Incentive Contracts in Complex Environments: Theory and Evidence on Effective Teacher Performance Incentives

Aaron Robert Phipps
Charlottesville, Virginia

M.A. Economics, University of Virginia, 2015
B.S. Economics and Statistics with Honors; Minor Mathematics and Philosophy, Brigham Young University, 2010

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of Doctor of Philosophy

Department of Economics

University of Virginia
May, 2018

_____

_____

_____

_____

Copyright

I certify that the dissertation I have presented for examination for the PhD degree of the University of Virginia is solely my own work unless otherwise clearly indicated.

Statement of Conjoint Work

One out of the three chapters that form this dissertation involve conjoint work:

Chapter 2 was co-authored with Emily Wiseman. Overall, my contribution amounts to 75% of the paper.

Abstract

The intent of incentive-based contracts – which tie compensation to performance in professions like teaching – is to improve productivity. In practice, the effects of such contracts have diverged markedly from predictions. The intent of this dissertation is to expand contract theory and provide empirical evidence from both the laboratory and real-world incentive programs on how contracts in complex environments, such as teaching, may be substantially improved. An innovation of this work is to present a theoretical model that considers the effects of output-based incentives when agents lack knowledge of the production function. In the context of incentive contracts for teachers, I expand on contract theory by adding uncertainty around the marginal productivity of inputs – such as different classroom activities – towards student test outcomes. I test the theoretical predictions of this model using variation in the implementation of evaluations in the Washington DC teacher incentive program and in the setting of a laboratory experiment.

In my first paper, "Personnel Contracts under Production Uncertainty: Theory and Evidence from Teacher Performance Incentives," I test the prediction that, due to production uncertainty, teacher incentives based on in-class evaluations may be substantially more effective than test-based incentives by separately identifying how two types of teacher incentives affect student outcomes. In the IMPACT program, teachers can be fired or receive large bonuses based on a combination of observational measures in unannounced in-class evaluations – which can be thought of as measures of teacher inputs – and test-based measures of the effect of teachers on student outcomes. I measure how teachers modify their behavior when they have no threat of an evaluation, and how those changes affect student test scores. Because the timing of in-class observations is random, the assignment of treatment – how many days a teacher has the threat of an evaluation – is exogenous. I find that increasing the number of days without the possibility of an evaluation leads to a decline in students' tested scores, which is inconsistent with a model in which agents know the production function, but consistent with my model of production uncertainty. I demonstrate that all of the positive effects of the IMPACT program can be explained entirely by the effect of a possible in-class evaluation, suggesting the test-based incentive has little or no effect. A takeaway from this analysis is that incentive-based compensation targeting production inputs may yield significant gains in the effectiveness of incentive contracts.

In my second paper, "Teacher Improvements in Windows of High-stakes Observation," I look deeper at the specific behavioral changes caused by the IMPACT incentive in Washington DC. I map how teachers modify specific components of their practice, as measured on their evaluations, in response to the daily probability of an in-class evaluation. In so doing, I illustrate the predictions of the Holmstrom-Milgrom multi-tasking model by showing teachers make the most improvements on teaching practices that are easily adjusted first. In the standard multi-tasking model, if employees devote all their attention on a single incentive, overall productivity may fall. However, overall teacher responses to the possibility of an evaluation still induce meaningful improvements in student outcomes, demonstrating the small cost of Holmstrom-Milgrom-style multi-tasking relative to the large gains from reducing employee production uncertainty by using an input-based incentive.

In my third paper, "Teacher Performance Pay through the Lens of Production Uncertainty: Theory and Evidence from a Real-Effort Laboratory Experiment," I test theoretical predictions of the production uncertainty model in a laboratory setting, which allows for controlled randomization in the production function in order to causally identify the effects of production uncertainty. I imitate uncertainty in the marginal value of inputs – analogous to inputs for student test scores – by asking participants to solve easy or hard problems to earn financial rewards, but the marginal payoffs are drawn from known distributions. Treatments vary by changing the variance of the marginal payoff to each task. I find that, as predicted, increased production uncertainty induces participants to favor inputs with lower variance in marginal productivity, even while holding all other things constant (including average marginal payoff).

For my wife, Jaimie, my children, Izzy and Dewey, and those children that may yet come.

Acknowledgments

Completing this dissertation has only been possible due to the generous support of many people. First and foremost, I am grateful and indebted to Sarah Turner, Leora Friedberg, and James Wyckoff, without whom this project would not have been possible. I would also like to thank William Johnson for his critically important insight and encouragement. There are a great many others who have provided helpful comments and ideas at various presentations or in personal meetings. In particular, I would like to thank Amalia Miller, Derek Neal, Brian Jacob, Susan Dynarski, and Susanna Loeb. I would also like to especially thank Steven Stern for the invaluable lessons he taught me and the extra time he devoted to my progress.

A great deal of thanks are due to the administrative staff at the UVA Economics department and the EdPolicyWorks Research Center, with special thanks to Patty Futrell, Sage Bradburn, and Leslie Booren, whose hard and sometimes tedious work made this process much smoother than it otherwise would have been.

A special note belongs here for my older siblings and closest friends who have quietly provided support for me and my family throughout these years. Their many kind acts would be difficult to enumerate, but they have not gone unnoticed. Perhaps most importantly, my happiness and sanity over the past five years were preserved by the loving and kind faces surrounding me at home. For this, I cannot overstate the debt owed to my wife, Jaimie, for her patience towards me and her tender care of our two lovely children, Izzy and Dewey. In answer to Izzy's prayers, I am hopeful that we can now get a house and a dog.

Funding Acknowledgments

The research reported here was supported by the Institute of Education Sciences, US Department of Education, through Grant #R305B140026 to the Rectors and Visitors of the University of Virginia. The opinions expressed are those of the authors and do not represent views of the institute or the US Department of Education. Additional support was provided by the Bankard fund. The laboratory experiment was funded by the generous donation of the Steer Family Endowment.

Contents

List of Tables

List of Figures

Chapter 1

Personnel Contracts under Production Uncertainty: Theory and Evidence from Teacher Performance Incentives

Performance incentive mechanisms for teachers and other professionals typically reward outcomes such as student test scores. Yet, incentive contracts may induce inefficient effort allocation if agents are uncertain about the production function. In such circumstances, incentives targeting production inputs may improve student outcomes more effectively than those targeting only the outcomes. The Washington DC teacher incentive program, IMPACT, provides a laboratory for testing how input-based and output-based incentives for teachers affect student test scores. Teachers face large incentives that are determined by their score on a combination of test-based measures of the effect of teachers on student outcomes and observational measures in unannounced in-class evaluations. Some teachers randomly experience days in which they are guaranteed not to receive an unannounced evaluation, which I use to identify how teacher behavior on these days affects student test outcomes. I find that increasing the number of days without the possibility of an evaluation leads to a decline in students' tested scores. I argue that my empirical results confirm the hypothesis that input incentives are more effective because teachers are uncertain about the test score production function.

1.1 Introduction

There is an apparent paradox in compensating teachers based on their performance. If teachers are best informed about how to teach their students, bonuses for improving student test outcomes should be more effective than those for observed teacher behavior. Yet the available evidence demonstrates that, among equally sized bonuses, only incentives which include rewards based on in-class performance evaluations improve student test outcomes. I address this paradox by expanding a principal-agent model to incorporate teacher uncertainty about the marginal productivity of inputs ("production uncertainty"). Outcome-based incentives can induce risk-averse teachers to inefficiently allocate their effort away from effective teaching practices because of uncertainty about their marginal productivity. How much students learn from conducting a science experiment may vary considerably, making it a riskier teaching choice than directly lecturing on material contained in the test, if compensation relies on student test scores. Rewarding teachers for particular classroom practices, however, does not induce the same uncertainty and may lead to greater improvements in student outcomes. I demonstrate the theoretical result that in-class observations should be a key mechanism of successful teacher incentive programs. The structure of the DC Public School (DCPS) teacher incentive program, called IMPACT, provides a natural experiment in which teachers randomly experience more days with the possibility of an unannounced evaluation. Using administrative data from IMPACT, I show how teacher responses to the possibility of an in-class evaluation substantially improve student test scores. I can also distinguish this from the positive feedback effect following in-class evaluations.

Policy makers worldwide are turning to teacher performance incentives to improve teacher quality. The motivation to improve teacher quality is a direct result of mounting evidence that good teachers matter. Good teachers measurably improve student tested outcomes and a variety of life outcomes ranging from decreasing the likelihood of teen pregnancy to increasing college attendance and lifelong earnings (Rockoff, 2004; Rivkin et al., 2005; Kane and Staiger, 2008; Aaronson et al., 2007; Chetty et al., 2014). Yet traditional programs to improve teacher quality, such as teacher training and increased teacher education, are largely ineffective (Weisberg et al., 2009; Hanushek, 2007). Instead, new approaches to

improving teacher quality often rely on economic theory to design incentives intended to motivate and retain effective teachers. The US federal government has spent $6.5 billion since 2010 on programs that reward teachers based on their performance, though the measured effects of such programs are lackluster. Of the effective programs, previous empirical studies are unable to identify the mechanism of their success (Wellington et al., 2016).

In the US, performance pay programs typically reward teachers for their estimated contribution to student test score growth. Conventional principal-agent models predict that outcome-based incentives should be more effective than rewarding specific inputs because teachers have additional information about their specific talents and the needs of their students. For the same reason, the educational community at large supports teacher autonomy in the classroom. Notwithstanding, some performance incentive programs use teacher scores on unannounced in-class observations to determine bonuses. Rigorous in-class evaluations measure how well a teacher implements teaching best practices on a variety of dimensions such as her classroom management techniques or her use of student assessment. In theory, if a teacher knows how these practices would improve student test outcomes, an output-based incentive should improve student test scores at least as much as input-based incentives. But the available empirical evidence shows that without in-class observations, teacher incentive programs are ineffective. This observation reveals a paradox: Why would incentives for specific inputs be more effective than incentives for outputs?

I expand a standard principal agent model to show that output-based incentive contracts may induce agents to allocate their effort inefficiently if they are uncertain about the marginal productivity of inputs. In response to an incentive based on student math scores, a teacher may resort to rote memorization instead of using a more pedagogically sound approach, such as using a manipulative like counting cubes. Even if using a manipulative is more effective on average, if its effect appears more uncertain to the teacher, she would favor rote memorization. In this case, rote memorization has a greater return in certainty equivalence terms. Even though she may increase her overall effort, her students would improve considerably more if she directed her increased effort towards the most productive inputs. Incentives based on in-class observations would provide the motivation to increase her effort but without inducing her to allocate it inefficiently.

The Washington DC IMPACT program provides a natural experiment to test how input-based and output-based incentives for teachers affect student out-

comes. Teachers face one-off bonuses of up to $25,000 and possible permanent pay increases, while exceptionally low-performing teachers are dismissed. Rewards are determined by a teacher's IMPACT score, which is the weighted average of their value-added score and the average of five in-class evaluations. The structured time frame of evaluations means that some teachers randomly experience spells in which they are guaranteed no evaluation will occur. I use this variation to causally identify how changes in teacher behavior on these days affect student outcomes. However, teachers who experience more days without the possibility of an evaluation likely had their evaluations earlier than their peers. As a result, these teachers will also have more time than their peers to enact evaluation feedback, potentially improving their students' test outcomes and masking any observed effect from the incentive (Taylor and Tyler, 2012). I can separately identify the effect of evaluation feedback by using the random variation in the length of time between the day of an evaluation and the day a teacher discusses her results with her evaluator.

I find that both the possibility of an evaluation and the feedback a teacher receives substantially improve student test outcomes. For each additional day in which it was possible for a teacher to be evaluated, her students scored up to 0.005 standard deviations higher in reading and math, depending on the season. Overall, in a back-of-the-envelope calculation, I estimate the possibility of an evaluation improved student reading scores up to 0.074 standard deviations and math scores up to 0.089 standard deviations, though these are upper bound estimates. These are large effects relative to other teacher incentive programs, but they are consistent with the estimated effect of in-class evaluations in high-demand charter schools (Dobbie and Fryer, 2013).

While my empirical approach cannot specifically identify why teachers decrease their effectiveness during days without the possibility of an evaluation, I argue my results support the narrative that teachers are uncertain about the test score production function. The IMPACT program is unique because a teacher's in-class evaluation score and value-added score jointly determine her eligibility for a reward, as opposed to having a separate bonus for each. Using a stepwise incentive model similar to Lazear and Rosen (1981), I show how this feature makes it likely that, if the teacher knows the production function, she would increase her student test scores during no-risk days. However, with production uncertainty, the uncertainty costs become more salient after her evaluation, de-

creasing her overall effort and reducing the marginal benefit of inputs that would improve student test scores.

1.2 Motivation and Research Context

Improving teacher quality is an effective tool for reaching a variety of policy objectives. A one standard deviation increase in teacher quality, as measured by her contribution to student test score growth (value-added), increases lifetime earnings by roughly $39,000 after a single year of school (Chetty et al., 2014). Hanushek (1992) estimates that students with teachers in the 95th percentile gain a full year's worth more than students with teachers at the 5th percentile, measured by standard test scores. Over time, a student who consistently has low quality teachers will be significantly behind her peers. The wide variation in teacher quality motivates the effort to improve the pool of teachers. School districts historically attempted to improve teacher quality by promoting more teacher training and experience by using what is called a "steps and lanes" system (or "single salary"), which makes salaries depend only on education, certification, and teaching experience. The available evidence shows that these factors do not translate into improved student outcomes (Rivkin et al., 2005). Such pay systems are discouraging to young but effective teachers looking to distinguish themselves in their career, and they provide no credible method for acknowledging and celebrating effective teachers.[1]

An alternative payment scheme would seek to identify effective teachers based on their performance and reward them accordingly. This payment scheme, in theory, would encourage high-quality teachers to self-select into teaching and provide appropriate rewards for working being productive. A handful of school districts in the US have attempted to create performance incentives throughout the 20th century, but few of these programs survived for more than a couple years, at least until recently (Murnane and Cohen, 1986). Programs implemented in the last decade have had mixed effects on student outcomes, and there is little empirical evidence to determine what makes some incentives effective. This paper presents both theoretical and empirical arguments that in-class evaluations are a key mechanism of effective teacher performance incentives.

[1] Hoxby and Leigh (2004) estimate that the share of teachers in the highest aptitude category fell from 5% to 1% from 1963 to 2000, and they estimate that 80% of this decline can be attributed to high quality teachers being pushed out of the profession because of the lack of pay differentiation.

1.2.1 The Theory of Incentive Contracts as Applied to Teachers

The problem of creating optimal incentives for teachers falls within contract theory. The basic intent of incentive contracts is to reduce employee moral hazard given their asymmetric information about their own effort and talents. An incentive contract is intended to align the incentives of an employee such that they choose the same level of effort and combination of inputs that the principal would, which is the first-best solution. A variety of obstacles complicate optimal contract design. One basic problem is that creating incentives for specific workplace behaviors (inputs) can create inefficiencies because of the differences in comparative advantage. For example, an employer seeking to improve sales could reward a salesperson for each call he makes instead of per sale. But such an incentive may be inefficient because it is possible that the employee has an unusually high marginal productivity per minute spent writing emails. The incentive to increase calls will induce him to inefficiently spend his time on calls instead of emails. To avoid this problem, the employer can instead reward the salesperson per sale.[2] This basic conclusion motivates the paradox of teacher incentives: why would incentives based on specific teacher behaviors lead to greater gains in student test scores than rewarding teachers for improving test scores?

