{\partial e_1 \partial e_2} < 0$*.*

Proof. Start with any pair of comparison threshold policies p and $\tilde{p}$, i.e., $\tilde{p}$ is an ex post screening policy with threshold $\overline{\xi}$ that retains a fraction ρ of individuals and p is an announced screening policy with threshold ξ that also retains a fraction ρ of individuals. Again, from Lemma A.5 we know this pair exists for any ρ.

We will then define the sets A and B such that A is the set of individuals retained under announced screening policy but removed under the ex post screening policy and set B to be the set of individuals that are removed under announced screening policy but retained under the ex post screening policy. Formally, we get that:

$$A = \{\theta | x_1^*(\theta) < \overline{\xi} \ \& \ x_1^*(\theta|p) \geq \xi\}$$
$$B = \{\theta | x_1^*(\theta) \geq \overline{\xi} \ \& \ x_1^*(\theta|p) < \xi\}.$$

Now consider the case in which $\frac{\partial^2 c}{\partial e_1 \partial e_2} > 0$ and take any two points $\theta^A \in A, \theta^B \in B$.[47] From the definition of A and B, we know that $x_1^*(\theta^A) < x_1^*(\theta^B)$ and from Lemma A.8 we know that $x_1^*(\theta)$ is increasing in θ_1 and decreasing in θ_2. Thus it cannot be the case that $\theta_1^A > \theta_1^B$ and $\theta_2^A < \theta_2^B$. Similarly, from the definition of A and B we know that $x_1^*(\theta^A|p) > x_1^*(\theta^B|p)$ and from Lemma A.8 we know that $x_1^*(\theta|p)$ is also increasing in θ_1 and decreasing in θ_2. Thus, it also cannot be the case that $\theta_1^A < \theta_1^B$ and $\theta_2^A > \theta_2^B$ and we can conclude that either $\theta^A > \theta^B$ or $\theta^A < \theta^B$.

To show that $\theta^A > \theta^B$ we use a proof by contradiction. Assume that $\theta_1^A < \theta_1^B$ which – from the argument above – also implies that $\theta_2^A < \theta_2^B$. Then define $\hat{\theta}$ such that $\hat{\theta}_1 = \theta_1^B$ and $\hat{\theta}_2 = \theta_2^A$. From the comparative statics on $x^*(\theta)$ we then get that: $x_1^*(\hat{\theta}) \geq x_1^*(\theta^B)$. By the intermediate value theorem, we can then conclude that there exists $\tilde{\theta}$ such that $x_1^*(\tilde{\theta}) = x_1^*(\theta^B)$, $\tilde{\theta}_2 = \theta_2^A < \theta_2^B$, and $\tilde{\theta}_1 \leq \theta_1^B$. From Subsection A.3, we then get that $x_1^*(\tilde{\theta}|p) \leq x_1^*(\theta^B|p)$ and from the fact that $x_1^*(\theta|p)$ is increasing in θ we get that $x_1^*(\tilde{\theta}|p) \geq x_1^*(\theta^A|p)$. Thus, $x_1^*(\theta^B|p) \geq x_1^*(\tilde{\theta}|p) \geq x_1^*(\theta^A|p)$, which is a contradiction based on the definitions of A and B. We can therefore conclude that $\theta_1^A > \theta_1^B$, which from the argument above also implies that $\theta_2^A > \theta_2^B$. Since $V(\theta)$ is increasing in θ, this implies that $V(\theta^A) > V(\theta^B)$ and since this is true for every $\theta^A \in A, \theta^B \in B$, it follows that the endogenous response makes screening more efficient.

We next consider the case that $\frac{\partial^2 c}{\partial e_1 \partial e_2} < 0$ and again take any two points $\theta^A \in A, \theta^B \in B$ and from the definition of A and B we know that $x_1^*(\theta^A) < x_1^*(\theta^B)$ and so our approach

[47] We will assume that A and B are not empty. If they are, the announcement has no effect on screening at all, in which case the inequality in our definition of efficiency clearly holds.

is to show that it is also the case that $x_2^*(\theta^A) < x_2^*(\theta^B)$. First, we have that $x_1^*(\theta|p)$ is increasing in $x_1^*(\theta)$; see Lemma A.10. Next, we note that from Subsection A.3, we get that $x_1^*(\theta|p)$ is decreasing in $x_2^*(\theta)$ conditional on $x_1^*(\theta)$. Thus, if $x_2^*(\theta^A) > x_2^*(\theta^B)$ it would follow that $x_1^*(\theta^A|p) \leq x_1^*(\theta^B|p)$, which is false from the definition of A and B. It must therefore must be the case that $x_2^*(\theta^A) < x_2^*(\theta^B)$. This implies that $V(x^*(\theta^A)) < V(x^*(\theta^B)))$ and, since this is true for every $\theta^A \in A, \theta^B \in B$, it follows that the endogenous response makes screening less efficient.

□

A.2 Screening on Ability vs Screening on Output

Much of the intuition for Theorem A.1 can be seen in a comparison between two infeasible screening regimes. One – which we refer to as "ex post screening" – screens on the teachers' output without the additional incentive of the screening policy in place, i.e., on $x_1^*(\theta)$, and the other – which we refer to as "ability screening" – screens on the first dimension of ability, i.e., on θ_1. These two policies are referred to as $p_{\overline{\xi}}$ and p_∞, respectively, in Theorem A.1. We show in the Lemma below that which one is more efficient at screening depends on the cross-derivatives of the cost function.

Lemma A.1. *Define $p_{\overline{\xi}}$ and p_∞ as in Theorem A.1. Then:*

- *If $\frac{\partial^2 c}{\partial e_1 \partial e_2} > 0$, then ability screening is more efficient than ex post screening, i.e., $\mathbb{E}_\Theta[V(\theta) \cdot p_\infty(\theta)] \geq \mathbb{E}[V(\theta) \cdot p_{\overline{\xi}}(x_1^*(\theta))]$.*
- *If $\frac{\partial^2 c}{\partial e_1 \partial e_2} < 0$, then ability screening is less efficient than ex post screening, i.e., $\mathbb{E}_\Theta[V(\theta) \cdot p_\infty(\theta)] \leq \mathbb{E}[V(\theta) \cdot p_{\overline{\xi}}(x_1^*(\theta))]$.*

Proof. We start by defining the sets A and B such that A is the set of individuals removed under $p_{\overline{\xi}}$ but retained under p_∞ and set B to be the set of individuals that are retained under $p_{\overline{\xi}}$ but removed under p_∞.[48] Formally, we get that:

$$A = \{\theta | x_1^*(\theta) < \overline{\xi} \ \& \ \theta_1 \geq F_{\theta_1}^{-1}(1-\rho)\}$$
$$B = \{\theta | x_1^*(\theta) \geq \overline{\xi} \ \& \ \theta_1 < F_{\theta_1}^{-1}(1-\rho)\}$$

We then consider any two points $\theta^A \in A, \theta^B \in B$. Note that from the definition of the sets we clearly get that $\theta_1^A > \theta_1^B$ and $x_1^*(\theta^A) < x_1^*(\theta^B)$. Under the assumption that $\frac{\partial^2 c}{\partial e_1 \partial e_2} > 0$, it follows that $x_1^*(\theta)$ is increasing in θ_1 and decreasing in θ_2. Since $x_1^*(\theta^A) < x_1^*(\theta^B)$ and $\theta_1^A > \theta_1^B$, it must then be the case that $\theta_2^A > \theta_2^B$. Thus, $\theta^A > \theta^B$ and so, since $V(\theta)$ is

[48] We assume that the sets are not empty; otherwise the policies are identical and the weak inequalities hold with equality.

increasing in θ, it follows that $V(\theta^A) > V(\theta^B)$. Since this is true for all $\theta^A \in A, \theta^B \in B$, we clearly get that $\mathbb{E}_\Theta[V(\theta) \cdot p_\infty(\theta)] > \mathbb{E}[V(\theta) \cdot p_{\bar{\xi}}\big(x_1^*(\theta)\big)]$.

The proof of the the case in which $\frac{\partial^2 c}{\partial e_1 \partial e_2} < 0$ proceeds similarly, although it requires transforming the points to X^* space. Again, we consider two points $\theta^A \in A, \theta^B \in B$, which means that $\theta_1^A > \theta_1^B$ and $x_1^*(\theta^A) < x_1^*(\theta^B)$. We then consider the inverse of $x^*(\theta)$, i.e. a function from X^* space to Θ-space which we denote as $\theta(x^*)$. As we show in Lemma A.9, if such a function exists and is differentiable and $\frac{\partial^2 c}{\partial e_1 \partial e_2} < 0$ then $\theta_1(x^*)$ is increasing in x_1^* and decreasing in x_2^*. Thus, since $\theta_1^A > \theta_1^B$ and $x_1^*(\theta^A) < x_1^*(\theta^B)$ it must be the case that $x_2^*(\theta^A) < x_2^*(\theta^B)$ and so $V\big(x^*(\theta^A)\big) < V\big(x^*(\theta^B)\big)$. Since this is true for all $\theta^A \in A, \theta^B \in B$, we clearly get that $\mathbb{E}_\Theta[V(\theta) \cdot p_\infty(\theta)] < \mathbb{E}[V(\theta) \cdot p_{\bar{\xi}}\big(x_1^*(\theta)\big)]$.

□

We can think of this lemma as being roughly akin to a screening version of the traditional multitasking problem. The traditional multitasking problem highlights that when the principal can only add an incentive to a single output it also affect production of the other outcomes and that whether it has a positive or negative impact on the other outcomes depends on the cross-derivative of the cost function. The screening version outlined in the lemma above instead highlights that when the principal can only screen on a single output, she also implicitly screens on the individuals' ability to improve the other outcomes (since their choice of effort allocation depends on both abilities). Furthermore, whether the policy positively or negative selects on the individuals' ability to improve the other outcomes (conditional on their ability to improve the outcome being screened on) also depends on cross-derivative of the cost function.