Theoretically it is not clear that incentives for teachers should be a feasible means for improving student outcomes. Teaching has well-known characteristics that complicate incentive design (Murnane and Cohen, 1986; Dixit, 2002). Teachers are motivated agents, which makes their response to incentives more inelastic (see Dixit, 2002; Francois, 2000, for example). Motivated teachers may prioritize certain outcomes that do not necessarily align with those of the general public (Neal, 2011). Teachers are also responsible for improving a variety of outcomes that are hard to define, making it unclear which outcome should be used for determining bonus payments, but rewarding teachers for multiple outcomes is likely inefficient (Holmstrom and Milgrom, 1991). The most important teaching outcomes could be measured, in theory, with well-designed student standardized tests, but should teachers be paid for their individual contributions to student outcomes or for their team's contribution? Teachers often collaborate, which implies individual, rank-based incentives could reduce collaboration, but a bonus

[2]This is a basic result discussed in the comprehensive contract theory reviews of Lazear and Oyer (2012) and Prendergast (1999), which also highlight several concerns that are key for my analysis: how well the measured output aligns with the desired outcome (Akerlof and Kranton, 2005; Neal, 2011; Benabou, 2016), potential gaming of the outcome measure (Baker, 1992), and the use of subjective measures of employee output (Levin, 2003; MacLeod, 2003; Gibbs et al., 2004).

for group achievement introduces moral hazard (Holmstrom, 1982; Kandel and Lazear, 1992).[3] A teacher's contribution to student test scores ("value-added") is the average difference between her students' predicted scores and realized scores. Teacher value-added scores have inherent noise, which arises from random events that affect students while taking their tests, such as illness or distractions. Teachers may become less responsive to an incentive as measurement noise in tests increases (Lazear and Rosen, 1981). Yet in spite of these concerns, some teacher performance incentives have measurably improved student outcomes. The question remaining is what was the driving difference between effective and ineffective programs?

A common explanation for the ineffectiveness of some teacher incentive programs is that the incentive size was too small. In theory, increasing the incentive size should eventually overcome the incentive design problems resulting from noisy value-added scores and motivated agents. Yet incentive size does not appear to explain the variation in the observed effects of teacher incentive programs. For example, the incentive size of the Denver ProComp teacher performance incentive is comparable to the possible bonuses in the Washington DC incentive program, yet only the DC program measurably improved student outcomes (Dee and Wyckoff, 2015; Briggs et al., 2014). Not only were the incentive sizes similar, but both programs used value-added to assign bonuses. A key distinction in the DC program is its use of in-class observations in addition to teacher value-added.

1.2.2 Teacher Performance Incentive Programs in the US and Their Effects

Teacher performance incentives (also called "merit pay" or "performance pay") are an old idea that has reappeared several times since the early 20th century (Murnane and Cohen, 1986). Recently, US and international policy makers have actively encouraged public schools to use performance incentives with large-scale federal programs like Race to the Top and the Teachers Incentive Fund. School districts responded by implementing teacher performance incentives that vary considerably in their implementation details. What follows is a discussion of the

[3] As evidence of the moral hazard of group incentives, Imberman and Lovenheim (2015) evaluate the effect of grade-based teacher incentives and find that increasing a teacher's portion of the students in a grade leads to higher test outcomes (to a point). In a randomized trial in New York, Fryer (2013) found that a school-level incentive had no effect on any measured outcome.

most recent developments in the empirical literature of incentives for individual teacher performance.

Empirical evidence of the effects of teacher-level performance incentives has been mixed, but there have been distinctly effective programs. Dee and Keys (2004) exploit the random student assignment from the Tennessee STAR experiment, which overlapped with the implementation of the Career Ladder. This program awarded career advancement and bonuses to teachers for achieving milestones in their in-class evaluations. The authors find that students of teachers enrolled in the incentive program improved math and reading scores. Dee and Wyckoff (2015) use a regression discontinuity approach around the sharp cutoffs in the IMPACT program. They find that dismissal threat improved a teacher's effect on student test outcomes by up to 0.05 standard deviations when compared to teachers just above the threshold in the previous year. Other work has found some evidence that the Teacher Advancement Program (TAP), which is a comprehensive teacher review and reward system, improved student outcomes, but these results are less robust (Mann et al., 2013).

Other programs had only minor positive effects, but in these cases the policy analyzed did not dictate specifically how a school district was to create teacher incentives. For example, the Minnesota Quality Compensation (Q-Comp) program, started in 2005 and still ongoing, only requires that school districts implement an incentive program but the legislation enacted does not specify the structure. Q-Comp had some small positive effects, but it is unclear what incentive design elements lead to these positive effects (Sojourner et al., 2014). As for the results of the many programs funded by the Teacher Incentive Fund (TIF), the Department of Education's Institute of Education Sciences produced a report that shows small positive effects after three years, but like Q-Comp, there is no clear mechanism (Wellington et al., 2016).

There have been other sizable programs that showed little or no effects on student test scores. The Tennessee Project on Incentives in Teaching (POINT) was a three-year experiment started in 2006. While selection into the experiment was voluntary, assignment to treatment was randomized. Treated teachers would receive bonuses based solely on the test score improvements of their students. There were no significant positive effects from the incentive (Springer, 2010). The Denver Professional Compensation program (ProComp), started in 2007, created several routes for teachers to receive bonuses, but by far the largest bonus was awarded to teachers with large gains in student test scores. A report from the

University of Colorado, Boulder finds that this incentive had no positive effects on student test scores (Briggs et al., 2014).

In all, these empirical studies provide mixed evidence of the effectiveness of teacher performance pay. In other contexts, there is strong evidence that performance incentives improve employee effort. For example, Lazear (2000) finds that a piece-rate wage in the Safelite Glass Corporation led to significant improvements in output. In a firm-level randomized experiment, Bandiera et al. (2007) show that managers receiving a performance incentive increase the productivity of their team.

In education, the successful application of performance incentives in some school districts suggest that teaching can be an appropriate context for performance incentives, but what element defines a program's success? Of the incentive programs described above, the incentive sizes are mostly comparable. The ineffective Denver ProComp and POINT programs both had bonuses ranging between $5,000 and $15,000, and sometimes even more. Yet the effective Career Ladder program, which did not use any test-based measures of teacher effectiveness, had pay increases ranging from roughly $2,000 to $4,000 in 2017 dollars. These observations alone suggest a large disparity in how effective test-based incentives are relative to input-based incentives.

The use of in-class evaluations is a distinguishing characteristic of effective teacher performance pay. To illustrate, I divide up the teacher incentive programs detailed above in Table 3.1 based on the use of incentives for in-class evaluation scores and any measurable incentive effects.[4] All of these incentive programs have teacher-level measurements and incentives with minimal school-level incentives. Other performance programs not included in Table 3.1 rely almost entirely on group-level achievement.[5] Six of the seven programs use measures of how a teacher improves student test scores. While not a causal argument, Table 3.1 highlights a positive relationship between a performance incentive's effect and the use of in-class evaluations.

[4]The Denver ProComp program uses pass-fail in-class evaluations. Over 99 percent of teachers pass these exams, making them perfunctory exercises at best (Briggs et al., 2014), which is why I do not consider the ProComp program to have any rigorous or meaningful in-class evaluations. Such pass-fail in-class evaluations are common, and are known to have no meaningful effects on teacher behavior (Weisberg et al., 2009).

[5]Notable programs are the School-wide Bonus Program in New York (Fryer, 2013), the Dallas School Accountability and Incentive Program (DSAIP) (Ladd, 1999), the Kentucky Instructional Results Information System (KIRIS) (Koretz and Barron, 1998), the North Carolina ABC program (Vigdor, 2008), the Chicago version of TAP (Glazerman and Seifullah, 2012), and Houston's ASPIRE program (Imberman and Lovenheim, 2015)

If teachers can predictably improve student test scores by engaging in behaviors measured by in-class evaluations, why not employ these techniques when there is only a test-based incentive? One possible explanation is that the observed effects are not driven by incentives at all: perhaps programs that utilize in-class observations are effective because they provide valuable feedback that teachers use to improve. In Ohio, evaluation feedback improved teacher quality in a district which staggered the roll-out of a teacher evaluation program that had no financial incentives (Taylor and Tyler, 2012). This result suggests that the effects of the Career Ladder incentive program in Tennessee may have been driven in large part by the positive effects of evaluation feedback. The analysis in Dee and Keys (2004) is unable to identify the mechanism of teacher improvements. In my empirical approach, I estimate both the incentive effects of in-class evaluations and the positive effects from evaluation feedback.

This paper addresses the somewhat paradoxical results of teacher incentives both theoretically and empirically. First, I present a theoretical argument why input incentives may be more effective than output incentives in complex contexts such as teaching. Second, I present empirical evidence identifying the substantial positive effects of unannounced in-class observations. A distinctive feature of my empirical approach is that it separates the effects of teacher improvements as a result of the incentive from teacher improvements as a result of evaluation feedback.

1.3 Agent Production Uncertainty

In what follows, I present a principal-agent model that expands on the multitasking model in Holmstrom and Milgrom (1991), with two modifications. First, to clarify the theoretical results, I remove measurement error. Second, I add uncertainty about the marginal productivity of inputs for agents. The result is a rich model with distinctive predictions. To preview, I show that even if all outputs are measured without noise, designing incentives around production inputs can have substantially greater effects than designing incentives around the final output when agents are uncertain about the production function. The intuition is that, with production uncertainty, some inputs have low rewards in certainty equivalent terms, even if they have high average productivity. As a result, agents may favor inefficient inputs, or become unresponsive to an incentive based on outputs.

The theoretical results are consistent with a growing empirical literature in teacher incentives. Incentives in other industries have been met with mixed results as well, such as hospital and doctor incentives (Kristensen et al., 2014; Van Herck et al., 2010; Flodgren et al., 2011). These incentives rely on measures of patient outcomes and health measures, which effectively have no measurement error. Without measurement error, these empirical findings are inconsistent with the predictions of a standard principal-agent model, but consistent with my model of production uncertainty.

1.3.1 Basic Piece-Rate Incentive with Production Uncertainty

Consider the teacher's daily problem of allocating her time and effort both in preparation and execution of the next day's lessons.[6] Assume there are n possible inputs. These can include different lesson types or instructional practices. The teacher exerts x_i effort and time for input i. Her effort vector is $x = [x_1, x_2, \ldots, x_n]^T$. Her effort comes at a cost, $C(x)$, which I assume exhibits increasing marginal costs. She receives a bonus that is proportional to the amount she has improved her students' test scores (her value-added), denoted y. Then her wage is defined as $W = A + B \cdot y$, where A is her guaranteed wage and B is the piece-rate bonus for improving her value-added.

A teacher's daily optimization problem requires her to allocate effort and improve her value-added score, $y(x)$.[7] Value-added is a measure of how much a teacher has improved her students' test outcomes from the previous year, beyond what was projected for each student. Let $f(x)$ be the production function of value-added scores, then assuming no measurement error, $y(x) = f(x)$.[8] To model a teacher's uncertainty about the test score production function, I use a Taylor Expansion around some reference input vector, $\tilde{x}$. From the teacher's perspective, the production function is as follows:

$$y(x) = f(\tilde{x}) + (x - \tilde{x})^T \nabla f(\tilde{x}) + \xi(x - \tilde{x}) \tag{1.1}$$

[6] This version of agent production uncertainty is meant to illustrate how an output-based incentive may cause inefficient effort allocation. Many of the assumptions made are used to ease the presentation, but the results are robust to a variety of realistic modifications. I will extend the model to a step-wise incentive scheme in the next section.

[7] Because the teacher only observes her value-added score at the end of the year, her daily effort allocation will be identical for every day throughout the year.

[8] This notation is similar to Holmstrom and Milgrom (1991). To make a multi-dimensional incentive, allow y to be a vector of measured outcomes and ε_y a vector of measurement noise for each, then the setup is identical to Holmstrom and Milgrom (1991): $y(x) = f(x) + \varepsilon_y$.

where $\xi(x - \tilde{x})$ is a measure of her inaccuracy about the production function as she considers input levels farther from her reference point, $\tilde{x}$, and is the usual error term of a Taylor expansion.

Then introduce the teacher's production uncertainty by allowing $\nabla f(\tilde{x})$ to be a random variable with a normal distribution with mean $\mu = \nabla f(\tilde{x})$. That is, let $\nabla f(\tilde{x}) = \mu + \gamma$ where γ is a mean-zero normally distributed random variable with covariance matrix Σ. Assuming an exponential utility function with coefficient of risk aversion, r, her daily expected utility is

$$EU = E\Big[- \exp\Big\{-r\big(A + B \cdot (f(\tilde{x}) + (x - \tilde{x})^T(\mu + \gamma) + \xi(x - \tilde{x})) - C(x)\big)\Big\}\Big]. \tag{1.2}$$

For now, let $\xi(x - \tilde{x})$ be negligible.[9] For simplicity, I assume $\tilde{x} = 0$, which can be thought of as designating $\tilde{x}$ to be the amount of effort with no performance pay. Then x is deviations from this "default" contract. The moment generating function of a normal distribution simplifies Equation 1.2 considerably:[10]

$$EU = - \exp\left\{-r\big(A + B \cdot (f(\tilde{x}) + x^T\mu) - C(x) - \frac{1}{2}rB^2x^T\Sigma x\big)\right\}. \tag{1.3}$$

The teacher's maximization problem is then:

$$\max_x \; A + B \cdot \underbrace{(f(\tilde{x}) + x^T\mu)}_{\substack{\text{Expected} \\ \text{Production}}} - C(x) - \underbrace{\frac{1}{2}rB^2x^T\Sigma x}_{\substack{\text{Uncertainty} \\ \text{Cost}}}. \tag{1.4}$$

The teacher has uncertainty costs as highlighted, which increase as her uncertainty in the marginal productivity of inputs increases, represented by Σ. Her uncertainty costs are also increasing in her coefficient of risk aversion r, the size of her bonus B, and her allocation of effort x. A key outcome is the uncertainty

[9] In other applications, the size of $\xi(x - \tilde{x})$ can be used to motivate aspects of prospect theory, such as reference points and framing.

[10] The moment generating function for a random variable X is defined as $M(t) = E[e^{tX}]$. The multivariate normal moment generating function is $M(t) = e^{t^T\left(\mu + \frac{1}{2}\Sigma t\right)}$. Then let $t = -rB(x - \tilde{x})$, then the utility function can be rewritten as $U = -e^{-r(A + B(y(\tilde{x}) - C(x))} \cdot e^{t^T\gamma}$, where the only random variable is γ in the last term. Substituting the moment generating function means

$$\begin{aligned} E[e^{t^T\gamma}] &= e^{t^T\mu + \frac{1}{2}t^T\Sigma t} \\ &= e^{-r(B(x-\tilde{x})\mu - \frac{1}{2}rB^2(x-\tilde{x})^T\Sigma(x-\tilde{x}))}. \end{aligned}$$

cost term highlighted. In the Holmstrom-Milgrom model, the uncertainty cost is linear in x, whereas now it has increasing marginal costs in x.

The teacher's first-order conditions show that increases in production uncertainty, Σ, or the bonus size B, will affect her marginal effort choice:

$$B\mu - rB^2\Sigma x = \nabla C(x). \tag{1.5}$$

If there is negligible agent production uncertainty ($\Sigma \approx 0$), then this is a first-best solution for a risk-neutral principal: the teacher optimally picks her inputs such that the marginal benefits, $B\mu$, equal the marginal costs, $\nabla C(x)$.[11] However, as elements of Σ become larger, production uncertainty will distort her effort choices away from the first-best solution.

Equation 1.5 provides a number of practical implications. The structure of Σ changes a teacher's overall effort and how she allocates it. Suppose there are only two inputs a and b, and suppose both have the same marginal cost. Let input a be a math lesson using a manipulative like counting cubes, whereas input b is using a worksheet that emphasizes information on the test. If using counting cubes improves tests more on average than using a worksheet, this would be the input preferred by the principal. Without production uncertainty, a piece-rate incentive on test scores would provide encouragement to a teacher for using counting cubes. But if the teacher is uncertain about the marginal product of counting cubes, she may choose the worksheet instead. There are a number of uncontrollable factors that may make the effect of one approach more uncertain than another. There may be variation in how well students receive one approach, or an approach may depend on how well students interact with each other.

Even if a teacher increases the intensity of her effort in response to an incentive, she may allocate it inefficiently by preferring a worksheet. This distortion in her allocation of effort is an effect I call "friction." In essence, she is compensating for her uncertainty about the counting cubes by increasing her use of the worksheet. If she is uncertain about the marginal effect of both inputs, she cannot compensate in this manner. In this case, the fact that all inputs have a highly variable effect on student outcomes makes the teacher believe she has no control over her students'

[11] Even if $y(x)$ is measured with noise such that $y(x) = f(x) + \varepsilon$, with a linear compensation scheme, her optimal effort allocation is not affected if there are constant marginal costs. With a step-wise reward, measurement noise will reduce $\sum x_i$, but it will not change how she allocates total effort among her inputs. If instead of interpreting x as teacher effort, it is how she allocates time within a fixed-length day, measurement noise will have no effect on how she allocates her time throughout the day.

outcomes. As a result, the incentive will not increase her effort on either teaching style, an effect I call "futility." Because an incentive based on an output with an uncertain production function can induce friction and futility, it is possible that an incentive based on an input will be more effective, in spite of the potential inefficiencies caused by asymmetric information and Holmstrom-Milgrom style multitasking.

In a principal-agent problem, the first-best outcome is defined as the effort allocation the principal would choose if it were possible to dictate the agent's actions knowing the agent's preferences over income and labor costs. When output is measured perfectly, a linear payment scheme based on output achieves the first-best (Lazear, 1986). Under production uncertainty, this is rarely true. Furthermore, this result does not depend on differences in risk preferences between the principal and the agent.

To illustrate, assume each class is a new draw of the marginal productivity of inputs, γ, for a teacher. The assumption implies there are important interactions between students and their peers that will influence the productivity of different teaching styles, making γ depend on the classroom composition, not just individual student characteristics. For example, some classrooms may function exceptionally well under peer-led group learning, while others certainly will not. Principals oversee multiple classrooms, giving them multiple draws of γ each year, effectively reducing its covariance. As a result, even if principals and teachers completely agree on the average marginal effect of each input, μ, the test-based incentive will still induce teachers to allocate their effort differently than the principal would dictate if possible. The marginal benefit of inputs in certainty equivalent terms is different between principals and teachers, but not necessarily because of differences in risk preferences.

Alternatively, production uncertainty can be thought of as a measure of teacher expertise, where more experienced teachers have smaller variance about the marginal productivity of each input. Supposing the principal is a highly experienced teacher, she would have less uncertainty about the marginal productivity of inputs. Yet even if the teachers in her school agreed on what constitutes "best practices" in teaching, their uncertainty about the marginal productivity of inputs could still induce them to allocate their effort inefficiently from the principal's perspective. When paid based on student test outcomes, teachers' lack of con-

fidence in effective teaching practices can induce them to resort to less effective teaching approaches.[12]

1.3.2 Extending to Step-Wise Incentive

In application, many teacher incentive programs use a step-wise reward system instead of the piece-rate, linear reward described above. In order for the theoretical results to apply to the IMPACT incentive structure more readily, I modify the model for a step-wise incentive and add measurement noise.

As before, a teacher chooses her daily effort vector, x, which produces $f(x)$ in student test scores. Then her observed contribution to student test scores is $S(x) = f(x) + \varepsilon$ with mean-zero measurement error ε. The teacher knows the distribution of ε. She will receive her bonus B if her value-added score, S , passes some threshold S^*.

Unlike before, I assume she is risk neutral.[13] Then she choose her inputs to solve the following maximization problem:

$$\max_x P(x)B - C(x). \tag{1.6}$$

Her choice of inputs determine her probability of earning a bonus, which is $P(x)$. The probability depends on the measurement noise of test scores and evaluation scores, as well as the presence of production uncertainty. The teacher's first-order

[12]There is growing evidence that in some cases, teachers' attempts to "teach to the test" are less effective at improving test scores than other pedagogically sound approaches (Hill et al., 2015; Blazar, 2015; Blazar and Kraft, 2016; Blazar and Pollard, 2017). This is not to suggest that teaching to the test is always ineffective, but that teachers' ineffective attempts demonstrate a lack of confidence in established teaching best practices.

[13]As in Lazear and Rosen (1981), the step-wise payment function generates diminishing marginal returns to effort because of measurement noise, which drives the theoretical results. Allowing for risk aversion does not qualitatively change the results, though it does complicate the model.

condition for each input x_i is as follows:[14]

$$\frac{dP}{dx_i}B = \frac{dC}{dx_i} \tag{1.7}$$

Her probability of earning a bonus will depend on the distribution of measurement noise, ε, and whether or not there is production uncertainty.

Following a linear estimate of the test score production function, the teacher's value-added score is represented by:

$$S(x) = f(\tilde{x}) + (x - \tilde{x})^T(\mu_f + \gamma_f) + \varepsilon \tag{1.8}$$

which includes both production function uncertainty and measurement error, ε, with assumed zero mean and variance σ^2. As before, γ_f is a multivariate normal random variable with covariance matrix Σ.

To simplify, again assume $\tilde{x} = 0$. Let her expected score be $\bar{S}$, which can be expressed as $\bar{S} = x^T\mu_f$. Her score, S, is a random variable that is normally distributed and centered around $\bar{S}$. That is, define the error term $\theta = S - \bar{S} = \varepsilon + x^T\gamma_f$, then her final score is a random variable that can be rewritten as $S = \bar{S} + \theta$ and θ is normally distributed with mean zero and variance $\rho(x)$. The functional form of $\rho(x)$ is:

$$\rho(x) = \sigma^2 + x^T\Sigma x. \tag{1.9}$$

If there is no correlation in the error term for the marginal productivity of inputs, then the off-diagonal elements of Σ are zero.

Given that a teacher's score is a normally distributed random variable with mean $\bar{S}$ and variance $\rho(x)$, the probability of earning her bonus is

$$\begin{aligned} P(x) &= Pr[\varepsilon + x^T\gamma_f > S^* - \bar{S}] \\ &= 1 - \Phi\left(\frac{S^* - \bar{S}}{\sqrt{\rho(x)}}\right) \end{aligned} \tag{1.10}$$

[14] Allowing the teacher to be risk-averse with an exponential utility function changes Equation 1.7 only slightly. The term B is replaced with a weighting function that is increasing in B, call it $\Gamma(B)$:

$$\frac{dP}{dx_i}\frac{1}{r}\Gamma(B) = \frac{dC}{dx_i}$$

For values of r and B such that $rB > -\ln\left(\frac{1}{2}\right)$, the weighting function is greater than 1, $\Gamma(B) > 1$, as expected. This means that increasing the bonus B will increase the weight placed on improving the probability of earning the bonus.

where $\Phi(\cdot)$ is the standard normal distribution function.

The teacher's first-order conditions for each input i, Equation 1.7, are determined by the marginal effect of x_i on the probability of earning a bonus. Equation 1.10 previews how production uncertainty will change the first-order conditions. If there is no production uncertainty, the variance of her score no longer depends on her choice of inputs: $\rho = \sigma^2$. The intuition is that, without production uncertainty, a teacher is only shifting $\bar{S}$ such that her marginal benefits equal her marginal costs. However, with production uncertainty, her choice of x also influences the distribution's shape. Increasing x increases the variance, $\rho(x)$, which flattens the distribution of S around $\bar{S}$. To see this, the marginal effect of x_i on the probability of earning the bonus is

$$\frac{dP}{dx_i} = \phi\left(\frac{S^* - \bar{S}}{\sqrt{\rho(x)}}\right)\frac{1}{\sqrt{\rho(x)}}\frac{d\bar{S}}{dx_i} - \psi_i(x, \Sigma) \tag{1.11}$$

where $\phi(\cdot)$ is the probability distribution function of the standard normal distribution. The term $\psi_i(x, \Sigma)$ is the cost of flattening the distribution and is a direct result of production uncertainty. Before discussing $\psi_i(x, \Sigma)$, note that the first term in Equation 1.11, $\phi\left(\frac{S^*-\bar{S}}{\sqrt{\rho(x)}}\right)\frac{1}{\sqrt{\rho(X)}}$, is decreasing in $\rho(x)$ when $\bar{S}$ is suitably close to S^*.[15] That is, the marginal benefit of increasing x_i decreases as measurement noise increases. As a result, increasing measurement noise will reduce a teacher's response to an incentive as in Lazear and Rosen (1981).

The function $\psi_i(x, \Sigma)$ in Equation 1.11 is a "flattening cost" from increasing input i.[16] The flattening cost reflects the reduction in the marginal benefit of increasing an input because doing so will increase the variance of S and stretch its distribution. To understand the key properties of ψ_i, let the diagonal elements of the covariance matrix, Σ, be labeled δ_i. The first key property is that the flattening cost for input i is always zero if there is no production uncertainty for input i. That is, $\psi_i(x, \Sigma | \delta_i = 0) = 0$. Increasing an input that has no production uncertainty

[15] Intuitively, if $\bar{S} = S^*$, increasing the variance of S pushes the pdf down at the mean. Of course, in the tails of the distribution, increasing the variance will increase ϕ. The exact condition for $\phi(\cdot)\frac{1}{\sqrt{\rho(x)}}$ to be decreasing in $\rho(x)$ is that $S^* - \bar{S} < \sqrt{\frac{1}{2}\rho}$

[16] While some key properties are discussed here, the details about the properties of $\psi_i(x, \Sigma)$ are beyond the scope of this paper. The function is

$$\psi_i(x, \Sigma_\gamma) = \frac{1}{2}\phi\left(\frac{S^* - \bar{S}}{\sqrt{\rho}}\right)\frac{\bar{S}}{\rho^{\frac{3}{2}}}\frac{d\rho}{dx_i}$$

will only shift the distribution but it will not add any variance to S. The second key property is that for two inputs i and j, if the production uncertainty of i is greater than that of j, then the flattening cost of i will be greater than that of j. In short, if $\delta_i > \delta_j$ then $\psi_i > \psi_j$, under reasonable conditions.[17] The flattening cost is continuous in x_i and zero for $x_i = 0$. This guarantees that there is always a value x_i such that $\frac{dP}{dx_i} > 0$. Under reasonable parameters, the flattening cost is smaller than the first, positive term in Equation 1.11.[18]

The key result of the model with step-wise incentives and production uncertainty is that production uncertainty will distort how a teacher responds to an incentive. In general, Mirrlees (1971) shows that if x has a sufficiently large effect on the outcome of interest around S^*, then a step-wise incentive will approximate the first-best solution. My theoretical result is that if the principal is seeking to maximize average productivity, an output with high production uncertainty will not achieve the first-best allocation of effort.

1.4 Empirical Evidence of Production Uncertainty in Personnel Contracts

This paper's empirical contributions are twofold: I first establish that input teacher incentives improve student tested outcomes. I then argue that the observed effects are evidence of production uncertainty given the structure of the IMPACT program. A secondary but distinguishing feature of my empirical approach is its capacity to separately identify the incentive effect of high-stakes evaluations from positive effect of evaluation feedback. This is possible because evaluations are randomly timed and because there is random variation in when teachers received feedback following their evaluation.

To preview, the activities teachers employ when concerned about a potential unannounced in-class evaluation substantially improve student test scores. I also find that evaluation feedback can significantly improve student outcomes. Another feature of my approach is that I can identify discontinuous changes in the effect of teacher effort when the probability of an evaluation is zero. Discontinuous changes in the effect of teacher effort are difficult to reconcile with a model in

[17] This latter result depends only on $\frac{d\rho(x)}{dx_i} > \frac{d\rho(x)}{dx_j}$, which is true assuming the off-diagonal elements of Σ are zero and as long as $\frac{\delta_i}{\delta_j} > \frac{x_j}{x_i}$.

[18] This result is driven by the fact that $\psi_i(x, \Sigma)$ decreases in ρ by a factor of $\frac{1}{\rho}$ more than the positive term.

which the teacher knows the production function. If teachers know the production function, then teachers with a very small probability of an evaluation should have roughly the same daily effect on student test outcomes as a teacher with no possibility of an evaluation. However, if teachers have uncertainty about the test score production function, the uncertainty cost of test-directed effort becomes more salient after an evaluation, causing a discontinuity in their effort allocation. My results then function as test to demonstrate my hypothesis of production uncertainty in teaching.

1.4.1 Setting and Data Source

The IMPACT program began in the 2009-10 school year and its structure was unchanged for the first three years. Over this time period, DCPS has between 128 and 133 elementary, middle and high schools, with roughly 3,500 teachers each year. Of these teachers, about 13 percent (475 each year) teach grades and subjects for which a teacher's value-added score can be calculated.[19] Slightly less than half (42 percent or 200 teachers) of these teach both math and reading. The remaining 275 teachers are evenly split between teaching only math and only reading.

As part of the IMPACT incentive program, teachers receive evaluations from both principals and district employees called "Master Educators." Principals conduct three evaluations throughout the year, and master educators conduct only two. Principals are required to inform teachers a day in advance of their first evaluation, but the remaining evaluations are unannounced. Similarly, master educators must announce their first evaluation but not their second. The in-class observation uses a well-defined observation rubric called the "Teaching and Learning Framework" (TLF). TLF is a 9-dimensional grading rubric derived from the Danielson Framework. For each dimension, teachers receive a score between 1 and 4. The final TLF score is the average of the scores for all 9 dimensions.