A.3 Multitasking Heterogeneity

In the proof of Theorem A.2, we show that the question of whether the endogenous response increases or decreases screening efficiency can be boiled down to a question of how individuals who produce the same output on the first dimension without additional incentives respond differentially to the incentive. Here, we show that the answer to that question depends on what what amounts to a single-crossing condition on the marginal utility functions.

We start by defining $\tilde{u}(x_1, \theta)$ – as we do above – to be the optimal utility individual θ can get when constrained to produce at least x_1, i.e.,

$$\tilde{u}(x_1, \theta) \equiv \begin{cases} \max_{e \in \mathcal{E}} u(e, \theta) \text{ s.t. } e_1\theta_1 \geq x_1 & \text{if } e_1^{max}\theta_1 \geq x_1 \\ -\infty & \text{if } e_1^{max}\theta_1 < x_1. \end{cases} \tag{26}$$

We use $\tilde{u}'(x_1,\theta)$ to denote $\frac{\partial \tilde{u}(x_1,\theta)}{\partial x_1}$ where $\tilde{u}$ is differentiable and $\tilde{u}''(x_1,\theta)$ to be defined similarly. While differentiability is not an necessary assumption in our results, it makes the notation and intuition more straightforward. We then have the following Lemma:

Lemma A.2. *Consider any θ,θ' with $x_1^*(\theta) = x_1^*(\theta')$ and suppose that $\tilde{u}''(x_1,\theta') - \tilde{u}''(x_1,\theta)$ is increasing in x_1. Then for every weakly increasing function $f(x_1)$, we have $x_1^*(\theta'|f) \geq x_1^*(\theta|f)$. Similarly, for every weakly decreasing function $x_1^*(\theta'|f) \leq x_1^*(\theta|f)$.*

Proof. Consider any weakly increasing function $f(x_1)$. Clearly, $x_1^*(\theta|f) \geq x_1^*$ and $x_1^*(\theta'|f) \geq x_1^*$. If $x_1^*(\theta|f) = x_1^*$, then we are done, so we will assume in what follows that $x_1^*(\theta|f) > x_1^*$. We then consider any $x_1 \in (x_1^*, x_1^*(\theta|f))$. Since $x_1^*(\theta|f)$ is an optimizer, we get that $\tilde{u}(x_1,\theta) + f(x_1) \leq \tilde{u}(x_1^*(\theta|f),\theta) + f(x_1^*(\theta|f))$ or that $\tilde{u}(x_1^*(\theta|f),\theta) - \tilde{u}(x_1,\theta) \geq f(x_1) - f(x_1^*(\theta|f))$. Further, since $x_1^*(\theta|f) > x_1 \geq x_1^*$ we get from the assumption that $\tilde{u}''(x_1,\theta') - \tilde{u}''(x_1,\theta)$ is increasing in x_1 and the fact that $\tilde{u}'(x_1^*,\theta') = \tilde{u}'(x_1^*,\theta)$ that: $\tilde{u}(x_1^*(\theta|f),\theta') - \tilde{u}(x_1^*(\theta|f),\theta) > \tilde{u}(x_1,\theta') - \tilde{u}(x_1,\theta)$. Rearranging, it follows that:

$$\tilde{u}(x_1^*(\theta|f),\theta') - \tilde{u}(x_1,\theta') > \tilde{u}(x_1^*(\theta|f),\theta) - \tilde{u}(x_1,\theta) \geq f(x_1) - f(x_1^*(\theta|f)).$$

Thus, individual θ' would choose $x_1^*(\theta|f)$ over x_1 for all $x_1 \in (x_1^*, x_1^*(\theta|f))$, which along with the fact that $x_1^*(\theta'|f) \geq x_1^*$, proves that $x_1^*(\theta'|f) \geq x_1^*(\theta|f)$. The proof that $x_1^*(\theta'|f) \leq x_1^*(\theta|f)$ for any weakly decreasing function f is identical.

□

To understand the economic intuition of this Lemma, we can contrast that result with the traditional results on comparative statics. While traditional comparatives statics aims at understanding how the optimal choice of x varies according to characteristics θ, we consider a slightly different question and aim to understand how the *change* in the optimal choice of x varies according to characteristics θ when the individuals are presented with an identical *change* in incentives. Interestingly, the result mirrors the result from comparative statics, although it now hinges on increasing differences in the *marginal* utility function rather than the utility function itself.

The condition that $\tilde{u}''(x_1,\theta') - \tilde{u}''(x_1,\theta)$ is increasing in x_1 is slightly stronger than necessary, in a way that mirrors the fact that increasing differences is a slightly stronger condition than one needs for traditional comparative statics. We can relax the assumption to be equivalent to a single crossing condition. Here, the the single crossing condition is on the marginal utility function $\tilde{u}'(x_1,\theta)$ and states that that for θ,θ' with $x_1^*(\theta) = x_1^*(\theta') \equiv x_1^*$, we get that $\tilde{u}(x_1,\theta') - \tilde{u}(x_1,\theta)$ is strictly decreasing for all $x_1 < x_1^*$ and $\tilde{u}(x_1,\theta') - \tilde{u}(x_1,\theta)$ is strictly increasing for all $x_1 > x_1^*$. With $\tilde{u}'$ this corresponds to the fact that $\tilde{u}'(x_1,\theta) > \tilde{u}'(x_1,\theta)$ for every $x_1 > x_1^*$ and $\tilde{u}'(x_1,\theta) < \tilde{u}'(x_1,\theta)$ for every $x_1 < x_1^*$, or that $\tilde{u}'(x_1,\theta)$ and

$\tilde{u}'(x_1, \theta')$ cross a single time. As we show below, this single-crossing condition is both a sufficient and – in a limited sense – necessary condition for our comparative statics.

Lemma A.3. *Consider any θ, θ' with $x_1^*(\theta) = x_1^*(\theta') \equiv x_1^*$ and assume that $\tilde{u}$ is differentiable. Say that the single crossing condition on $\tilde{u}'$ holds iff $\tilde{u}'(x_1, \theta) > \tilde{u}'(x_1, \theta)$ for every $x_1 > x_1^*$ and $\tilde{u}'(x_1, \theta) < \tilde{u}'(x_1, \theta)$ for every $x_1 < x_1^*$.*

1. *If the single-crossing condition on $\tilde{u}'$ holds, then $x_1^*(\theta'|f) > x_1^*(\theta|f)$ for every strictly increasing differentiable function $f(x_1)$ and $x_1^*(\theta'|f) < x_1^*(\theta|f)$ for any strictly decreasing differentiable function.*

2. *If the single-crossing condition on $\tilde{u}'$ does not hold, then there exists a strictly increasing function such that $x_1^*(\theta'|f) \leq x_1^*(\theta|f)$ or there exists a strictly decreasing function such that $x_1^*(\theta'|f) \geq x_1^*(\theta|f)$.*

Proof. To prove the first point, note that the single crossing condition implies: $\tilde{u}'(x_1^*(\theta|f), \theta') > \tilde{u}'(x_1^*(\theta|f), \theta)$. Furthermore, since $x_1^*(\theta|f)$ is an optimum, we get that $\tilde{u}'(x_1^*(\theta|f), \theta) = -f'(x_1^*(\theta|f))$ for interior $x_1^*(\theta|f)$. Together, this implies that $\tilde{u}'(x_1^*(\theta|f), \theta') + f'(x_1^*(\theta|f)) > 0$. Thus, for small enough $\epsilon > 0$, we get that $\tilde{u}(x_1^*(\theta|f) + \epsilon, \theta') + f(x_1^*(\theta|f) + \epsilon) > \tilde{u}(x_1^*(\theta|f), \theta') + f(x_1^*(\theta|f))$ and so θ' would choose $x_1^*(\theta|f) + \epsilon$ over $x_1^*(\theta|f)$. Together with the previous result that $x_1^*(\theta'|f) \geq x_1^*(\theta|f)$, we conclude that $x_1^*(\theta'|f) > x_1^*(\theta|f)$. Again, the proof is identical to show that $x_1^*(\theta'|f) < x_1^*(\theta|f)$ if f is strictly decreasing.

For the second point, assume that the failure of the single crossing condition on $\tilde{u}'$ occurs by there being some $\tilde{x}_1 > x_1^*$ such that $\tilde{u}'(\tilde{x}_1, \theta') = \tilde{u}'(\tilde{x}_1, \theta)$ for some $\theta' > \theta$ with $x_1^*(\theta) = x_1^*(\theta') \equiv x_1^*$. We will then show that there exists a strictly increasing differentiable $f(x_1)$ such that $x_1^*(\theta|f) = x_1^*(\theta'|f) = \tilde{x}_1$. Specifically, for $f(x_1) = \Delta x_1$ with $\Delta = -\tilde{u}'(\tilde{x}_1, \theta)$, we get that $\tilde{u}'(\tilde{x}_1, \theta) + f'(\tilde{x}_1) = \tilde{u}'(\tilde{x}_1, \theta') + f'(\tilde{x}_1) = 0$. From the assumption that $u(x, \theta)$ is concave in x and $\mathcal{E}$ is convex, $\tilde{x}_1$ is the optimal choice of x_1 for both θ' and θ under the added incentive $f(x_1)$ and so $x_1^*(\theta|f) = x_1^*(\theta'|f) = \tilde{x}_1$. Again, the proof is nearly identical if the failure of the single crossing condition on $\tilde{u}'$ is in the other direction.

□

Finally, we follow up on Lemma A.2 to understand better how the condition that $\tilde{u}''(x_1, \theta') > \tilde{u}''(x_1, \theta)$ is increasing in x_1 depends on the cross-partials of the cost function. We have the following Lemma:

Lemma A.4. *Suppose that $c(e)$ is quadratic and consider any θ, θ' such that $x_1^*(\theta) = x_1^*(\theta')$ with $x_2^*(\theta) < x_2^*(\theta')$ and consider any weakly increasing screening function p. Then:*

- $x_1^*(\theta|p) \leq x_1^*(\theta'|p)$ *if* $\frac{\partial^2 c}{\partial e_1 \partial e_2} > 0$

- $x_1^*(\theta|p) \geq x_1^*(\theta'|p)$ *if* $\frac{\partial^2 c}{\partial e_1 \partial e_2} < 0$.

Proof. Using the Envelope theorem, we can derive the fact that:

$$\frac{\partial^2 \tilde{u}}{\partial x_1^2} = -\frac{1}{\theta_1^2} \cdot \left[\left(\frac{\partial^2 c}{\partial e_1^2} \frac{\partial^2 c}{\partial e_2^2} - \left(\frac{\partial^2 c}{\partial e_1 \partial e_2} \right)^2 \right) \left(\frac{\partial^2 c}{\partial e_2^2} \right)^{-1} \right]. \tag{27}$$

Comparing how this expression differs for individual θ and θ' is challenging. Since they will have different effort levels, it depends on how $\left(\frac{\partial^2 c}{\partial e_1^2} \frac{\partial^2 c}{\partial e_2^2} - \left(\frac{\partial^2 c}{\partial e_1 \partial e_2} \right)^2 \right) \left(\frac{\partial^2 c}{\partial e_2^2} \right)^{-1}$ varies with e_1, e_2, which in turn depend on the third derivatives of $c(e_1, e_2)$. However, note that if $c(e_1, e_2)$ is quadratic, then all the third derivatives are zero and so the expression is larger for (i.e., less negative) for θ' than θ if $\theta_1' > \theta_1$.

From Lemma A.9 it follows that $\theta_1' > \theta_1$ if $\frac{\partial^2 c}{\partial e_1 \partial e_2} > 0$ and $\theta_1' < \theta_1$ if $\frac{\partial^2 c}{\partial e_1 \partial e_2} < 0$. Thus, the conclusion follows. □

A.4 Other Supporting Lemmas

Lemma A.5. *Suppose that θ is continuously distributed. Then for every $\rho \in (0,1)$ and $\lambda \Delta v$ there exists a ξ such that under the policy $p(x_1) = \mathbf{1}(x_1 \geq \xi)$, we get that $\mathbb{E}_\Theta\big[p\big(x_1^*(\theta|p)\big)\big] = \rho$.*

Proof. Define

$$g(\xi) \equiv \mathbb{E}_\Theta\Big[\tilde{p}\big(x_1^*(\theta|\tilde{p})\big)\Big] = \mathbb{E}_\Theta\Big[\mathbf{1}\Big(\tilde{u}(\xi,\theta) + \lambda\Delta v - v(\theta) \geq 0\Big)\Big]. \tag{28}$$

Clearly $g(0) = 1$ and $g(\overline{e}_1\overline{\theta}_1) = 0$ and so if g is a continuous function of ξ, the intermediate value theorem implies the result. To show that g is continuous, we consider a sequence that converges to ξ, i.e., $\xi_n \to \xi$. If $g(\xi_n) \to g(\xi)$, then it follows that g is continuous.

Define $h(\theta) \equiv \mathbf{1}\Big(\tilde{u}(\xi,\theta) + \lambda\Delta v - v(\theta) \geq 0\Big)$ and $h_n(\theta) \equiv \mathbf{1}\Big(\tilde{u}(\xi_n,\theta) + \lambda\Delta v - v(\theta) \geq 0\Big)$. Note that in h_n – unlike $p_n(\theta)$ defined in Theorem A.1 – $\lambda\Delta v$ is held fixed in this sequence. Furthermore, define e_1^{max} to be the maximum feasible effort level, i.e., $e_1^{max} = sup\{e_1 | c(e) < \infty\}$.

Next, consider some θ such that $\tilde{u}(\tilde{x}_1,\theta) + \lambda\Delta v(\theta) - v(\theta) \equiv \epsilon \neq 0$. We first consider the case in which $\tilde{x}_1 < e_1^{max}\theta_1$. We then know that $\tilde{u}(\tilde{x}_1,\theta)$ is continuous in $\tilde{x}_1$ and so there exists a $\delta > 0$ such that $|\tilde{u}(\tilde{x}_1,\theta) - \tilde{u}(x_1,\theta)| < \frac{|\epsilon|}{2}$ for every x_1 such that $|x_1 - \tilde{x}_1| < \delta$. This implies that $\mathbf{1}\Big(\tilde{u}(x_1,\theta) + \lambda\Delta v(\theta) - v(\theta) \geq 0\Big) = \mathbf{1}\Big(\tilde{u}(\tilde{x}_1,\theta) + \lambda\Delta v(\theta) - v(\theta) \geq 0\Big)$ for every x_1 such that $|x_1 - \tilde{x}_1| < \delta$. It thus follows that there exists an N such that $h_n(\theta) = h(\theta)$ for all $n > N$.[49] In the second case in which $\tilde{x}_1 > e_1^{max}\theta_1$, then $\tilde{u}(\tilde{x}_1,\theta) = -\infty$ and so again,

[49] This is because $\tilde{x}_{1,n} \to \tilde{x}_1$ means that there exists some N such that $|\tilde{x}_{1,n} - \tilde{x}_1| < \delta$ for all $n > N$.

there exists an N such that $h_n(\theta) = h(\theta)$ for all $n > N$, since for all $n > N$ we get that $\tilde{x}_{1,n} > e_1^{max}\theta_1$.

We therefore get that $h_n(\theta) \to h(\theta)$ at every θ such that $\tilde{u}(\xi, \theta) + \lambda\Delta v - v(\theta) \neq 0$ and $\theta_1 e_1^{max} \neq \xi$. Furthermore, from the envelope condition, we get that $\tilde{u}(\xi, \theta) + \lambda\Delta v - v(\theta)$ is strictly increasing in θ_1 at every point where $\tilde{u}(\xi, \theta) + \lambda\Delta v - v(\theta) = 0$. This means there is at most one θ_1 for every θ_2 such that $\tilde{u}(\xi, \theta) + \lambda\Delta v - v(\theta) = 0$. Under the assumption that θ is continuously distributed, this implies that the set $\{\theta | \tilde{u}(\xi, \theta) + \lambda\Delta v - v(\theta) = 0 \cup \theta_1 e_1^{max} = \xi\}$ has zero measure. Thus, $h_n \to h$ pointwise almost everywhere. From the bounded convergence theorem, it follows that $g_n \to g$ and so g is continuous. □

Lemma A.6. *Define the sequence ξ_n and ξ^{max} as in Theorem A.1. Then $\xi_n \leq \xi^{max}$ for every n and $\xi_n \to \xi^{max}$.*

Proof. By definition, we have that for every n it is the case that $\mathbb{E}\Big[\mathbf{1}\big(\tilde{u}(\xi_n, \theta) + \lambda\Delta v_n - v(\theta) \geq 0\big)\Big] = \rho$. Since $\tilde{u}(\xi_n, \theta) + \lambda\Delta v_n - v(\theta)$ is increasing in $\lambda\Delta v$ and decreasing in ξ, it therefore must be the case that get that ξ_n is increasing in n. Furthermore, ξ_n is clearly bounded by ξ^{max}; otherwise $\tilde{u}(\xi_n, \theta)$ for more than $1 - \rho$ of individuals and so $\mathbb{E}\Big[\mathbf{1}\big(\tilde{u}(\xi_n, \theta) + \lambda\Delta v_n - v(\theta) \geq 0\big)\Big] < \rho$.

We can therefore conclude that ξ_n converges to a point, which we will denote $\tilde{\xi}$. We need only ensure that $\tilde{\xi} \geq \xi^{max}$. To do so, suppose instead that $\tilde{\xi} < \xi^{max}$. By definition, the fraction of individuals for whom $\tilde{\xi}$ is feasible, i.e., for whom $\tilde{u}(\tilde{\xi}, \theta)$ is finite, is strictly more than ρ and, since $\xi_n < \tilde{\xi}$ it would then be the cases that for each one of these individuals there exists an N such that $\forall n > N$ we'd get that $\mathbf{1}\big(\tilde{u}(\xi_n, \theta) + \lambda\Delta v_n - v(\theta)\big) = 1$. Thus, we would have that there would exist an N such that $\mathbb{E}\big[\mathbf{1}\big(\tilde{u}(\xi_n, \theta) + \lambda\Delta v_n - v(\theta)\big)\big] > \rho$, a contradiction given the definition of ξ_n.

□

Lemma A.7. *Define the function p_n and p_∞ as in Theorem A.1. Then $p_n \to p_\infty$ pointwise at almost every θ.*

Proof. Consider a point θ such that $\theta_1 > F_{\theta_1}^{-1}(1 - \rho)$, which by definition implies that $p_\infty(\theta) = 1$. We then want to show that there exists an N such that $\forall n > N$ we get that $p_n(\theta) = 1$. To do so, we note that:

$$\tilde{u}(\xi_n, \theta) + \lambda\Delta v_n - v(\theta) \geq \tilde{u}(\xi^{max}, \theta) + \lambda\Delta v_n - v(\theta)$$

and from the assumption that $\theta_1 > F_{\theta_1}^{-1}(1 - \rho)$ it follows that $\tilde{u}(\xi^{max}, \theta) - v(\theta)$ is finite. Since $\lambda\Delta v_n \to \infty$, it then follows that there exists an N such that $\forall n > N$ we get that $\lambda\Delta v_n > \tilde{u}(\xi^{max}, \theta) - v(\theta)$, which implies that $p_n(\theta) = 1$.