A teacher's final IMPACT score is between 100 and 400. For teachers in grades 4 through 8, the IMPACT score assigns 50 percent weight to a teacher's value-added score and 35 or 45 percent weight to classroom evaluations, depending on the year. The remaining score depends on a teacher's rating on the "Commitment to School and Community" rating determined by the principal. Based on

[19] Value-added scores require a teacher's students have a prior test score available. These scores are only available starting from grade 3 through grade 8, and are only available for Math and English Language Arts (ELA). This means that only teachers in grades 4 through 8 in math and reading will have value-added scores available.

their numeric score, teachers receive a rating of "Ineffective" (score below 175), "Minimally Effective" (between 175 and 250), "Effective" (between 250 and 350), or "Highly Effective" (greater than 350). Teachers face large consequences based on their IMPACT rating. Highly Effective teachers receive one-off bonuses ranging from $5,000 to $25,000 depending on the school, grade, and subject taught. If teachers are Highly Effective a second year in a row, they receive permanent pay increases that range from $6,000 per year and possibly exceed $20,000 per year.[20] If a teacher is rated Minimally Effective, she experiences a pay freeze, meaning her salary does not increase as it normally would with each year of experience. She must also improve to Effective in the next year or be dismissed. Receiving a rating of Ineffective leads to immediate dismissal. Only 1.6 percent of teachers received a final rating of Ineffective in the years studied, whereas 12 percent received a Minimally Effective overall rating.

I observe a teacher's final value-added score, the date of each of her in-class performance evaluations, and the date on which she meets with her evaluator to review her performance. I also observe her years of experience and her highest degree earned. With this information, I calculate the number of days in which she is guaranteed not to receive an evaluation. I also calculate the daily probability of receiving an evaluation. The data are limited by the available teacher covariates. The only reliable covariate available for the entire sample is a teacher's years of experience.

1.4.2 Empirical Approach

My empirical model builds on the version of production uncertainty for a stepwise incentive. The first distinction is that, in DCPS, a teacher experience days in which she may be evaluated ("threat days") and days in which she is certain not to be evaluated ("no-threat days"). I assume her effort allocation on threat days will be different than her effort allocation on no-threat days. I also assume that her value-added score at the end of the year is the cumulative effect of her daily effort. Her choice of inputs is the solution to her utility maximization problem represented in Equation 1.4. Let N be the total number of instruction days and n be the number of no-threat days. Then let $\beta(x)$ be her daily marginal effect on student test scores as a function of her daily vector of inputs, x. An addi-

[20] Pay increases depend on a variety of factors, such as a teacher's current base pay, whether her school is a high-poverty school (60 percent or more of students receive free or reduced-price lunch), or if she teaches a high need subject. See Dee and Wyckoff (2015) for more details.

tional distinction from the theoretical model is that her choice of inputs before her evaluation will change as the probability of an evaluation changes, making x a function of p_t. Then her value-added score at the end of the year is

$$Y = n\beta(x') + \sum_{t=1}^{N-n} \beta(x(p_t)) + \varepsilon \tag{1.12}$$

Her individual score is measured with error ε, which I will assume is conditionally independent of the timing of her in-class evaluation. Equation 1.12 implicitly assumes that the probability of an evaluation affects a teacher's daily contribution, $\beta(x(p_t))$ linearly. There is some minimum daily effect for a teacher's choice of x, even if p_t is very small. Then I allow this effect to grow linearly with p_t. This assumption does not require that a teacher is modifying her effort linearly, only that its effects on value-added are approximately linear in p_t.[21]

When a teacher receives feedback, it may improve her daily effectiveness either by making good teaching less costly or by increasing the marginal productivity of her inputs. I can measure the cumulative effect of feedback on student outcomes. To do so, let v be an indicator for each evaluation, where $v = P1, P2, P3$ for her three principal evaluations and $v = M1, M2$ for her master educator evaluations. For feedback on evaluation v, she has m_v days in which her feedback affects her teaching by α_v per day. With this addition, her value-added is

$$Y = n\beta(x') + \sum_{t=1}^{N-n} \beta(x(p_t)) + \sum_{v} m_v\alpha_v + \varepsilon \tag{1.13}$$

Then a teacher's value-added at the end of the year is

$$Y = n\beta(x') + (N-n)\beta(x) + \beta(x) \sum_{t=1}^{N-n} p_t + \sum_{v} m_v\alpha_v + \varepsilon \tag{1.14}$$

This model specification allows me to identify a discontinuity in the effect of teacher effort at $p_t = 0$. If the effect of her inputs is continuous in p_t, then her daily effect on value-added as p_t approaches zero should be equal to her daily effect when she has no threat of an evaluation. In my specification, this would mean $\beta(x) = \beta(x')$.

[21] In other specifications, I allow the effect of evaluation probability to be quadratic or logarithmic, but neither specification alters my results qualitatively.

Key measures include how the probability of an evaluation affects teacher value-added and if there are discontinuous changes in her behavior after her evaluation. The parameter $\beta(x)$ represents the daily effect of an increase in p_t on student outcomes, and the difference $\beta(x) - \beta(x')$ is the discontinuous change in effort when a teacher is guaranteed to not have an evaluation. I can estimate $\beta(x)$, $\beta(x')$, and α_v using a standard OLS estimation.

1.4.3 Econometric Specification

In the simple case where a teacher receives only one evaluation in a year, identifying threat and no-threat days is straightforward. In the IMPACT program, evaluations occur in multiple pre-specified time windows. The structure of these windows provides two opportunities for no-threat time. This feature is useful because it provides two opportunities within the same year to measure how input incentives affect student test outcomes. Over the year, the curriculum and teaching priorities may change, potentially varying the observed effects of a pending evaluation. I use the two possible no-threat windows to allow the effect of high-stakes evaluation to change based on the season of the year.

My econometric specification builds directly off Equation 1.14. As before, the outcome variable of interest is a standardized value-added score on student reading and math tests, Y_{ijs}, for teacher i in year j at school s. I control for school-level characteristics using school fixed effects ϕ_s. The variable X_{ij} is a vector of annual experience dummies, up to 15 years of experience. Let $w = 1, 2$ indicate the no-threat window. The number of no-threat days in window w is n^w_{ijs}. The number of days between when a teacher receives feedback on evaluation v and student tests is m_v for $v \in \{P1, P2, P3, M1, M2\}$. I estimate the following equation:

$$Y_{ijs} = X_{ij}\Gamma + \phi_s + \sum_{w=1}^{2} \left(\beta^w(x') - \beta^w(x)\right) n^w_{ijs} + \sum_v m_v \alpha_v + \sum_{w=1}^{2} \beta^w(x) \sum_{t=1}^{N^w - n^w} p^w_t + \varepsilon_{ijs} \tag{1.15}$$

The error term in Equation 1.15 consists of the measurement error of a teacher's value-added and other unobservable factors that affect a teacher's value-added score. I assume that ε_{ijs} is conditionally independent of n^w_{ijs}. That is, $E[n^w_{ijs}\varepsilon_{ijs}|X_{ij}, \phi_s] = 0$. The assumption is that there are no unobservable characteristics of a teacher that are correlated with her value-added score and systematically change her

number of no-threat days. If evaluators systematically target low-quality teachers early in the year based on criteria that I cannot observe, then my results will be negatively biased. If the timing of evaluations is independent of the error term, the probability of an evaluation will be too. In order to separately identify the positive effects of evaluation feedback from the effects of no-threat time, I also assume that the space between when a teacher receives her evaluation and the day she receives feedback is conditionally independent of ε_{ijs}. Let d^v_{ijs} be the number of business days between when a principal received evaluation v and when she received her feedback. I assume $E[d^v_{ijs}\varepsilon_{ijs}|X_{ij}, \phi_s] = 0$. This is assuming that evaluators do not systematically change how long they wait to meet with teachers based on unobservable characteristics that correlate with teacher value-added. If evaluators meet sooner with good teachers, my estimated positive effects of receiving feedback will be biased upwards.

I estimate Equation 1.15 using ordinary least squares with clustered errors at the school-by-year level. I cluster at the school-by-year level because each year at each school is effectively a new random assignment to treatment. As a result, the error terms for individual teacher value-added scores are likely correlated within school and year.

1.4.4 Description and Calculation of Treatment Measures

In-class evaluations must occur within pre-specified time frames as depicted in Figure 1.1. The first principal evaluation must occur by December 1, the second must occur before March 15, and the third must occur before the end of the school year. The Master Educator evaluations split the school year: the first occurs before February 1 and the second occurs afterwards.

Once a teacher has received all of her possible evaluations in the current window, it is guaranteed that she will not have an evaluation until the next window begins. Figure 1.2 provides a simplified example. In the case depicted, a teacher must only receive two Master Educator evaluations. Her no-threat time is defined by the number of days from her first evaluation until the start of the next window. She again will have more no-threat days after her second evaluation. Because I am looking at how no-threat days affect student test outcomes, I only consider no-threat days that occur before students begin taking their tests for the year.

The actual possible no-threat windows in IMPACT are more complicated than Figure 1.2. No-threat time requires that *both* possible evaluations are completed, and no-threat days are only counted until the next possible window begins. Fig-

ure 1.3 provides an example of how no-threat time is calculated. Because the first principal and Master Educator evaluations are announced before-hand, I do not consider them to provide any evaluation threat. The first possible no-threat time starts from the time of the second principal evaluation and lasts until the start of the second Master Educator window. If a teacher does not receive her second principal evaluation before the start of the second Master Educator window, she will not have any no-threat days. The second no-threat window is possible from the end of her second Master Educator evaluation until the start of the window for the third principal evaluation. Because fewer than eight percent of teachers receive their third principal evaluation before the start of student testing, I do not consider the effects of possible no-threat time at this window. It is unclear if teachers have any real expectation of receiving an evaluation in this short time frame.

The probability of an evaluation changes throughout the school year. Teachers are effectively drawn without replacement, making the daily probability of an evaluation for a specific teacher increase as the school year progresses. I account for this by estimating the teacher's probability of being evaluated on each day t at her school s. Intuitively, if a teacher has not been evaluated by the last day of the window, she can be certain to receive her evaluation on the next day. I observe the date of each observation for each teacher, which I use to calculate how likely the remaining teachers are to be evaluated in each of the remaining days. Intuitively, a teacher knows that if tomorrow is the last day of an evaluation window and she has not been evaluated, she will be evaluated tomorrow. Two factors determine a teacher's estimate of the probability of being evaluated on any particular day: the number of teachers that remain to be evaluated and how many evaluations a teacher expects to be conducted. It is then straightforward to calculate evaluation probability if each remaining teacher has an equal probability.

As before, let v be an evaluation indicator, where v is $P1$, $P2$, or $P3$ for the principal evaluations and $M1$ or $M2$ for master educator evaluations. Then let a teacher's estimate of the number of evaluations to be conducted on day t at school s be $\hat{L}^v_{ts}$. If R^v_{ts} is the number of teachers who still need evaluation v on day t at school s, then each remaining teacher's probability of being evaluated is

$$p^v_{ts} = \frac{\hat{L}^v_{ts}}{R^v_{ts}}. \tag{1.16}$$

It is straightforward to determine the number of remaining teachers for an evaluation, R_{ts}^{v}, but estimating how many evaluations a teacher expects to be conducted, $\hat{L}_{ts}^{v}$, requires assumptions about what a teacher knows. If a teacher knew exactly how many evaluations would be conducted on every day, then $\hat{L}_{ts}^{v} = L_{ts}^{v}$. This is a strong assumption that is unlikely to be true, especially if evaluations are not evenly distributed within a window.

Principals tend to cluster their evaluations near the last third of the time window, which changes the expected number of daily evaluations to change over time as well. In the beginning of an observation window, teachers expect that principals will conduct few evaluations, but towards the end of the window, teachers expect more each day. On the other hand, master educators distribute their evaluations more evenly, so the expected number of evaluations remains constant. Figure 2.2 shows the overall distribution of evaluations across each window. While the master educators maintain a fairly uniform distribution, principals are very often conducting evaluations in the last third of the available time. The dip in evaluations in M2 around day 45 is a result of student testing days in April.

Instead of assuming teachers know exactly how many evaluations will be conducted on each day, I can allow a teacher to assume a uniform distribution of evaluations, or assume she is broadly aware of the trend in evaluations. I approximate the information available to a teacher by estimating the distribution of evaluations with a kernel density. The kernel smoothing approximates changes in the trend of daily evaluations that teachers notice. I estimate the model under both a uniform assumption and the kernel.

For many days in the year, a teacher has the possibility of either a principal evaluation or a master educator evaluation (or both). The two events are independent and in rare cases both occur on the same day for a single teacher. To determine the probability of any evaluation, I use the sum of their individual probabilities. For example, if a teacher has not yet had either $P1$ or $M1$ evaluations, her probability of *any* evaluation the next day is $p_{ts} = p_{ts}^{P1} + p_{ts}^{M1}$, but if she had already received her $P1$ evaluation, her probability is just $p_{ts} = p_{ts}^{M1}$.[22] Then in my specification, I use p_{ts}. Figure 1.5 shows the distribution of the cumulative probability of receiving any evaluation for each of the five evaluations. Two key features stand out. Because the measure is the cumulative probability across many days, my measure of evaluation probability is often larger than one.

[22] These additive probabilities are capped at 1, though the cap was rarely needed (it applied to 0.38 percent of all observations).

The other important feature is that the cumulative evaluation probability for principal evaluations is larger on average, which is the result of principals lumping evaluations near the end of the evaluation window.

1.5 Treatment Exogeneity

A primary object of the empirical analysis is to measure how "no-threat time" or the days without a potential evaluation affect teacher value-added. Central to this question is the exogeneity of no-threat time. While the statutory design of the policy would suggest that these measures should be exogenous, actual practice may produce variation in evaluation timing that is related to teacher quality. Targeted evaluation timing would invalidate the assumption that assignment to no-threat days is conditionally independent of teacher quality. If low-quality teachers systematically receive their evaluations early, assignment to no-threat days will correlate with lower student test scores and bias my results.

The first step in my empirical analysis is to assess the extent to which evaluators may target teachers. Anecdotally, principals are somewhat haphazard in determining when to evaluate each teacher, fitting in evaluations as time allows. While this may suggest they are not purposefully targeting specific teachers, principals may still do so inadvertently. Unlike principals, master educators are likely more methodical because their primary job is to conduct evaluations.

I can observe potential targeting by estimating the correlation between evaluation timing and measures of a teacher's quality, such as her previous year's value-added score. Even under a variety of tests, I find no evidence that either principals or master educators target teachers based on quality. If evaluation timing is independent of teacher quality, my treatments of no-threat days and the probability of an evaluation are exogenous.

My specification also identifies the positive effects of post-evaluation feedback by using variation in how much time elapses between an evaluation and when a teacher receives her feedback. If evaluators systematically provide prompt feedback to good teachers but are slow to give feedback to low-quality teachers, my estimates of the effect of evaluation feedback will be biased upward. I find no evidence of this form of targeting.

To test my identification assumptions, I use a variety of teacher characteristics to assess whether or not evaluations are timed based on teacher quality. I regress characteristics that are potentially observed by principals and master educators,

C_{ijs}, on the treatment variables, T_{ijs}. The regression specification is

$$T_{ijs} = \beta_0 + \phi_{sj} + \beta C_{ijs} + \varepsilon_{ijs}\,. \tag{1.17}$$

Because there may be school-by-year systematic differences in evaluation timing, ϕ_{sj} is a school-by-year fixed effect for school s. The characteristics I consider are a teacher's overall evaluation score in the previous year (calculated as the mean of all five evaluations), the teacher's value-added score in reading and math in the previous year, an indicator for whether she is a first-year teacher, and then observable scores from evaluations within the same year. The characteristics C_{ijs} are demeaned. The treatment variables T_{ijs} I consider are, first, no-threat time in the two possible windows. I also consider the specific principal and Master Educator evaluation timing, as well as targeting the space between an evaluation and the conference in which a teacher receives her feedback. Finally, I consider the treatment variable, the sum of the probabilities of evaluation across each window w: $\sum_t p^w_{tijs}$.