We next consider a point θ such that $\theta_1 < F_{\theta_1}^{-1}(1-\rho)$, which by definition implies that $p_\infty(\theta) = 0$. Now we want to show that there exists an N such that $\forall n > N$ we get that $p_n(\theta) = 0$. Since $p_\infty(\theta) = 0$, we know that $\theta_1 e_1^{max} < \xi^{max}$. Define $\xi^{max} - \theta_1 e_1^{max} = \epsilon$. Furthermore, since $\xi_n \to \xi^{max}$, we know that $\forall \epsilon > 0$ there exists an N such that $\forall n > N$ we have that $|\xi_n - \xi^{max}| < \epsilon$. Thus, there exists N such that $\forall n > N$ we have that $\theta_1 e_1^{max} < \xi_n$ and so $\tilde{u}(\xi_n, \theta) = -\infty$. Thus, regardless of how large $\lambda \Delta v$ is, we have that $\tilde{u}(\xi_n, \theta) + \lambda \Delta v_n - v(\theta) = -\infty$ and so $p_n(\theta) = 0$.

□

Lemma A.8. *For $k \in \{1, 2\}$, we have that $x_k^*(\theta)$ and $x_k^*(\theta|p)$ are both increasing in θ_k. They are also both decreasing in θ_k' for $k' \neq k$ if $\frac{\partial^2 c}{\partial e_1 \partial e_2} > 0$ and increasing in θ_k' if $\frac{\partial^2 c}{\partial e_1 \partial e_2} < 0$.*

Proof. Ignoring, at first, the incentive to stay we can write the utility function as:

$$u(e, \theta) = e_1\theta_1 + e_2\theta_2 - c(e_1, e_2). \tag{29}$$

Looking pairwise, it follows that $u(e, \theta)$ is supermodular in $(e_1, -e_2, \theta_1, -\theta_2)$ if $\frac{\partial^2 c}{\partial e_1 \partial e_2} > 0$. Thus, we get that $e_1^*(\theta)$ is increasing in θ_1 and decreasing in θ_2, which implies that $x_1^*(\theta)$ is increasing in θ_1 and decreasing in θ_2.

Similarly, if $\frac{\partial^2 c}{\partial e_1 \partial e_2} < 0$, then $u(e, \theta)$ is supermodular in $(e_1, e_2, \theta_1, \theta_2)$. Thus, we get that $e_1^*(\theta)$ is increasing in both θ_1 and θ_2, which implies that $x_1^*(\theta)$ is increasing in both θ_1 and θ_2.

Finally, since adding $\lambda \Delta v p(e_1\theta_1)$ to $u(e, \theta)$ does not affect the supermodularity of the function it likewise does not affect the comparative statics.

□

Lemma A.9. *There exists an inverse function $\theta(x^*) : X^* \to \Theta$ such that $\theta(x^*(\theta)) = \theta$ with the following comparative statics:*

- *$\theta_k(x^*)$ is increasing in x_k^* for $k \in \{1, 2\}$.*
- *If $\frac{\partial^2 c}{\partial e_1 \partial e_2} > 0$ then $\theta_k(x^*)$ is increasing in $x_k'^*$ for $k \in \{1, 2\}$ and $k' \neq k$.*
- *If $\frac{\partial^2 c}{\partial e_1 \partial e_2} < 0$ then $\theta_k(x^*)$ is decreasing in $x_k'^*$ for $k \in \{1, 2\}$ and $k' \neq k$.*

Proof. To show that the inverse function exists, we start by noting that the Jacobian of $x^*(\theta)$– which we denote as $D_\theta x^*(\theta)$ – can be written as follows:

$$D_\theta x^*(\theta) = \begin{bmatrix} e_1^*(\theta) + \theta_1 \frac{\partial e_1^*(\theta)}{\partial \theta_1} & \theta_1 \frac{\partial e_1^*(\theta)}{\partial \theta_2} \\ \theta_2 \frac{\partial e_2^*(\theta)}{\partial \theta_1} & e_2^*(\theta) + \theta_2 \frac{\partial e_2^*(\theta)}{\partial \theta_2} \end{bmatrix} \tag{30}$$

Noting the determinant of some matrix A as $det(A)$, some algebra shows that $det\big(D_\theta x^*(\theta)\big) > \theta_1\theta_2 det\big(D_\theta e^*(\theta)\big)$. The fact that inequality is strict stems from our assumption that everyone has some ability to increase each output, i.e., $\underline{\theta}_k > 0$ for $k \in \{1,2\}$, and that everyone exerts at least some effort on each task, i.e., $\underline{e}_k > 0$ for $k \in \{1,2\}$. Thus, the determinant of $D_\theta x^*(\theta)$ is strictly positive and so there is an inverse function.

Once we know that the inverse function exists and $det\big(D_\theta x^*(\theta)\big) > 0$, the comparative statics follow directly from the inverse function theorem and the formula for inverting a 2x2 matrix. In particular, the inverse function theorem says that $D_{x^*}\theta(x^*) = \Big(D_\theta x^*(\theta)\Big)^{-1}$. Using the formula for inverting a 2x2 matrix, we therefore have that

$$D_{x^*}\theta(x^*) = \frac{1}{det\big(D_\theta x^*(\theta)\big)} \begin{bmatrix} e_2^*(\theta) + \theta_2 \frac{\partial e_2^*(\theta)}{\partial \theta_2} & -\theta_1 \frac{\partial e_1^*(\theta)}{\partial \theta_2} \\ -\theta_2 \frac{\partial e_2^*(\theta)}{\partial \theta_1} & e_1^*(\theta) + \theta_1 \frac{\partial e_1^*(\theta)}{\partial \theta_1} \end{bmatrix}. \tag{31}$$

The proof then follows from the fact that $det\big(D_\theta x^*(\theta)\big) > 0$ and the results – obtained in the same manner as Lemma A.8 or via the implicit function theorem – that:

- $\frac{\partial e_k^*(\theta)}{\partial \theta_k} > 0$ for $k \in \{1,2\}$.
- If $\frac{\partial^2 c}{\partial e_1 \partial e_2} > 0$ then $\frac{\partial e_k^*(\theta)}{\partial \theta_k'} < 0$ when $k' \neq k$.
- If $\frac{\partial^2 c}{\partial e_1 \partial e_2} < 0$ then $\frac{\partial e_k^*(\theta)}{\partial \theta_k'} > 0$ when $k' \neq k$.

□

Lemma A.10. *Suppose that $c(e)$ is quadratic. Then $x_1^*(\theta|p)$ is increasing in x_1^*.*

Proof. Consider θ, θ' with $x_1^*(\theta') > x_1^*(\theta)$ and $x_2^*(\theta') = x_2^*(\theta)$. From Lemma A.9, we have that $\theta_1' > \theta_1$, which using Equation 27 and the fact that $c(e)$ is quadratic, implies that $\tilde{u}''(x_1, \theta') - \tilde{u}''(x_1, \theta)$ is increasing in x_1. Since $x_1^*(\theta') > x_1^*(\theta)$, we also get that $\tilde{u}'(x_1^*(\theta), \theta') > \tilde{u}'(x_1^*(\theta), \theta)$ and since $x_1^*(\theta|p) \geq x_1^*(\theta)$ we have that $\tilde{u}'(x_1, \theta') > \tilde{u}'(x_1, \theta)$ for all $x_1 \in [x_1^*(\theta), x_1^*(\theta|p)]$. It therefore follows that $x_1^*(\theta'|p) \geq x_1^*(\theta|p)$, using the arguments in Subsection A.3. □

ONLINE APPENDIX: NOT FOR PUBLICATION

B Value-Added Estimation

In this appendix, we describe the different forms of value-added we use in the paper and how we estimate them. Our estimation procedure follows Mulhern and Opper (2021), although Mulhern and Opper (2021) does not control for experience in their estimates.

B.1 Residualizing Outcomes

Let i index students, j index teachers, c index classrooms, and t index years. Let $\tau()$ be a function that describes when an outcome is realized. For contemporaneous outcomes, $\tau(k) = 0$, while for outcomes realized in the future, like next year's test scores, $\tau(k) > 0$. Our statistical model of outcomes, for a specific subject-level, is:

$$y_{i,t+\tau} = \Lambda X'_{it} + \sum_{e'} \rho_{e'} \mathbb{1}\{e_{jt} = e'\} + \mu_{jt} + \nu_{ct} + \phi_{c',t+1}\mathbb{1}(\tau \geq 1) + \phi_{c'',t+2}\mathbb{1}(\tau = 2) + \epsilon_{it} \quad (32)$$

where e_{jt} is a teacher's experience level (with all teachers with six or more years of prior experience grouped into one level).

We have 4 types of outcomes:

1. **Targeted outcome**: test scores in year t
2. **Untargeted outcomes**: test scores in year $t+1$, test scores in year $t+2$, attendance rate in year t, attendance rate in year $t+1$, grades in tested subject in year t, grades in tested subject in year $t+1$, grades in untested subjects in year t, grades in untested subjects in year $t+1$
3. **Index of untargeted outcome**: an index of the above outcomes (constructed below)
4. **Long-term outcome**: whether the student graduates high school on-time

For ease of exposition, label the four outcomes (at the student-level) y^1_{it}, $\vec{y}^2_{it}$, y^3_{it}, y^4_{it}.

In a first step, we standardize each outcome in y^1_{it} and $\vec{y}^2_{it}$ to have mean 0 and standard deviation 1 for each grade-year in NYC.

We then residualize outcomes in y^1_{it}, $\vec{y}^2_{it}$, and y^4_{it} by regressing them on a set of observable characteristics and teacher fixed effects:

$$y_{i,t+\tau} = \Lambda X'_{it} + \sum_{e'} \rho_{e'} \mathbb{1}\{e_{jt} = e'\} + \mu_j + \upsilon_{it}. \quad (33)$$

where $\upsilon_{it} = \mu_{jt} - \mu_j + \nu_{ct} + \phi_{c',t+1}\mathbb{1}(\tau \geq 1) + \phi_{c'',t+2}\mathbb{1}(\tau = 2) + \epsilon_{it}$. We run separate regressions for each outcome, subject (math or ELA), and level (elementary or middle) combination.

We let the set of controls, X'_{it}, vary by outcome. For all outcomes, we include year dummy variables and indicators for whether the student receives free or reduced price lunch and whether the student is an English language learner, male, Black, Hispanic, and Asian. For lagged outcomes we use:

- Cubic polynomials in $t-1$ test scores for each subject – used for test scores in t, $t+1$, $t+2$, subject grades in t and $t+1$, untested subject grades in t and $t+1$, graduation
- Cubic polynomial in $t-1$ attendance rate – used for attendance rate in t and $t+1$.

For each student, we construct two residuals:

1. $\hat{\upsilon}^1_{it} = y_{i,t+\tau} - \hat{\Lambda}X'_{it} - \sum_{e'} \hat{\rho}_{e'}\mathbb{1}\{e_{jt} = e'\}$
2. $\hat{\upsilon}^2_{it} = y_{i,t+\tau} - \hat{\Lambda}X'_{it}$

The residuals differ in whether the teacher's experience effects are included.

B.2 Constructing Measures of Teacher Output in Each Year

We then construct two (noisy) measures of a teacher's output in year t (for each subject-level) by taking the mean of the two student residuals over each teacher-year combination:

1. $\hat{\mu}^1_{jkt} = \frac{1}{N_{jkt}} \sum_{i \in \mathcal{I}_{jkt}} \hat{\upsilon}^1_{ikt}$
2. $\hat{\mu}^2_{jkt} = \frac{1}{N_{jkt}} \sum_{i \in \mathcal{I}_{jkt}} \hat{\upsilon}^2_{ikt}$

where $\mathcal{I}_{jkt}$ is the set of N_{jkt} students with outcome k who are taught by j in year t. We construct these measures for each outcome in y^1_{it}, $\vec{y}^2_{it}$, and y^4_{it}.

For analysis in Section V we use the measure that includes experience effects ($\hat{\mu}^2_{jkt}$) when it is the outcome variable. The exception is Figure 7, where we use the version without experience effects ($\hat{\mu}^1_{jkt}$) to show the flatness of the curve for unexposed cohorts. For analysis in Sections VI and VII, we use the version without experience effects ($\hat{\mu}^1_{jkt}$) as the outcome because we project it onto shrunken value-added measures that exclude the experience profile.

B.3 Constructing Forecasts of Teacher Output

The prior measures are noisy estimates of a teacher's realized output in a given year. For classifying teachers according to their unincentivized output, we construct forecasts that

incorporate data from multiple years. We construct forecasts for each outcome in y_{it}^1 and $\vec{y}_{it}^2$.

We follow Mulhern and Opper (2021) and refer there for the details. The key estimation points are:

- The estimates are from a joint Empirical Bayes procedure where the estimates are shrunken jointly.
- We estimate using data from unincentivized periods only. We produce estimates for all years as if they were unincentivized (even if they were actually incentivized).
- We estimate using the non-experience residuals ($\hat{\mu}_{jkt}^1$).
- We allow for the non-experience component of a teacher's effect to drift over time. We let drift rates vary depending on the difference in years between measures, where we estimate a constant drift rate for year differences at least 3.
- We keep teacher-subject-levels with at least ten students with an outcome.
- In forecasting a teacher's output in year t, we use data from all years except year t (i.e., a leave-out estimator).
- For each jt, some of the output measures may be missing. In these cases, we predict the missing measures with forecasts based on the non-missing measures. Specifically, we estimate a separate joint Empirical Bayes procedure for each combination of non-missing measures and use it to forecast the missing measures. The identifying assumption is that the missingness is random conditional on the forecast of the non-missing measures.
- Our inclusion of $\phi_{c',t+1}\mathbb{1}(\tau \geq 1) + \phi_{c'',t+2}\mathbb{1}(\tau = 2)$ in the model means that the correlation structure of a teacher's students' residuals varies with the overlap in their classes in $t+1$ and $t+2$. We incorporate this in constructing the forecasts.

We denote these forecasts as $\tilde{\mu}_{jkt}$ and use them in Sections VI and VII. The exception is the first four columns in Table 8, where we show how treatment effects vary with heterogeneity in *only* the targeted or untargeted forecasts, rather than jointly. For these columns, we construct forecasts from Emprical Bayes procedures that only include the relevant outcomes (y_{it}^1 or $\vec{y}_{it}^2$).

B.4 Constructing the Index of Untargeted Output

We create the untargeted index by anchoring the measures to their relative predictiveness of whether a student graduates from high school:

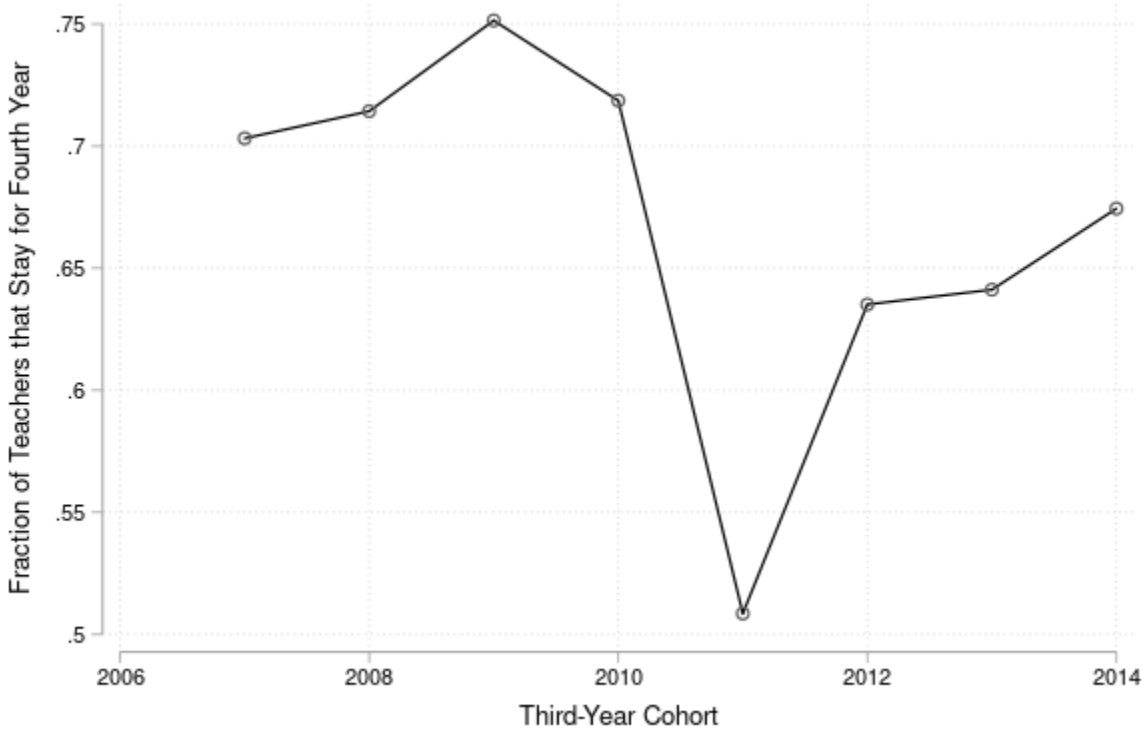
$$\hat{\mu}^{1}_{jlt} = \omega' \tilde{\mu}_{jt} + \upsilon_{ij}, \tag{34}$$

where l corresponds to the graduation outcome and ω is a vector of anchoring weights. We estimate using data from the unincentivized period.

We use the estimated weights to construct two measures: targeted output ($\tilde{\mu}^{T}_{jt} = \tilde{\mu}_{j1t}$) and an index of untargeted output ($\tilde{\mu}^{U}_{jt} = \frac{1}{\hat{\omega}_1} \sum_{k=2}^{K} \hat{\omega}_k \tilde{\mu}_{jkt}$). We also apply these weights to the unshrunken measures for further indices, $\hat{\mu}^{T}_{jt} = \hat{\mu}_{jkt}$ and $\hat{\mu}^{U}_{jt} = \frac{1}{\hat{\omega}_1} \sum_{k=2}^{K} \hat{\omega}_k \hat{\mu}_{jkt}$.