The results in Table 2.4 show that there is no evidence that principals or master educators are targeting specific teachers early. None of the coefficients have been adjusted for multiple hypothesis testing. The exogeneity checks in Table 2.4 may be limited by the relatively small sample size. My analysis is restricted to teachers in math and reading in grades four through eight. Using the same administrative data, Phipps and Wiseman (2017) conduct the same exogeneity checks on the full range of Washington DC teachers and also find no targeting.

1.6 Results

Table 1.4 shows the estimated effects of evaluation probability, no-threat time, and post-feedback time on student test outcomes for reading and math. The outcome variable is standardized teacher value-added scores.[23] The models vary by the method used to calculate the probability of evaluation, either assuming

[23]It is possible to use student test outcomes instead, and the qualitative results are the same. Using student test data introduces considerable measurement error in the independent variables, however, due to students switching teachers. Roughly 20 percent of students switch teachers within a year, but the exact timing of their switch is not known. In calculating no-threat days for a specific student's teacher, I am forced to loosely approximate when that student switched. Using teacher value-added instead keeps measurement error in the dependent variable, which does not introduce bias and unknown effects but it decreases my precision.

evaluations are uniformly distributed or using a kernel density estimate of their distribution.

Standardized teacher value-added scores are on a different scale than student test scores. Value-added is calculated by first predicting each student's test scores based on their previous year performance and other observable characteristics, and then calculating how much, on average, a teacher's students' performance differed from their predicted score. In other words, value-added is an annual estimation of a teacher's fixed effect on student outcomes. These fixed effects are then standardized, making the scale of teacher value-added different than the scale of standardized student test scores. In Washington DC, I find that a one standard deviation increase in teacher value-added corresponds to a 0.10 standard deviation increase in student reading and 0.13 standard deviation increase in student math. This estimate agrees with the larger empirical literature, where a one standard deviation increase in teacher value-added score usually corresponds to an increase of 0.11 standard deviations in student reading test scores and 0.14 standard deviations in student math scores (Hanushek and Rivkin, 2010).

The effects of no-threat time are negative for reading in Window 2, but not significant for Window 1. As seen in the second row of the first two columns of Table 1.4, an increase in the number of no-threat days reduces teacher value-added by 0.046 standard deviations, which is a 0.005 standard deviation decrease in student reading scores per day of no-threat time. For math, Window 1 no-threat days reduce teacher value-added by 0.036 standard deviations (row one of columns three and four), which is a decrease of about 0.005 student standard deviations. There is no statistically significant effect in math for Window 2.

The counterfactual for teachers with no-threat days is teachers with threat-days, adjusted for the probability of an evaluation. This implies that the measured effects of no-threat days will depend on the activities and effectiveness of teachers with the possibility of an evaluation. A feature of my empirical approach is that it allows the counterfactual teacher effects to differ between the two windows. The different effects of Windows 1 and 2 likely reflect differences in the effect of curriculum on test scores depending on the season. In mathematics, concepts build, which could explain why Window 1 is more important for math preparation. For reading, the proximity of Window 2 to the test may reflect the importance of the time just before test taking.

Feedback from an evaluator has positive effects, though it is not always significant or meaningfully large. For reading, feedback from the last Master Educator

evaluation has a significantly positive effect on student test scores, as seen in row four of columns one and two in Table 1.4. Looking at Figure 1.3, the second Master Educator evaluation is the evaluation that is most likely to distinguish between threat and no-threat days. That is, all teachers will receive their second principal evaluation by March 15, but many teachers will not receive their Master Educator evaluation until after the start of the third principal evaluation window (and possibly only after student tests). As a result, there is greater variation between no-threat time and Master Educator feedback in Window 2 than between no-threat time and principal feedback. The increased variation may contribute to the observed positive training effects by more clearly separating the two effects. It is also possible that feedback from the Master Educator just a month before students take their tests may have a particularly salient effect. For math, there is a similar pattern. Feedback from a principal in the second principal evaluation has a significant, positive effect on student math scores, which also happens to be the evaluation that most distinguishes no-threat time in the first window.

The average teacher has 31 school days between when her students take their standardized test and when she received feedback for her second principal evaluation. This implies a total increase of 0.068 student standard deviations in math due to improvements teachers make after receiving detailed feedback from their evaluation. In Taylor and Tyler (2012), the authors find that student math scores improve by 0.064 standard deviations for the year in which teachers were evaluated. Unlike much of the literature on teacher interventions, my results also show a positive effect on student reading test scores.[24] I find that the second master educator evaluation feedback improves student reading scores. Not all teachers receive their second master educator evaluation before their students take standardized tests. Among those that do, teachers receive their feedback 17 days before the test date, on average. The overall improvement in student reading scores is 0.031 standard deviations.

The cumulative probability of an evaluation has no significant effect on student outcomes, regardless of the estimation method. To put the effects into perspective, the median daily probability of any evaluation is 10 percent. An additional day of threat time adds roughly 10 percentage points to the cumulative probability of evaluation. A 10 percentage point increase in the cumulative probability of being evaluated increases or decreases student reading scores by 0.0001 standard devi-

[24] Taylor and Tyler (2012) do not find positive effects in reading, which coincides with much of the literature on teacher interventions.

ations and decreases math scores by 0.0006 standard deviations, at most. In other specifications, I allow the effect of cumulative probability to vary by evaluation, but there is no meaningful difference from the results shown in Table 1.4.

The small effects that I observe from evaluation probability suggest that teachers are responding to evaluation threat along an extensive margin, not necessarily an intensive margin. If a teacher's effectiveness varied based on the intensity of her effort, the intensive margin would be sensitive to the probability of an evaluation. On the other hand, discontinuous changes in her effectiveness suggest changes to the types of activities she employs in her classroom. The coefficient for no-threat days is a discontinuity in the effect of a teacher's effort bundle after her evaluation.

It is possible to measure how teachers modify their practice in preparation for an evaluation. Phipps and Wiseman (2017) use the same DCPS data to estimate that a 10 percentage point increase in the probability of an evaluation leads to a 0.03-0.06 standard deviation increase in a teacher's evaluation score. This result supports the notion that teachers are responding to the probability of an evaluation, even though those responses have no measured effect on student test outcomes. In all, I interpret these results to mean that teachers select a set of inputs when there is a possible evaluation, and then make minor improvements to those inputs as the probability of an evaluation increases. When they are certain there will be no evaluation, teachers modify their selection of inputs. The change in their selection of inputs after their evaluation has a negative effect on student tested outcomes.

The results presented are a within-year effect, which is a novel contribution to the empirical literature on teacher incentives. My results imply that a teacher's daily behavior can have large within-year effects on student outcomes. For an upper bound on what students gain per day, Goodman (2015) finds that a single absence can cause a student's test score to fall by 0.05 standard deviations. In a nationally representative sample, Hill et al. (2008) estimate that, over the grades in this sample, the average student gain in student math scores ranges between 0.32 to 0.52 standard deviations and between 0.24 and 0.40 in reading. There are usually 122 school days before tests. If learning were uniformly distributed across the school year, students should gain between 0.003 and 0.004 standard deviations per day in math and 0.002 and 0.003 in reading. The students of a teacher with an additional no-threat day will score 0.005 standard deviations less than their peers, depending on the time of year.

My results, combined with Hill et al. (2008), suggests potentially large fluctuations in how a teacher's daily activity affects student outcomes. These results also touch on the potential effect of extending school years and school days; the wide variation in daily learning implies school day and year extensions should be approached with caution and thought towards the quality of those time extensions as well as their quantity.

The structure of IMPACT implies that some teachers face very severe consequences for a Minimally Effective rating this year if they were Minimally Effective in the previous year. Such differences in the implicit incentive for a teacher may change her response to no-threat time. To test for this possibility, I could allow the effect of no-threat days to vary based on the teacher's IMPACT rating in the previous year. However, the small sample size and the relatively few treated teachers prevents me from conducting meaningful heterogeneous effect tests.

In my preferred specification, experience is a vector of dummy variables for each year of experience. Other literature uses a quadratic form (Taylor and Tyler, 2012). My results are not sensitive to either specification. I do not include other teacher covariates because of fairly restrictive data limitations on other characteristics like teacher race or highest degree obtained. It is also possible to estimate the results by including a teacher fixed-effect. Doing so will control for unobserved teacher characteristics that are persistent across years. The trade-off is that my sample will only include teachers that appear at least twice. The results of this specification are qualitatively the same and statistically significant, but the magnitude of no-threat days increases slightly. The increase is not statistically significant. I interpret these results as further evidence supporting my identification assumption. If evaluators were targeting teachers based on persistent unobservable characteristics, the fixed-effect results would be zero. The fact that my estimates do not meaningfully change supports the assumption that evaluators are not targeting teachers based on characteristics that I cannot observe in the data.

Most studies on teacher interventions consider the effect across the entire school year. To facilitate a comparison with other work, I calculate the overall effect of evaluations in DCPS by multiplying the number of days in which an evaluation was possible in a window by the daily effect. The average teacher has 16.2 threat days in Window 2, which implies student reading test scores improved by up to 0.074 standard deviations. For Window 1, the average teacher in an average year had 19 threat days, implying student math scores improved up to 0.089

standard deviations. For comparison, Dobbie and Fryer (2013) find that among high-demand charter schools, students in schools that use unannounced in-class teacher evaluations scored 0.048 standard deviations higher in reading and 0.044 standard deviations higher in math.

My results and those of Dee and Wyckoff (2015) are complements. Using the same administrative data, they find that teachers under the threat of dismissal improve their students' scores an average of 0.029 standard deviations relative to similar teachers that were just above the Minimally Effective threshold in the previous year. However, their results are only statistically significant for the last year of data, which is the 2011-12 school year. In the last year, the authors find that the threat of dismissal improved student test scores by roughly 0.066 standard deviations. This is also the only year in which the authors find that dismissal threat caused an improvement in teacher in-class evaluation scores. These results are consistent with the idea that as teachers learned how to improve their practice in response to the in-class observation rubric, their student outcomes also improved. Together with their analysis, my results imply that much of the observed effect of IMPACT on student outcomes is driven by teacher responses to the possibility of an in-class observation.

1.7 Production Uncertainty and Alternative Explanations

My empirical results confirm the implication in the broader empirical literature that in-class evaluations are a key mechanism for improving student test outcomes. My empirical results also show a discontinuity in the effect of teacher effort when the probability of an evaluation is zero. In what follows, I argue that the discontinuity represents a shift in teacher inputs along the extensive margin. But given the design of IMPACT, where a teacher's bonus still depends on her value-added score, changes along the extensive margin should have a weakly positive effect on student test outcomes if she knows the production function. Without production uncertainty, what she does when there's an extremely low probability of an evaluation should not differ much from what she does when there is zero probability of an evaluation.

I use the discontinuity of teacher effort and the design of the IMPACT program to argue that my results support the narrative that teachers are uncertain about

the test score production function. Using the theoretical model of a step-wise incentive presented earlier, I show how a sudden change in the marginal benefit of improving student test scores could explain the discontinuous jump in her daily effectiveness. Before her evaluations, the possibility of a randomly good evaluation effectively diminishes the marginal benefit of improving her test score. But this effect is gone after her evaluation. If she knows the test score production function, her effort after her evaluation would likely have either no effect or an increased effect on student test scores. However, if she is uncertain about the production function for test scores, no-threat time should have a negative effect on student test outcomes. This argument would not be true if there were two separate bonuses for in-class observation scores and value-added scores, as in some performance pay programs.

1.7.1 Empirical Results as Evidence of Production Uncertainty

The results in Table 1.4 show that teacher responses to the probability of an evaluation have no measurable effect on student outcomes except when the probability is zero. Yet teachers modify their behavior in response to the probability of an evaluation. Phipps and Wiseman (2017) use the same data and show that teachers improve their in-class observation score when there is a higher probability of an evaluation. Anecdotal evidence suggests that in preparation for a possible evaluation, teachers first select a teaching style that will perform well in an in-class evaluation. As Phipps and Wiseman (2017) explain, teachers may then make minor adjustments in their daily activities as an evaluation becomes more likely. For example, it is common for teachers to spend some time in the morning with their students as they eat breakfast at school. If an evaluation is very likely, teachers may forgo meeting students at breakfast and instead work to prepare their classroom to meet specific components on the evaluation rubric, such as clearly writing and stating the day's objectives on the whiteboard. Or teachers may spend additional time preparing specific questions to "check for understanding," a specific component of their evaluation score.

I consider adjustments that do not alter the teaching style or overall pedagogical approach to be adjustments along the intensive margin: teachers are not altering their choice of inputs but only the intensity of an input. Adjustments along the extensive margin are changes to the teaching style and approach that reflect a fundamental change in which tools a teacher selects from her toolkit. My empirical results show that changes along the intensive margin have no measur-

able effect on student outcomes, but changes in teaching style when there is no possibility of an evaluation have meaningful effects on student tested outcomes. The majority of the effects I find are driven by teacher choices along the extensive margin, not the intensive margin.

To be consistent with the empirical finding that teachers have a discontinuous jump in their effect on student test scores after an evaluation, the theory needs to have a discontinuous change in a teacher's incentives. During no-threat days, the incentive structure for a teacher changes in ways that change the marginal benefit of an input i. In a model without production uncertainty, these changes should cause teachers to switch to inputs that improve student test scores. To see why, consider two teaching styles, a and b. Let a be "teaching to the test" and b be "objective-based learning." Before their evaluation, teachers choose their inputs in order to improve their student test scores, $f(a,b)$, and their in-class observation score, $g(a,b)$. Let f_a and f_b be the marginal effect of teaching to the test and objective-based learning on student test scores, and g_a and g_b are their effect on the teacher's in-class observation score. Understandably, teaching to the test does not improve an in-class evaluation score ($g_a = 0$), but it does improve student test outcomes by some amount f_a. I also allow the marginal benefit of objective-based learning to depend on the probability of an evaluation, λ, where there is no marginal benefit if there is no possibility of an evaluation, $g_b(\lambda = 0) = 0$. Finally, assume her value-added and in-class observation are measured with noise with variance σ_f and σ_g, respectively. Equation 1.7 provides her first-order conditions. To simplify the expression, let $\alpha = \phi\left(\frac{S^* - \bar{S}}{\sqrt{\sigma_f + \sigma_g}}\right)$, which is the standard normal probability distribution function. Then each input affects the probability of earning a bonus in the following way:

$$\frac{dP}{da} = \frac{\alpha}{\sqrt{\sigma_f + \sigma_g}} f_a \tag{1.18}$$

$$\frac{dP}{db} = \frac{\alpha}{\sqrt{\sigma_f + \sigma_g}} \left[f_b + g_b(\lambda)\right] \tag{1.19}$$

After her evaluation, there is no longer any measurement error about her in-class observation score ($\sigma_g = 0$). In addition, her inputs no longer affect her in-class observation score ($g_b = 0$). The marginal effect of each input on the

probability of earning a bonus is:

$$\frac{dP}{da} = \frac{\alpha'}{\sqrt{\sigma_f}} f_a \tag{1.20}$$

$$\frac{dP}{db} = \frac{\alpha'}{\sqrt{\sigma_f}} f_b \tag{1.21}$$

Under reasonable conditions, this represents an increase in the marginal benefit of inputs that improve student test scores. That is, $\frac{\alpha}{\sqrt{\sigma_f+\sigma_g}} < \frac{\alpha'}{\sqrt{\sigma_f}}$.[25] The reason is that the measurement noise from her in-class evaluation score provided the possibility that she would get a randomly high in-class observation score. After her evaluation, this is no longer a possibility and increasing her student test scores has a greater effect on the probability of earning a bonus.