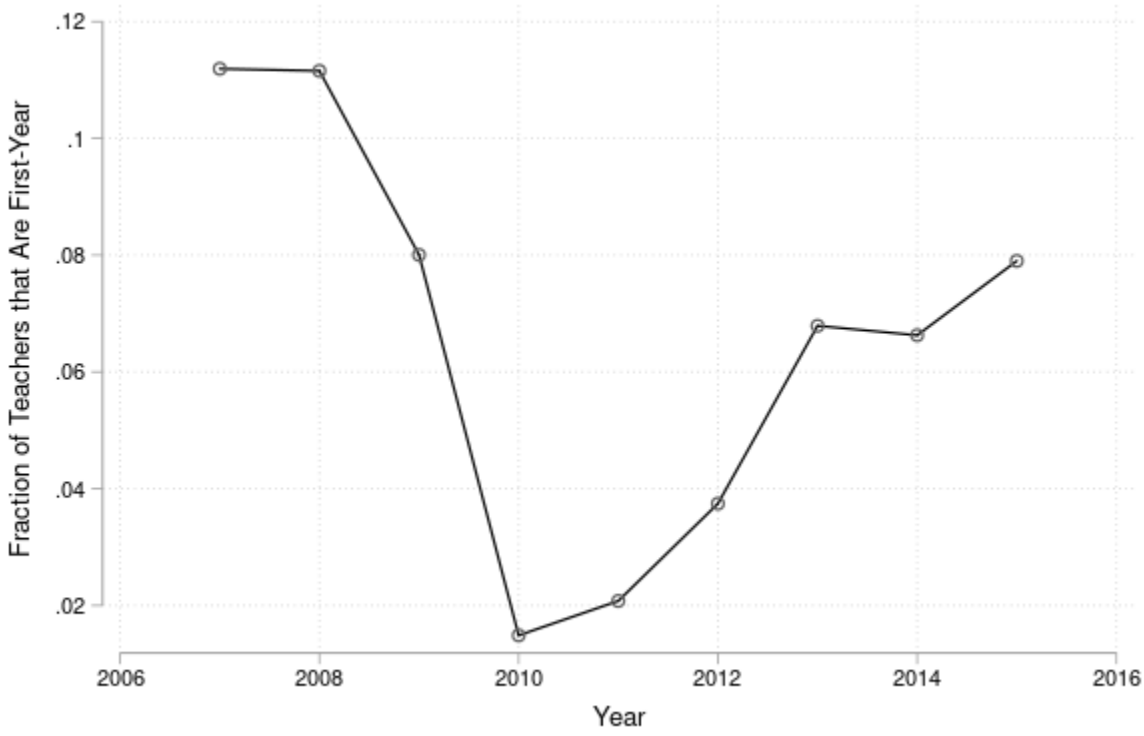
C Appendix Figures

Figure A1: Teachers' Persistence to Fourth Year of Teaching



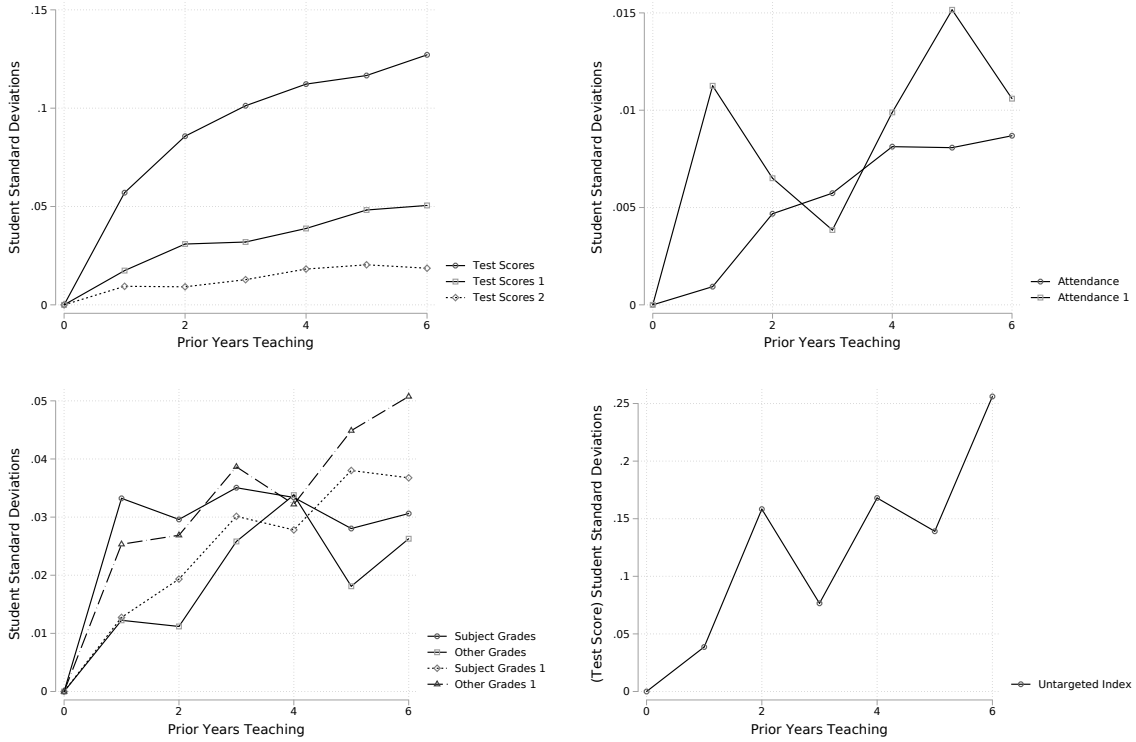
Note: This figure shows the fraction of teachers who were in the district in their third year of teaching who remain in the district their fourth year. The x-axis classifies cohorts based on the academic year when their third year of experience occurred.

Figure A2: New Teachers' Fraction of Workforce



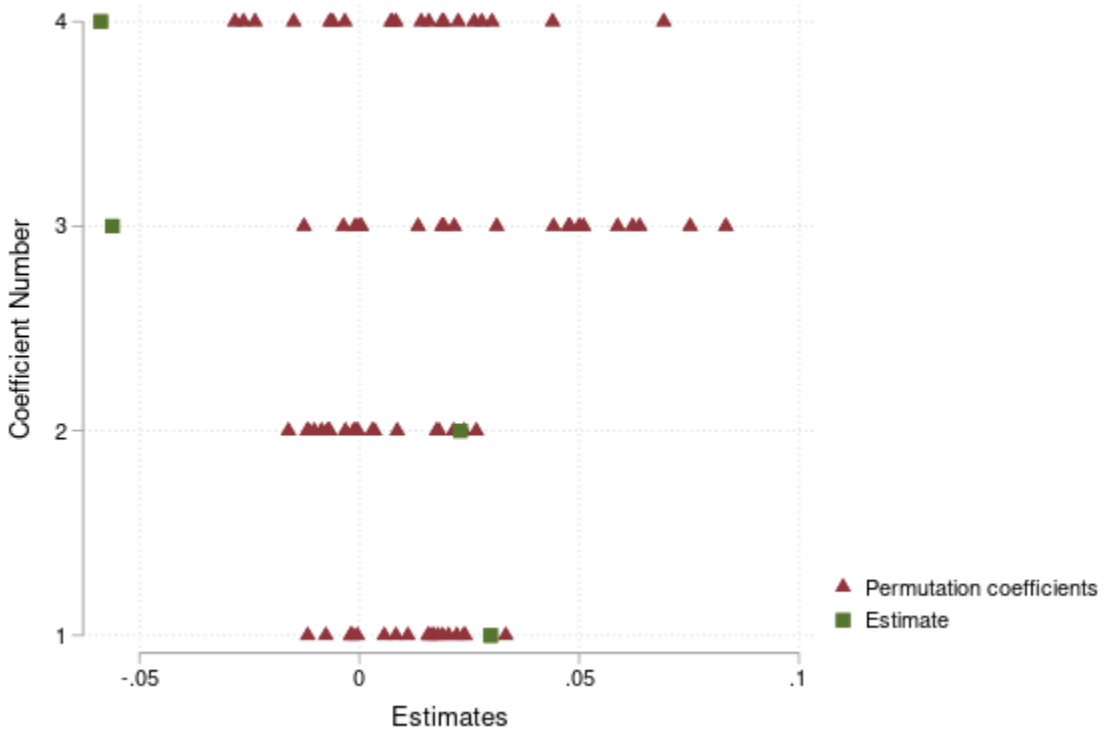
Note: This figure shows the fraction of each year's teacher workforce that first-year teachers comprise in NYC.

Figure A3: Experience profiles



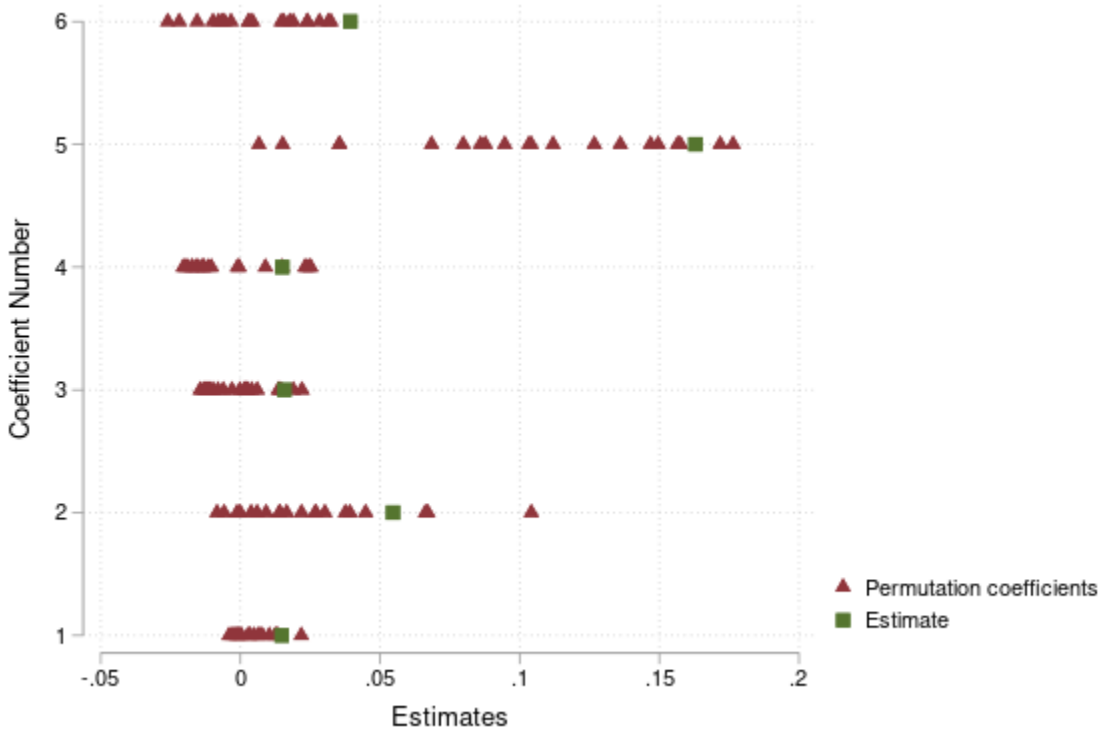
Note: This figure shows the estimated experience profile for our measures, where output in the first year of teaching is normalized to 0 and teachers with six or more years of experience are grouped into one category. The top left panel shows the score measures (current and future), the top right panel shows the attendance measures, the bottom left panel shows the grades measures, and the bottom right panel shows the untargeted index. In the first three panels, units are student standard deviations on each measure. In the bottom right panel, the units are student standard deviations on test scores.