Before her evaluation, changes in the probability of an evaluation (λ) would have continuous changes on her choice of inputs. However, due to the abrupt decrease in her uncertainty about her overall IMPACT score upon completing her evaluation, there would be a discontinuous change in her effort, as observed. If a teacher chooses to switch her inputs (i.e. changes along the extensive margin), her input changes are likely to have a positive effect on student test scores, not negative. Importantly, if there is no measurement noise for in-class observation scores, there should be no discontinuous changes in a teacher's daily effect on student test scores. In the scenario provided, a teacher may choose objective-centered teaching more than teaching to the test before her evaluation because objective-centered teaching would have larger returns on her in-class evaluation. After her evaluation, she is going to increase teaching to the test as long as it has a better marginal effect on student test scores: $f_a > f_b$.[26]

[25] $\bar{S}$ needs to be sufficiently close to S^* as a function of σ_f for this condition to be true. If I assume all teachers have the same skill, I can use estimates of the non-persistent randomness in value-added scores to estimate how many teachers choose $\bar{S}$ sufficiently close to S^*. Importantly, this assumes no heterogeneity in teacher skill, which is extremely conservative. Yet even under this conservative assumption, I estimate that at least 75 percent of teachers choose $\bar{S}$ close enough to the cutoff score for either "Minimally Effective" and "Highly Effective" such that the condition is met.

[26] I have implicitly assumed that the marginal costs of teaching to the test and object-centered teaching are the same in this discussion. Regardless of which approach a teacher uses, there will be lesson preparation time. For veteran teachers, lesson preparation time is likely to be minimal, and switching to a teach-to-the-test approach may even incur more planning costs. Barring extremely unhelpful teaching behaviors (like watching non-educational movies), the difference in preparation costs will arguably be small. Furthermore, considering teachers as motivated agents, they have an implicit desire to improve student knowledge. In general, teachers do not consider

My empirical results show a discontinuous *negative* effect after a teacher's evaluation. This outcome is perfectly consistent with my model of production uncertainty. To see why, suppose there is only production uncertainty for how objective-centered teaching affects student test scores. Let δ be the variance in the marginal effect of objective-centered teaching. Equation 1.7 provides her marginal benefit for each input:

$$\frac{dP}{da} = \frac{\alpha}{\sqrt{\sigma_f + \sigma_b + b^2\delta}}(f_a) \tag{1.22}$$

$$\frac{dP}{db} = \frac{\alpha}{\sqrt{\sigma_f + \sigma_b + b^2\delta}}\left[f_b + g_b(\lambda) - \bar{S}B\frac{\delta}{\sigma_f + \sigma_g + b^2\delta}\right] \tag{1.23}$$

After her evaluation there is no measurement noise for her in-class observation score. This has the same effect on α. The key difference is the abrupt change in the flattening cost. After her evaluation, the uncertainty about how much objective-oriented teaching improves test scores becomes more relevant: the second term has a sharp increase. After her evaluation, her marginal benefit for each input is now:

$$\frac{dP}{da} = \frac{\alpha}{\sqrt{\sigma_f + b^2\delta}}(f_a) \tag{1.24}$$

$$\frac{dP}{db} = \frac{\alpha}{\sqrt{\sigma_f + b^2\delta}}\left[f_b - \bar{S}B\frac{\delta}{\sigma_f + b^2\delta}\right] \tag{1.25}$$

As B increases, the discontinuous effect after an evaluation is more pronounced. Again, if there is no measurement noise for in-class evaluations, there should be no discontinuous response after an evaluation.

The key insight from production uncertainty is that a teacher may switch to an input that has *lower* marginal productivity. If teaching to the test has a smaller average effect on student outcomes than objective-oriented teaching ($f_a < f_b$), a teacher may still choose to teach to the test after her evaluation because there is little uncertainty in its marginal effect on student test outcomes. Before her

teaching to the test to be the right way to teach. For motivated agents, this would suggest that teaching to the test has additional costs to a teachers sense of identity as a quality teacher.

evaluation, as long as $g_b(\lambda)$ is sufficiently large, she will use the objective-oriented style. But after her evaluation, when the only salient incentive is her students' test scores, her uncertainty about the benefits of using the objective-oriented style may induce her to switch to a less effective teaching style. This switch can occur without decreasing her overall effort and without decreasing her own personal costs, making the incentive's effect unintentionally perverse.

The theoretical conclusion that post-evaluation effort should improve student test scores when there is no production uncertainty is an outcome that is unique to the structure of the IMPACT program. If there were two separate bonuses, one for high value-added scores and another for high in-class observation scores, the theoretical hypothesis would not be so clear. But because the IMPACT bonus is jointly determined by a linear combination of value-added and in-class observation scores, the two measures become near-perfect substitutes in affecting the probability of receiving a bonus. Without this feature, there would be no change in the marginal benefit of inputs pre- and post-evaluation, and the theoretical results above would not hold.

1.7.2 Alternative Explanations

While my results are consistent with my model of production uncertainty, there are other potential reasons why in-class evaluations would have a disproportionate effect on student outcomes. Outside the IMPACT context, even given the demonstrated positive effects of in-class evaluations, there are a variety of alternative explanations. I address the most common explanations both within the context of IMPACT and more broadly. In general, the evidence appears to support my hypothesis that uncertainty about the test production function is a driving factor in the effectiveness of high-stakes in-class observations.

Differences in Measurement Noise:

More generally, the pattern observed in the empirical literature that incentive programs improve test scores more when they include an in-class component might be explained without production uncertainty if there is a large difference in measurement noise between value-added and in-class observation scores. In Washington DC and more generally, it is not clear that in-class evaluations are measured with less noise than teacher value-added scores.

To assess the reliability of in-class observation scores, psychometricians use generalizability theory to deconstruct score variance into its contributing sources, similar to analysis of variance (Cronbach, 1972). In an internal report on the reliability of evaluation scores within DCPS, researchers found that a teacher accounts for 20 to 30 percent of evaluation score variance, which implies a reliability rating of 0.20 to 0.30 for each evaluation. This means that for the same teacher observed on different days with a different evaluator, her scores will have a correlation of 0.20 to 0.30. With the same evaluator, the reliability increases to roughly 0.28 to 0.38 (Meyer, 2016).[27] The reliability of in-class evaluations in DCPS is very similar to other, larger studies in other school districts (Kane and Staiger, 2012).[28]

I use an approach similar to Chetty et al. (2014) to assess the reliability of value-added scores within DCPS. I look at the autocorrelation in value-added scores for each teacher across years and find that reading scores have a 35 percent correlation with the previous year and 30 percent correlation with two years back. In math, value-added correlates 33 percent within one year and 28 percent after two years. In DCPS, the sample size is relatively small, but the results are similar to what Chetty et al. (2014) find.[29] Overall, value-added scores are not clearly a noisier measure than in-class observations.

Salience and Commitment:

It is also possible that teachers struggle to follow through on their commitment to improving student test scores. Relatedly, in-class evaluations may be more salient to teachers – seeing other teachers receive evaluations or seeing the principal in the halls – such that teachers are more responsive to the incentive. However, this

[27]The reliability ratings shown are for a single evaluation. It is also common to consider the reliability of the year's measures, which is the average of five evaluations. This provides a between-year reliability measure. With five observations, the reliability between years increases considerably. However, because each observation has equal weight on the final IMPACT score, a teacher's effort response depends on the noisiness of each evaluation, not the noisiness of the average.

[28]The Measuring Effective Teaching (MET) project, funded by the Bill and Melinda Gates Foundation, is a carefully designed, large-scale randomized experiment to test several in-class observation rubrics and how well their measures correlate with teacher value-added. The MET study finds that only 15-30 percent of all score variation can be explained by teacher fixed effects (Kane and Staiger, 2012). This implies a per-evaluation reliability rating of only 0.15-0.30.

[29]Chetty et al. (2014) use 20 years of student-level data in a large urban school district. They predict the value-added for each teacher using test score data from other years to test the reliability of value-added for a single teacher across years. For math, the authors estimate that value-added scores from the previous year have a 0.45 correlation with the current year, and this correlation decreases over time leveling off at 0.25 6-10 years out. In reading, the previous year value-added scores have a 0.30 correlation with the current year, which plateaus at 0.15 after six years.

does not explain why the observed effects of no-threat time occur in the months just prior to tests. Teachers should be aware of upcoming standardized testing in the months where my observed effects are strongest.

A detailed survey administered to teachers at DCPS asked about school-wide efforts to use interim assessments ("formative assessments") of student work against district-wide standards.[30] In response to the statement, "Teachers at my school track the performance of their students toward measurable standards," 75 percent of teachers said they agreed or strongly agreed, and 18 percent said they somewhat agreed. Other questions asked if teachers have materials for formative assessments and if they are provided time to conduct and analyze the results of formative assessments. The responses to these questions are similarly positive. The overall implication of the survey results is teachers are evaluating student progress towards learning standards and they are provided the materials and time to do so. It would seem unlikely that teachers lack appropriate reminders or that student test scores are not sufficiently salient.[31]

Culture Effects:

Teachers may be sensitive to their ranking relative to other teachers, and it is possible that teachers give greater weight to in-class observations than to value-added measures because it may represent a truer measure of their efficacy, something commonly referred to as face validity. If this were the case, a school's culture could induce greater responses to in-class observations beyond the financial incentive. There are two reasons this seems unlikely. In-class observation scores are noisy, meaning many events outside the teacher's control affect their score. In a DCPS survey, teachers largely disagreed with the statement "At my school, evaluation ratings are accurate reflections of teacher effectiveness." Roughly 40 percent of teachers disagreed to some extent, which is indicative of a culture that does not whole-heartedly consider in-class evaluations a reflection of teacher quality.

[30] The survey was administered in the 2015-16 school year. While I use data from the 2009-10 through 2011-12 years because changes to the IMPACT system make it infeasible to identify and use no-threat time in years after 2012, formative assessments are a practice that have been used for years in DCPS.

[31] Salience can refer to an inability to make the multiple calculations from daily activity to final test scores, as when consumers fail to accurately account for sales tax when making purchasing decisions (Chetty et al., 2009). In this regard, agent production uncertainty is a mathematical representation of salience. The added steps in deciphering the production function of test scores leads to uncertainty about the marginal productivity of inputs.

The second reason culture effects do not seem to account for the disproportionate effect of in-class observations comes from a paper analyzing the same DCPS data. In Phipps and Wiseman (2017), the authors find a significant decrease in teacher preparation for an evaluation when it is their last.[32] This effect is stronger for teachers with higher previous evaluation ratings, but does not appear for teachers who are under dismissal threat. Teachers appear to respond strategically to their last evaluation depending on their success on prior evaluations and depending on the stakes of the evaluation. This evidence suggests that teacher responses to in-class observations are, to a measurable extent, driven by incentives, not culture alone.

1.8 Conclusion

While teachers play an important role in student outcomes, improving teacher quality through policy has proven difficult. As part of their TIF award agreement, ten school districts randomly selected schools to implement a teacher performance pay program. The characteristics of these programs differed by district, potentially allowing researchers the opportunity to identify key characteristics of effective performance pay. Wellington et al. (2016) compare programs along six dimensions ranging from incentive size to teachers' understanding of the performance program. Notably, the use of high-stakes unannounced in-class observations was not a dimension they considered. They conclude that "...none of the characteristics [they] examined could help explain observed differences in student achievement impacts across districts." If teacher performance incentives are to be a viable policy solution, the research priority is to identify what actually works.

In general, the available evidence shows teacher performance incentives are more effective when they include rigorous in-class observations. This result is paradoxical when assuming teachers know the test score production function. By relaxing this assumption, I provide theoretical evidence why incentives based on in-class observations may be more effective than test-based incentives. In an innovative approach, my empirical results show substantial gains in student learning as a result of high-stakes unannounced in-class evaluations. A feature of my approach is that it separately identifies the incentive effects of in-class evalua-

[32]The authors find no similar effect for earlier evaluations in the year, which is important for the purposes of this paper. For the vast majority of teachers, the third principal evaluation is their last evaluation of the year. As such, strategic responses do not appear to affect teacher preparations or weight given to the evaluations that are part of this paper's sample.

tions from their training effects. In addition, while other policy initiatives rarely measurably improve student reading test scores, I find significant effects in both reading and math.

Overall, the theoretical argument and empirical evidence presented in this study indicate high-powered incentives based on rigorous in-class evaluations are effective tools for improving teacher quality. Yet the potential downsides of such high-stakes teacher evaluation systems still remain. If poorly designed, teacher evaluations can encourage gaming or over-emphasis on a single component of the evaluation rubric. Teachers (and their unions) may be reticent to forfeit autonomy over their teaching style, creating costly controversy. Rigorous evaluations can be costly, as demonstrated in recent modifications to the IMPACT program reducing the number of formal teacher evaluations. This study will aid policy makers in deciding how to make such trade-offs.

Table 1.1: Teacher performance incentives in the US by effectiveness and use of in-class evaluations.

Differentiated In-Class Evaluations	**Improved Student Scores**		
	Yes	Mixed	No
Yes	3	0	0
Some	0	2	0
No	0	0	2

Notes - There is an apparent correlation between effective teacher incentive programs and the use of in-class evaluations. This is not a comprehensive review of teacher incentive programs in the US. The programs in this table were selected based on their emphasis on individual-level measures of teacher performance. Many other programs use large school-level bonuses, which I have excluded. The analyses in the Yes/Yes cell are Dee and Wyckoff (2015); Dee and Keys (2004); Hudson (2010); in the Mixed/Some cell are Sojourner et al. (2014); Wellington et al. (2016); in the No/No cell are Briggs et al. (2014); Springer (2010).

Table 1.2: Summary statistics of no-threat time and cumulative evaluation probability.

	n^1	n^2	$\sum\sum p_t^w$	$\sum p_t^{P2}$	$\sum p_t^{P3}$	$\sum p_t^{M2}$
Overall Mean	2.09	1.17	2.05	0.72	0.13	0.58
$\% > 0$	21%	19%	100%	100%	81%	100%
If treatment > 0						
Mean	9.87	6.23	2.05	0.72	0.16	0.58
Std Dev	7.49	3.96	1.10	0.59	0.20	0.40
Minimum	1	1	0.20	0.01	0.00	0.01
Median	8	5	1.88	0.56	0.08	0.54
Max	29	17	7	4	2	2
N	1182	1182	1182	1182	1182	1182

Notes - n_1 indicates the number of no-threat days – days in which a teacher is guaranteed no evaluation – in Window 1, which starts December 1 and ends February 1. The variable n_2 is no-threat days in Window 2, which starts February 1 and ends March 15. $\sum\sum p_t^w$ is the cumulative probability of receiving an evaluation in both windows. Because it is cumulative, this variable often exceeds 1. $\sum p_t^{P2}$, $\sum p_t^{P3}$, and $\sum p_t^{M1}$ are the cumulative probabilities for each evaluation. Because probability is only counted before students take their tests, $\sum p_t^{P3}$ is very small because its window begins only days before students take their tests. Notice that $\sum p_t^{P2}$ is larger than $\sum p_t^{M2}$ because principal evaluations are often clustered at the end of the window.

Table 1.3: Estimates of potential targeting by principals and master educators in evaluation timing.

	n_1	n_2	n^{P1}	n^{P2}	n^{M1}	n^{M2}	$t^{P1} - t^{P1}_{conf}$	$t^{P2} - t^{P2}_{conf}$	$t^{M1} - t^{M1}_{conf}$	$t^{M2} - t^{M2}_{conf}$
Previous Evaluation Score	0.41	-0.005	-0.833	0.282	0.121	0.542	0.388	0.768	2.48	-0.443
	(0.51)	(0.25)	(1.38)	(1.02)	(2.61)	(1.34)	(0.76)	(0.93)	(2.66)	(0.52)
Previous Reading VA	0.497	-0.163	-0.617	0.186	0.2	0.948	0.241	-0.253	1.75	-0.46
	(0.57)	(0.32)	(1.00)	(0.84)	(2.60)	(1.69)	(1.20)	(0.63)	(2.03)	(0.87)
Previous Math VA	0.712	0.111	-1.078	0.38	1.365	0.67	-0.265	0.336	2.754	-0.278
	(0.90)	(0.42)	(1.43)	(1.12)	(3.02)	(1.87)	(1.84)	(0.71)	(3.49)	(1.03)
First-Year	-0.368	-0.045	-0.571	-0.293	-0.39	-0.267	-0.61	-1.053	-1.034	-0.624
	(1.06)	(0.37)	(1.25)	(1.46)	(3.32)	(2.06)	(0.69)	(0.78)	(0.83)	(0.56)
1-3 years of experience	-0.266	-0.239	1.172	0.127	-0.405	-0.639	0.006	0.499	-1.818*	-0.299
	(0.62)	(0.27)	(0.89)	(0.98)	(2.24)	(1.47)	(0.70)	(0.51)	(0.98)	(0.48)
Score from 1st Principal Eval	0.619	-0.098	-0.197	0.681	0.868	0.388	0.107	0.105	1.036	-0.389
	(0.79)	(0.20)	(0.72)	(1.08)	(1.82)	(1.04)	(0.58)	(0.49)	(0.79)	(0.48)
Score from 2nd Prinicpal Eval	-0.126	-0.102	-0.101	-0.745	1.999	0.973	-0.423	-0.732	1.911	-0.522
	(0.48)	(0.18)	(0.67)	(0.85)	(1.80)	(0.99)	(0.45)	(0.52)	(1.54)	(0.43)
Score from 1st ME Eval	0.754*	0.054	0.931	1.173	-2.147	0.317	0.998*	0.135	0.974	-0.644*
	(0.44)	(0.17)	(0.58)	(0.71)	(1.59)	(0.88)	(0.54)	(0.40)	(1.25)	(0.39)

Notes - The values shown are the correlation between the treatment variable (columns) and the observable characteristic (rows). All variables except First-Year are centered around the school mean to control for school fixed effects. Significance levels do not make any multiple hypothesis corrections. n_1 and n_2 are the number of no-threat days in Windows 1 and 2. Window 1 starts December 1 and ends February 1; Window 2 starts February 1 and ends March 15. n^{P1}, n^{P2}, n^{P3}, n^{M1}, and n^{M2} are proxies for the timing of each respective evaluation within its designated window. The differences variables, $t^{P1} - t^{P1}_{conf}$, are the time gap between an evaluation and its respective feedback conference.

*** Significant at the 1 percent level

** Significant at the 5 percent level

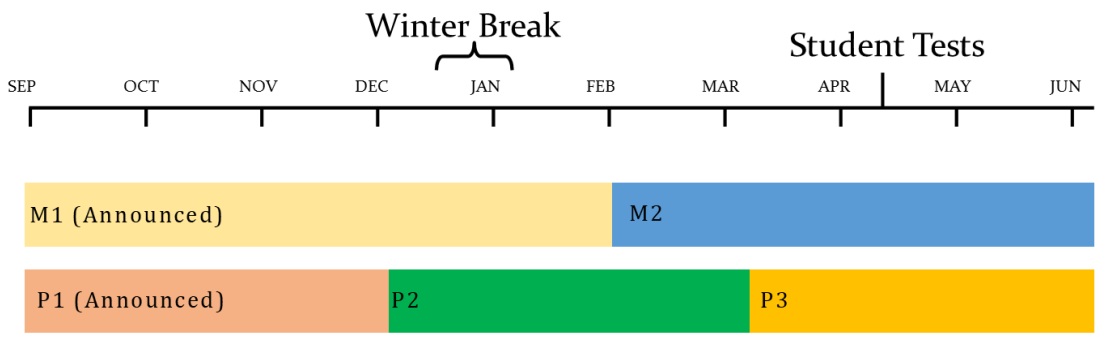
* Significant at the 10 percent level

Table 1.4: Effect of the possibility of an evaluation and evaluation feedback on teacher value-added in reading and math.

	Reading Value-Added		Math Value-Added	
Window 1 No-Threat Days, n_1	0.009	0.009	-0.036**	-0.034*
	(0.015)	(0.015)	(0.017)	(0.017)
Window 2 No-Threat Days n_2	-0.046**	-0.0456**	-0.0048	-0.0049
	(0.0183)	(0.0183)	(0.0221)	(0.0223)
Days after $P2$ feedback, m_{P2}	0.002	0.0016	0.0171*	0.0167*
	(0.0087)	(0.0088)	(0.0091)	(0.0090)
Days after $M1$ feedback, m_{M1}	0.0004	0.0002	0.0016	0.0019
	(0.0017)	(0.0017)	(0.0019)	(0.0019)
Days after $M2$ feedback, m_{M2}	0.0143***	0.0132**	0.0025	0.0044
	(0.0055)	(0.0053)	(0.0059)	(0.0059)
$\sum\sum p_t^w$	0.0148	-0.0107	-0.0633	-0.0261
	(0.0490)	(0.0407)	(0.0467)	(0.0416)
Probability Method	Uniform	Kernel	Uniform	Kernel
N	822	822	802	802

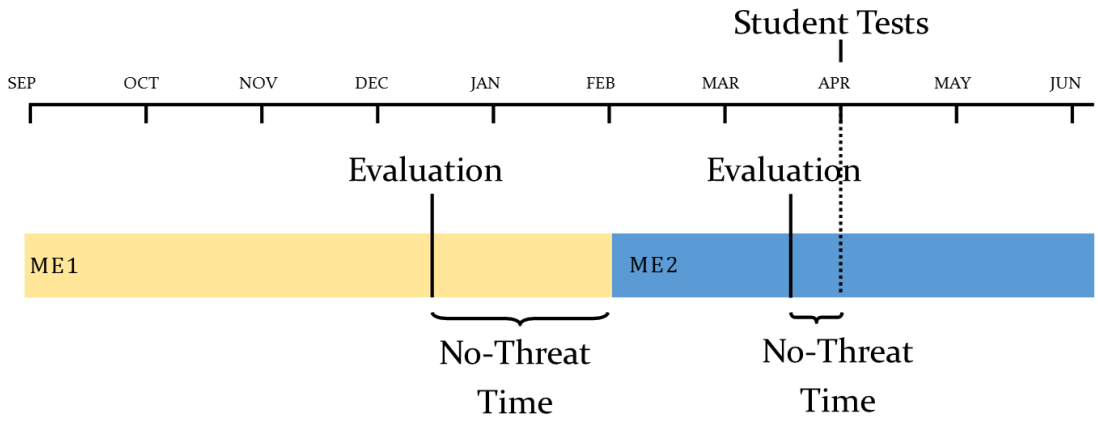
Notes - Effects are measured in standard deviations of teacher value-added scores, which is different than standard deviations in student test scores. A standard deviation increase in teacher value-added is roughly a 0.10 standard deviation increase in student reading scores and 0.13 standard deviation increase in student math scores.
*** Significant at the 1 percent level
** Significant at the 5 percent level
* Significant at the 10 percent level

Figure 1.1: Depiction of each evaluation window in DCPS.



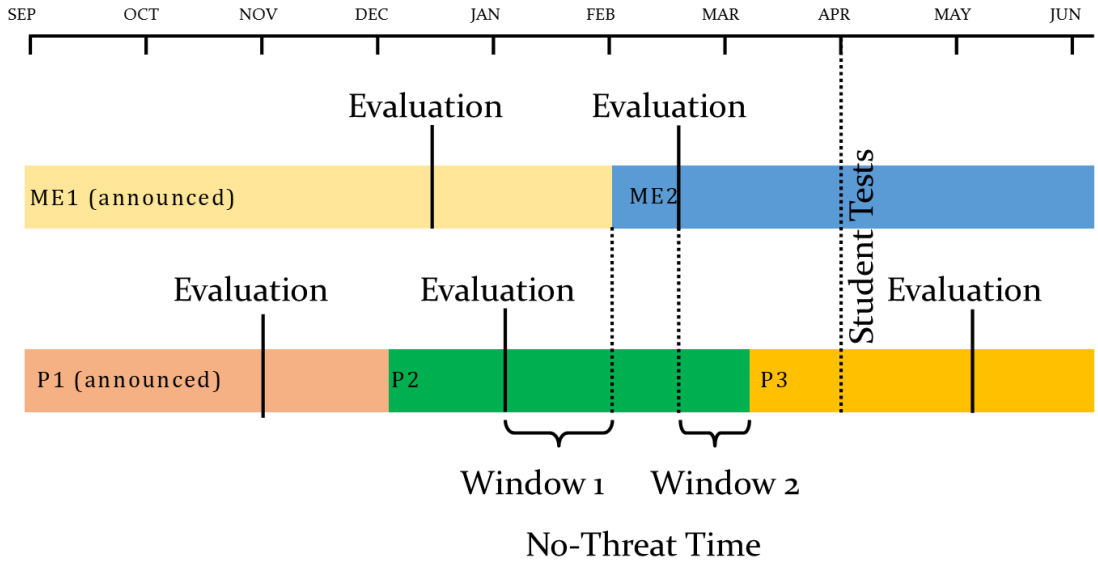
Notes - One evaluation must occur within each window. For announced evaluations, teachers are informed no later than the day before their evaluation

Figure 1.2: Simplified example of calculating no-threat time.



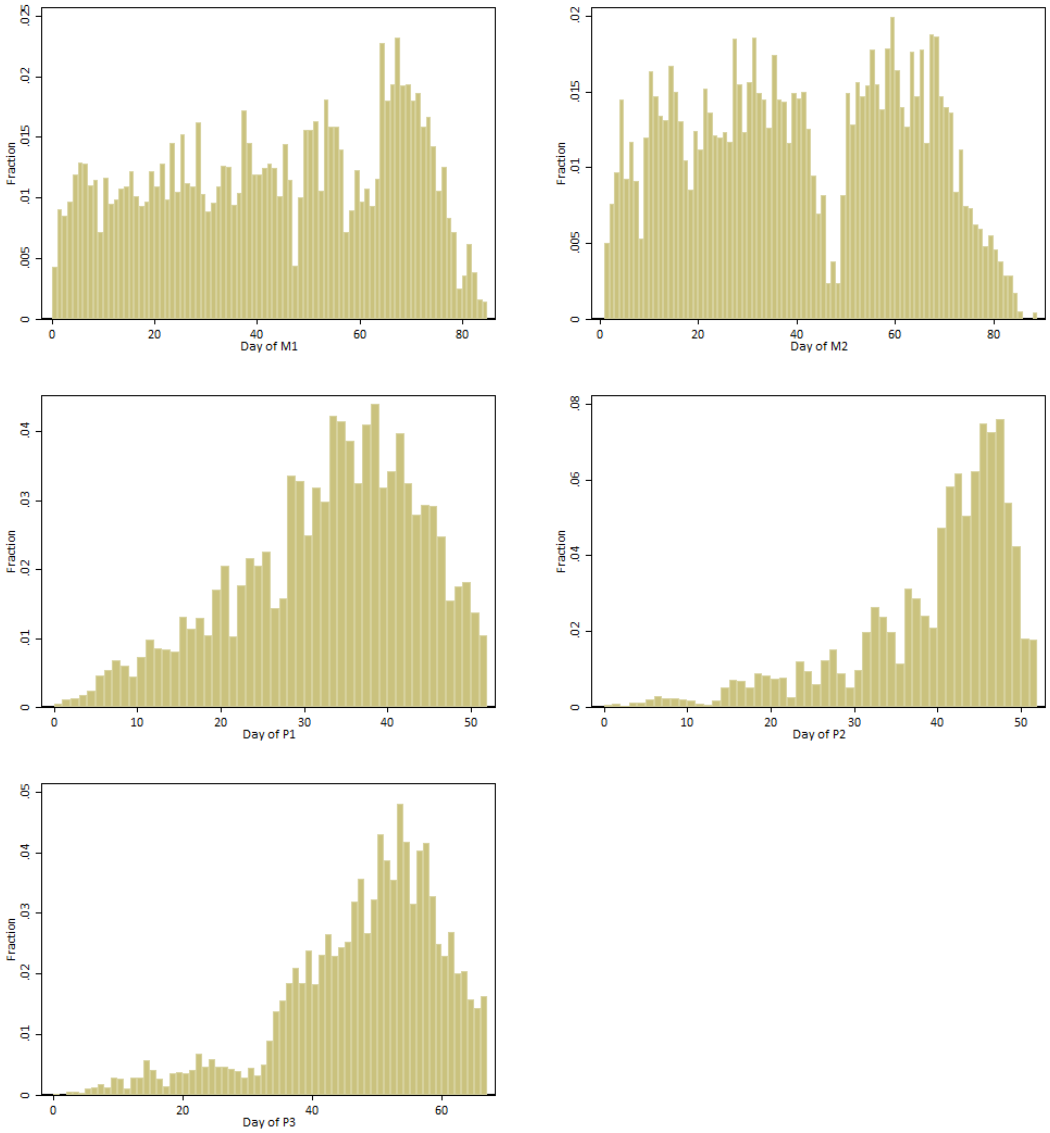
Notes - If teachers did not have overlapping windows, calculating no-threat time would be straightforward: it would be the time after the first evaluation until the start of the next window, and the time after the second evaluation until student tests. Note that no-threat days after student tests are not counted.

Figure 1.3: Calculating no-threat time in DCPS.