Figure A4: Permutation Tests for Table 4 Estimates



Note: This figure shows permutation tests for the coefficients on "Incentive" in Table 4. The labeled rows correspond to the columns in Table 4. For each test, we assign the policy change to a different set of cohorts. In the correctly specified timing, the 2010 policy affects the cohorts with 0-2 years of prior experience. In the placebo timings, we let the policy affect cohorts with 3-5, 4-6, 5-7, etc. years of prior experience (up to 22-24). We maintain the structure of the policy change and the sample restrictions we impose in our main analysis. The correctly specified regression is labeled with a green square while the placebo estimates are red triangles.

Figure A5: Permutation Tests for Table 8 Estimates



Note: This figure shows permutation tests for the coefficients in columns (5) and (6) of Table 8. Labeled rows (1)-(3) correspond to the estimate in column (5) in Table 8 (from top to bottom). Labeled rows (4)-(6) correspond to the estimate in column (6) in Table 8 (from top to bottom). For each test, we assign the policy change to a different set of cohorts. In the correctly specified timing, the 2010 policy affects the cohorts with 0-2 years of prior experience. In the placebo timings, we let the policy affect cohorts with 3-5, 4-6, 5-7, etc. years of prior experience (up to 22-24). We maintain the structure of the policy change and the sample restrictions we impose in our main analysis. The correctly specified regression is labeled with a green square while the placebo estimates are red triangles.

D Appendix Tables

Table A1: Correlation between Value-Added in t and $t-1$

	Corr b/t VA t and VA t-1	Corr b/t VA t and Test Score VA t-1	Corr b/t VA t and Index VA t-1
Test Score	0.436	0.436	0.161
Untargeted Index	0.560	0.160	0.560
Test Score t+1	0.357	0.196	0.201
Test Score t+2	0.446	0.118	0.249
Attendance	0.553	0.024	-0.121
Attendance t+1	0.265	0.028	0.052
Subject Grades	0.535	0.046	0.121
Other Grades	0.565	0.102	0.344
Subject Grades t+1	0.290	0.045	0.178
Other Grades t+1	0.453	0.074	0.357

This table shows correlations between a teacher's (shrunken) value-added measure in t and various (unshrunken) value-added measures in $t-1$. The columns show correlations with lagged value-added in (1) the same outcome, (2) the targeted measure ("Test Score"), and (3) the index of untargeted measures. We include the targeted measure ("Test Score"), the index of untargeted measures, and each untargeted measure separately.

Table A2: Outcomes' Univariate Relationship to Graduation

	Grad	Grad	Grad	Grad	Grad	Grad	Grad	Grad	Grad
Test Score VA	0.0879***								
	(0.00727)								
Test Score 1 VA		0.0635***							
		(0.00631)							
Test Score 2 VA			0.0716***						
			(0.00798)						
Attendance VA				0.0868***					
				(0.0119)					
Attendance 1 VA					0.0358***				
					(0.00611)				
Subject Grade VA						0.0618***			
						(0.00435)			
Other Grade VA							0.0697***		
							(0.00339)		
Subject Grade 1 VA								0.0923***	
								(0.00617)	
Other Grade 1 VA									0.106***
									(0.00481)
N Teachers	11689	11694	11689	11695	11695	11692	11670	11678	11676
Mean DV	-0.0154	-0.0154	-0.0153	-0.0154	-0.0155	-0.0153	-0.0153	-0.0153	-0.0153
N	32066	32089	32044	32157	32149	32078	31937	31997	31991

This table shows the univariate regression of (the residual of) whether a student graduated from high school on the targeted forecasted measure or on each untargeted forecasted measure. The forecasts come from our multi-year multi-dimensional value-added model that is estimated on unincentivized periods only. Forecasts are constructed for all periods and leave out data from that year. The regression includes only observations from the unincentivized period. In the variable labels, the number indicates the number of years in the future the outcome is realized. All variables are in (separate) standard deviation units.

Table A3: Anchoring Outcomes to Graduation

	Grad
Test Score VA	0.0334***
	(0.0113)
Test Score 1 VA	0.0300*
	(0.0173)
Test Score 2 VA	0.0246*
	(0.0129)
Attendance VA	0.231***
	(0.0211)
Attendance 1 VA	0.0134
	(0.0156)
Subject Grade VA	-0.0149**
	(0.00693)
Other Grade VA	0.0352***
	(0.00549)
Subject Grade 1 VA	-0.0295**
	(0.0142)
Other Grade 1 VA	0.111***
	(0.0108)
N Teachers	11670
Mean DV	-0.0153
N	31937

This table shows the regression of (the residual of) whether a student graduated from high school on the targeted and untargeted forecasted measures. The forecasts come from our multi-year multi-dimensional value-added model that is estimated on unincentivized periods only. Forecasts are constructed for all periods and leave out data from that year. The regression includes only observations from the unincentivized period. In the variable labels, the number indicates the number of years in the future the outcome is realized. All variables are in (separate) standard deviation units.

Table A4: Principal Component Analysis

	First Component	Second Component
Test Score 1 VA	0.156	0.376
Test Score 2 VA	0.151	0.346
Attendance VA	0.015	0.124
Attendance 1 VA	0.051	0.282
Subject Grades VA	0.422	-0.208
Other Grades VA	0.804	-0.350
Subject Grades 1 VA	0.186	0.450
Other Grades 1 VA	0.302	0.525

This table shows the first two components of a Principal Component Analysis on the value-added on each untargeted measure.

Table A5: Effect of Policy Change on Probationary Period Output – By Subject and Level

	Score	Score	Score	Score	Index	Index	Index	Index
Incentive	0.0148	0.0160*	0.0226*	0.00872	-0.0657*	-0.0578	-0.0653*	-0.0669
	(0.0103)	(0.00957)	(0.0130)	(0.00906)	(0.0386)	(0.0387)	(0.0372)	(0.0513)
Fixed Effects	Teacher	Teacher	Teacher	Teacher	Teacher	Teacher	Teacher	Teacher
Sample	Math	ELA	Elem	Middle	Math	ELA	Elem	Middle
N Teachers	13184	13868	13389	9909	11961	12494	9036	8030
Mean DV	0.123	0.0810	0.121	0.0542	0.0290	0.0255	-0.0152	0.0915
N	57796	59586	84107	42934	48665	49973	67931	32268

This table shows the causal effect of the tenure policy change on targeted and untargeted output in the probationary period. The columns show the effects in different subsamples, split by the tested subject (Math or ELA) and the level of school (elementary or middle). All regressions include teacher fixed effects. An observation is a teacher-subject-year. Standard errors are clustered by teacher. The sample covers years 2006 to 2014. The teachers with more than 3 years of experience are only included if they finished the standard probationary period before the tenure policy change. All outcome units are test score student standard deviations.

Table A6: Effect of Policy Change on Specific Untargeted Outcomes, Cohort Fixed Effects

	Score 1	Score 2	Attend	Attend 1	Grades	Other Grades	Grades 1	Other Grades 1
Incentive	-0.00572	0.0209	0.00491**	0.000419	0.0139	0.0218	-0.0351*	-0.0188
	(0.00981)	(0.0137)	(0.00212)	(0.00900)	(0.0186)	(0.0159)	(0.0181)	(0.0175)
Teacher FEs	No	No	No	No	No	No	No	No
Mean DV	0.0496	0.0548	0.00417	0.0314	0.0251	-0.00873	0.0398	0.0409
N	110038	87893	132253	110572	36847	48078	60916	74374

This table shows the causal effect of the tenure policy change on individual untargeted (residualized) outcomes in the probationary period. The "1" or "2" in the column headers indicate the measure's number of years into the future. "Grades" are in the tested subject while "Other" grades are in untested subjects. All variables are standardized at the grade-year level to have mean 0 and standard deviation 1 in the full population. All regressions include cohort fixed effects. An observation is a teacher-subject-year. Standard errors are clustered by teacher. The sample covers years 2006 to 2014. The teachers with more than 3 years of experience are only included if they finished the standard probationary period before the tenure policy change.

Table A7: Effect of Policy Change on Specific Untargeted Raw Outcomes, Teacher Fixed Effects

	Score	Score 1	Score 2	Attend	Attend 1	Grades	Other Grades	Grades 1	Other Grades 1
Incentive	0.0135	-0.0252	-0.0151	0.00435	-0.0242*	0.0337	0.000386	-0.0652***	-0.0429*
	(0.0133)	(0.0156)	(0.0192)	(0.0109)	(0.0134)	(0.0214)	(0.0199)	(0.0230)	(0.0219)
Teacher FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean DV	-0.0533	-0.0483	-0.0432	-0.0445	-0.0429	-0.0741	-0.102	-0.0613	-0.0719
N	127678	106647	85244	127480	106756	33217	43512	57508	71371

This table shows the causal effect of the tenure policy change on individual untargeted (non-residualized) outcomes in the probationary period. The "1" or "2" in the column headers indicate the measure's number of years into the future. "Grades" are in the tested subject while "Other" grades are in untested subjects. All variables are standardized at the grade-year level to have mean 0 and standard deviation 1 in the full population. All regressions include teacher fixed effects. An observation is a teacher-subject-year. Standard errors are clustered by teacher. The sample covers years 2006 to 2014. The teachers with more than 3 years of experience are only included if they finished the standard probationary period before the tenure policy change.

Table A8: Effect of Policy Change on Outcomes, Controlling for Treatment Status of Next Two Years' Teacher

	Untargeted Index	Untargeted Index	Untargeted Index	Untargeted Index	Score 2	Score 2
Incentive	-0.0749*	-0.0794**	-0.0837**	-0.0837**	-0.00476	-0.00582
	(0.0389)	(0.0389)	(0.0364)	(0.0364)	(0.0135)	(0.0135)
Incentive 1		-0.00913		-0.0102		-0.0267***
		(0.0253)		(0.0198)		(0.00814)
Incentive 2		-0.170***		0.00720		-0.0163**
		(0.0239)		(0.0168)		(0.00740)
Fixed Effects	Cohort	Cohort	Teacher	Teacher	Teacher	Teacher
N Teachers	11324	11324	11324	11324	11202	11202
Mean DV	0.0487	0.0487	0.0487	0.0487	0.0277	0.0277
N	63796	63796	63796	63796	63361	63361

This table shows how our estimates of the effect of the policy change on teachers' untargeted outcomes varies depending on whether we control for the treatment status of the teachers in the two subsequent years. Adjacent columns show the regression with and without these controls, where we restrict the samples to the teachers for whom we can classify the two subsequent teachers. "Score 2" is the test score in year $t+2$ and is in test score $t+2$ student standard deviation units while the untargeted index includes all of the untargeted outcomes and is in test sore t student standard deviation units. "Incentive 1" is the fraction of teacher j's students in year t that have an incentivized $t+1$ teacher, and "Incentive 2" is the fraction that have an incentivized $t+2$ teacher.

Table A9: Effect of Policy Change on Outcomes, Controlling for Treatment Status of Next Year's Teacher

	Score 1	Score 1	Attend 1	Attend 1	Grades 1	Grades 1	Other Grades 1	Other Grades 1
Incentive	-0.0262**	-0.0264**	-0.0348***	-0.0346***	-0.0523**	-0.0508**	-0.0428**	-0.0424**
	(0.0116)	(0.0116)	(0.0114)	(0.0114)	(0.0235)	(0.0235)	(0.0201)	(0.0201)
Incentive 1		-0.0116*		0.00868		0.0522***		0.0169**
		(0.00672)		(0.00625)		(0.0154)		(0.00848)
Fixed Effects	Teacher	Teacher	Teacher	Teacher	Teacher	Teacher	Teacher	Teacher
N Teachers	11299	11299	11310	11310	7407	7407	8231	8231
Mean DV	0.0266	0.0266	0.00370	0.00370	0.0234	0.0234	0.0276	0.0276
N	63719	63719	63726	63726	31770	31770	36882	36882

This table shows how our estimates of the effect of the policy change on teachers' untargeted outcomes varies depending on whether we control for the treatment status of the teacher in the subsequent year. Adjacent columns show the regression with and without these controls, where we restrict the samples to the teachers for whom we can classify the subsequent teacher. All outcomes are realized in year $t+1$. Units are student standard deviations for the respective outcome. "Incentive 1" is the fraction of teacher j's students in year t that have an incentivized $t+1$ teacher.

Table A10: Effect of Policy Change on Outcomes, Controlling for Identity of Next Year's Teacher

	Untargeted Index	Untargeted Index	Untargeted Index
Incentive	-0.0399***	-0.0386**	-0.0349*
	(0.0152)	(0.0152)	(0.0180)
Fixed Effects	Teacher	Teacher, Future Teacher	Teacher, Future Teacher-Year
N Teachers	16195	16177	16134
Mean DV	0.117	0.117	0.117
N	1127279	1126039	1122822

This table shows how our estimates of the effect of the policy change on the index of teachers' untargeted outcomes varies depending on whether we control for the identity of the teacher in the subsequent year. Adjacent columns show the regression with current year teacher fixed effects, with current and subsequent year teacher fixed effects, and with current and subsequent (year-specific) year teacher fixed effects, where we restrict the samples to the teachers for whom we can classify the subsequent teacher. All outcomes are realized in year $t+1$. Standard errors are clustered by teacher.

Table A11: Main Results, Using First Principal Component for Untargeted Measures

	Untargeted (PCA)	Untargeted (PCA)	Untargeted (PCA)	Score	Score
Incentive	-0.0822*	-0.0580	-0.0312	0.0101	-0.000103
	(0.0425)	(0.0402)	(0.0362)	(0.00776)	(0.0103)
Post-Incentive			0.0481		
			(0.0486)		
Incentive * Targeted VA				0.0613	0.159**
				(0.0516)	(0.0725)
Incentive * Untargeted VA (PCA)				0.0154	0.0794***
				(0.0187)	(0.0288)
Fixed Effects	Cohort	Teacher	Teacher	Cohort	Teacher
N Teachers	5967	3781	3897	16755	16755
Mean DV	-0.0460	-0.0541	-0.0503	0.00589	0.00589
N	14491	12305	12837	84288	84288

This table shows how our main results change if instead of using the untargeted measure index anchored to graduation rates we use the first principal component. The first two columns correspond to columns (3) and (4) of Table 4. The third column corresponds to column (2) of Table 7. The last two columns correspond to columns (5) and (6) of Table 8. Standard errors are clustered by teacher.

Table A12: Main Results, Restricting to Teachers with No Missing Measures

	Untargeted Index	Untargeted Index	Untargeted Index	Score	Score
Incentive	-0.0874	-0.110	-0.0897	0.0170	0.0220
	(0.0732)	(0.0832)	(0.0785)	(0.0175)	(0.0258)
Post-Incentive			-0.0225		
			(0.102)		
Incentive * Targeted VA				-0.160*	0.122
				(0.0960)	(0.177)
Incentive * Untargeted VA				0.0191	0.0760**
				(0.0191)	(0.0386)
Fixed Effects	Cohort	Teacher	Teacher	Cohort	Teacher
N Teachers	6366	3245	3338	4646	4646
Mean DV	0.204	0.206	0.209	-0.0243	-0.0243
N	13041	9920	10332	10021	10021

This table shows how our main results change if we restrict our sample to teachers with no missing measures instead of predicting missing measures with non-missing measures. The first two columns correspond to columns (3) and (4) of Table 4. The third column corresponds to column (2) of Table 7. The last two columns correspond to columns (5) and (6) of Table 8. Standard errors are clustered by teacher.

Table A13: Voluntary Attrition by Targeted and Untargeted Value-Added

	Voluntary Attrition
Targeted VA	-0.278***
	(0.0430)
Untargeted VA	-0.0210**
	(0.00914)
N Teachers	5834
Mean DV	0.394
N	8693

This table shows how teachers' voluntary attrition rates vary with their targeted and untargeted forecasted value-added. We consider tenured teachers in their seventh year of teaching in the district and determine whether the teachers left the sample before the end of our data. If so, we label them as voluntary attrition. Targeted and untargeted forecasted value-added are estimated using data from unincentivized periods only and are both in student test score standard deviation units.

Table A14: Output under Different Screening Regimes – 2006 Cohort

	Obs.	Mean Targeted Output	Mean Untargeted Output	Mean Total Output
Teachers' Tenure under Different Responses				
Never Tenured	477	-0.119	-0.771	-0.890
Only Tenured w/o Behavioral Response	50	-0.009	-1.262	-1.271
Only Tenured w/ Behavioral Response	50	-0.041	0.239	0.198
Always Tenured	1,024	0.111	-0.222	-0.111
Tenured Teachers under Different Policies				
Screening w/o Behavioral Response	1,074	0.106	-0.271	-0.165
Screening w/ Behavioral Response	1,074	0.104	-0.201	-0.097
(Infeasible) Screening on Both Dimensions	1,074	0.065	0.010	0.074
Gains Relative to Infeasible First-Best				
Gains (Fraction)				0.285

This table shows the mean output for different groups of teachers and under different policies. The sample is teachers who started in the district in 2006. The top panel splits teachers into four groups based on whether they would receive tenure in a regime without a behavioral response and whether they would receive tenure in a regime with a behavioral response. The middle panel shows the mean output associated with the set of teachers receiving tenure under different policies. The first two policies are screening on the targeted measure, without and with a behavioral response. The last policy is an infeasible policy screening on the sum of output across both dimensions. The final panel shows the fraction of gains the behavioral response achieves relative to the distance between the screening without behavioral response regime and the infeasible policy screening on the sum of output. "Mean Targeted" output is the mean forecasted test score value-added. "Mean Untargeted Output" is the mean forecasted value-added on the untargeted index. "Mean Total Output" is the sum of the mean forecasted value-added across the targeted and untargeted measures. All outcome units are test score student standard deviations.