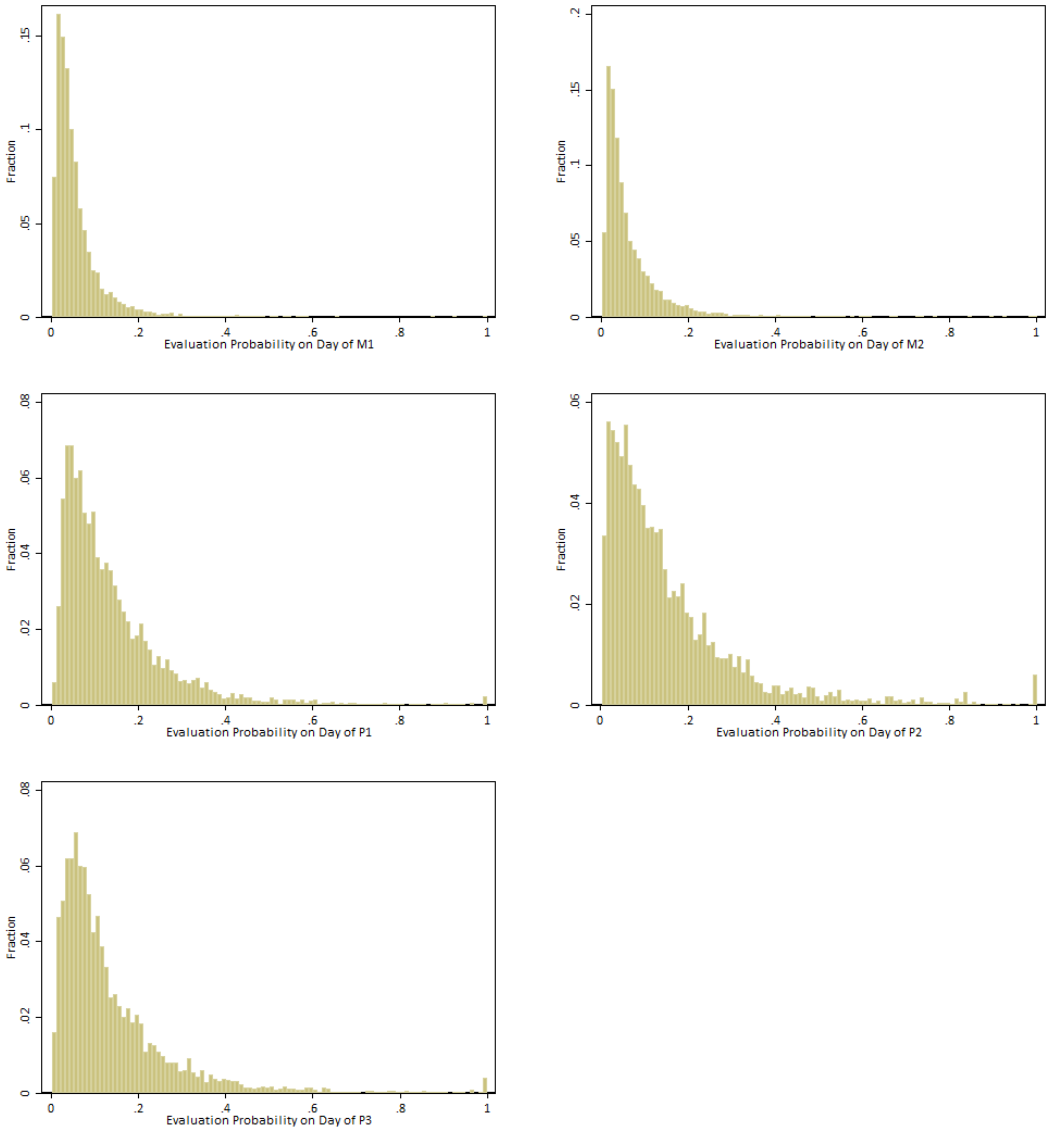
Notes - Because evaluation windows in DCPS overlap, no-threat time is defined as days in which there is no threat of an unannounced evaluation from either evaluator. Window 1 starts December 1 and ends February 1, and Window 2 starts February 1 and ends March 15. No-threat time is not calculated for *P*3 because very few teachers receive their evaluation before student tests, and there are selection concerns.

Figure 1.4: Timing of evaluations within evaluation window.



Notes - Days are measured as instruction days, which excludes in-service days, weekends, and holidays. Master educator evaluations, $M1$ and $M2$, are distributed uniformly across the window. Principal evaluations – $P1$, $P2$ and $P3$ – are often clustered near the end of each window.

Figure 1.5: Histograms of the cumulative probability of an evaluation.



Notes - For principal evaluations ($P1$, $P2$, $P3$), the probability is usually higher because these evaluations occur late in their assigned window.

Chapter 2

Teacher Improvements in Windows of High-stakes Observation

Policy makers are turning to multiple measure teacher evaluation systems in an effort to improve teacher practice. Yet, little evidence establishes the causal link between the evaluation program and daily teacher responses. Teachers may respond to such programs in ways not intended by the policy, such as over-emphasizing unimportant elements on a teaching evaluation rubric at the expense of more holistic teaching improvements, or using a perfectly prepared lesson-in-a-box reserved for when an evaluator arrives. Using administrative data from the Washington DC teacher incentive program, called IMPACT, we identify specific teacher behavioral responses to a pending unannounced in-class evaluation. Random variation in the probability of being observed identifies how teachers prepare for an evaluation as it becomes more likely. Our key finding is that teachers respond to possible in-class evaluations as predicted by a multi-tasking principal-agent model, making incremental improvements across a variety of teaching practices but most notably within areas that are easiest to improve. This rules out the possibility that the average teacher circumvents the intent of evaluations with practices like a "lesson-in-a-box."

2.1 Introduction

High quality teachers substantially improve student academic and life outcomes (Aaronson et al., 2007; Rivkin et al., 2005; Rockoff, 2004; Chetty et al., 2014). In an attempt to improve teacher quality, policy makers in the last decade are increasingly using multiple measure teacher evaluation systems, which are comprehensive in-class evaluations that measure a variety of teaching practices. Seminal work highlights teacher evaluation as a major policy lever for affecting end-of-year student and teaching outcomes (Dee and Wyckoff, 2015; Taylor and Tyler, 2012; Steinberg and Donaldson, 2016). Many of these new programs rely heavily on standards-based classroom observations because of their potential to bridge the gap between teacher training and practice through the provision of rich performance information and cohesive instructional standards (Taylor and Tyler, 2012; Papay, 2012; Cohen and Goldhaber, 2016). Yet, there exists little evidence demonstrating how evaluation systems improve teacher practice across instructional domains and throughout the year to achieve these effects. Teachers may respond to such programs in ways not intended by the policy, such as over-emphasizing unimportant elements on a teaching evaluation rubric at the expense of more holistic teaching improvements, or using a perfect prepared lesson reserved for when an evaluator arrives (a "lesson-in-a-box"). Our analysis unpacks the specific routes through which the evaluation program improves teacher performance on evaluations and student outcomes. Using administrative data from the Washington DC teacher incentive program, called IMPACT, we find that the average teacher improves her practice along multiple measured dimensions in response to a pending observation, as intended. Our results uniquely demonstrate that classroom observations in high-stakes systems encourage effective teaching through the uptake of standards-based instructional practices.

A few studies demonstrate that teachers change their practice in response to evaluation systems, but these studies cannot determine the mechanism by which these improvements occur. For example, using the same administrative data used in our analysis, Dee and Wyckoff (2015) show that low-performing teachers subject to dismissal threat in IMPACT improve their evaluation scores. However, their approach does not unpack the nature of these improvements throughout the year, nor does it teach us about the behavior of teachers whose ability put them far

from critical incentive thresholds. Similarly, Taylor and Tyler (2012) find improved student achievement as a result of a low-stakes observation system for mid-career teachers in Cincinnati, OH, and Steinberg and Sartain (2015) find small student achievement gains in a similar system in Chicago, IL. The authors hypothesize that particular elements of these policies may have supported teacher development, like peer evaluators or structured post-observation conferencing. However, the research design allows just for the conclusion that teacher evaluation can improve outcomes for teachers and students. Our work builds on this finding by revealing the mechanisms by which these results occur in Washington, DC.

Our empirical approach identifies how a teacher modifies her practice as an unannounced evaluation becomes more likely. Teachers in DC Public Schools experience multiple classroom observations per year, which take place during predetermined windows of time. In each evaluation, teachers are scored along nine different teaching practices, ranging from measures of classroom management to student engagement. We observe teachers' scores along all nine components for each evaluation, which we use to identify improvements in teacher practice as a response to the probability of an evaluation. Due to the structure of IMPACT, there is random variation in the daily probability of an evaluation within each window, allowing us to causally identify how the increased likelihood of an in-class observation affects a teacher's score. We show that teachers score higher when an evaluation is more likely, providing unique evidence that they are cognizant of observation standards and use that knowledge to enhance their practice in a high-stakes program. These responses substantively improve student outcomes (Phipps, 2017).

There is a substantial body of literature about the effective design of a teacher evaluation system, ranging from the type of language to use in defining evaluation components and how they are scored (the rubric) to whether or not principals are ideal evaluators. Our findings add needed empirical evidence to this literature by showing that the IMPACT evaluation rubric guides teacher behavior. The causal link between evaluation probability and teacher score establishes that teachers are aware of the practices required to improve their evaluation score, using that knowledge to enact the rubric and improve their evaluation score. A secondary finding in our analysis is that evaluations conducted by principals appear less responsive to evaluation probability than those conducted by outside observers. This adds to the empirical literature documenting the differences between principals and outside observers. Our results are consistent with

the literature, which shows principal evaluations have a more compressed distribution of teacher evaluation, particularly in high-stakes contexts (Kraft and Gilmour, 2016b,a).

Our results uniquely demonstrate that teachers do not appear to circumvent the purpose of the evaluation rubric. One way a teacher could do so would be to use a lesson-in-a-box. To assess this possibility, we compare teacher responses to announced evaluations with their responses to unannounced evaluations. When teachers are warned at least a day in advance of their evaluation, we find that increased evaluation probability has no effect on their score, which is not the case for unannounced evaluations. We would expect the results to be the same if teachers were able to prepare a lesson-in-a-box. It is also possible that a high-stakes evaluation system could lead teachers to over-emphasize easily manipulatable rubric components at the expense of more holistic teaching improvement, which we rule out by demonstrating substantive effects across many teaching domains.

This paper provides unique quantitative evidence that multiple measure teacher evaluation systems encourage instructional best practices, empirically validating the theoretical link between information-rich observation rubrics and improved teacher practice. In this paper, we review the DC context, provide an economic framework for conceptualizing teacher responses, and review the existing literature about classroom observation in evaluation systems. Then, we discuss our findings and the relevant policy questions that persist.

2.2 Study Setting

As part of the growing demand for increased school and teacher accountability, Washington DC Public Schools (DCPS) implemented a high-stakes teacher evaluation program called IMPACT starting in the 2009-2010 school year. All teachers in DCPS have large financial incentives that depend on a weighted combination of elements, which mirror many multiple measure teacher evaluation systems of this decade. For most teachers, the largest component of their IMPACT score comes from scored classroom observations based on the district's Teaching and Learning Framework (TLF). The TLF is intended to define criteria that establish effective teaching, derived largely from the Danielson framework. The TLF addresses various domains, such as maximizing instructional time and checking for student understanding. A teacher's final TLF score is the average of five evaluations conducted throughout the year. Where possible, IMPACT scores also in-

clude a student test-based value-added score for tested subjects (Math and ELA) in grades four through eight, which is only 18% of teachers in DCPS. For these teachers, value-added scores make up 50 percent of the final IMPACT score, while in-class evaluations comprise only 35 percent of their final IMPACT score. For the remaining majority of teachers, classroom observation scores comprise 75 percent of their final IMPACT score. The remaining components of a teacher's final IMPACT are small and include an overall school measure of student test score growth, a principal-assessed score of commitment to the school and community, and a teacher's success in reaching instructional goals for grades and subjects ineligible for value-added measures (for more details on the IMPACT program structure, see Dee and Wyckoff, 2015).

The high-stakes nature of this system makes DC a unique context, and previous analyses have comprehensively attended to these program elements. Dee and Wyckoff (2015) use the discontinuities in the IMPACT program's reward structure to show teachers facing dismissal threat significantly improve student achievement gains and in-class evaluation scores. Adnot (2016) uses latent profile analysis to identify components of the in-class evaluation rubric most sensitive to adjustment when teachers face a dismissal threat. Her results suggest that low-performing teachers adapt to highly specific language in the rubric to improve, while high-performing teachers demonstrate more variety in their uptake of rubric practices. Adnot et al. (2017) use a quasi-experimental differences-in-differences approach to find statistically significant student achievement increases of 0.14 standard deviations in reading and 0.21 standard deviations in math as a result of low-performing teacher exits under the IMPACT evaluation policy.

A few other policy components make DCPS an interesting context for policy analysis. In DCPS, the overall in-class evaluation score is the average of five in-class evaluations conducted by principals and district employees called "master educators," each of which is weighted equally. In the first year of the program, the first principal and master educator evaluations are announced at least a day in advance, though this was changed in subsequent years so that only the first principal evaluation was announced. The remaining three observations are conducted without notice within a pre-defined time period, or observation window. Each evaluation lasts roughly 30 minutes, and teachers are given a score from one to four on each of nine equally weighted components. At the beginning of each year, the rubric guidebook is published publicly for teacher review.

A complementary analysis uses a similar identification strategy as ours to disentangle the relative effects of the high-powered incentives and the feedback provided on observation rubrics from principals and master educators. Phipps (2017) shows that, by the structure of the IMPACT program, some teachers randomly experience days in which they are guaranteed not to have an evaluation, which the author uses to identify the effect of teacher responses to a potential unannounced observation on student test outcomes. He finds that the possibility of an evaluation has substantive effects on student test outcomes in both reading and math. The author's analysis is also able to separately identify the effects of evaluation feedback and demonstrates that teachers who receive feedback earlier have better student outcomes. This paper adds to the literature on the IMPACT system by documenting the behavioral adjustments teachers make in preparation for an evaluation.

2.3 Literature Review

Our analysis touches on a variety of classroom evaluation system design questions. In showing teachers enact the evaluation rubric as designed, our analysis raises the question as to whether or not these behaviors improve student outcomes. A related concern is, which of the elements measured during in-class evaluations are most important, and how does the language and design of the rubric effect teacher responses? In our results, we also identify significant differences between the two types of evaluators, principals and master educators. Whether or not principals are good candidates for conducting in-class observations remains an open question. Recent research addresses these questions in a variety of contexts, which we review here.

2.3.1 Standards-Based Observations

Most evaluation systems in the US use standards-based observation as a primary measure largely due to the growing evidence that these standards are linked to meaningful student achievement gains (Steinberg and Donaldson, 2016). While several high-profile teacher performance incentive programs use value-added measures, a survey of such programs in the US implies that without in-class observations, these programs do not improve student outcomes (Phipps, 2017). Teacher improvements on an observation rubric correlate with student

achievement, both in this context as well as others. Phipps (2017) shows that teacher responses during observation windows in the IMPACT program causes higher student achievement, suggesting that enacting the practices contained in the Teaching and Learning Framework (TLF) causally improves teaching quality. Standards based observation rubrics in other contexts correlate with student outcomes as well, namely the CLASS framework (Hamre and Pianta, 2001), the Framework for Effective Teaching (Kane et al., 2013), and other content-specific rubrics used in the Measures of Effective Teaching (MET) study (Kane and Staiger, 2012).

Given the variety of possible observation rubrics, scholars continue to engage in a conceptual discussion regarding the content of observation rubrics. Yet only a few empirical studies demonstrate the ways in which this content matters. Two studies in particular tackle the notion that rubric specificity can influence teacher practice. Adnot (2016) found that low performing teachers facing a dismissal threat in DCPS improved the most on highly specific rubric practices, ranging from a statistically significant effect size of 0.22 to 0.62 standard deviations. Another study found that "concrete suggestions for improvement" in the rubric language correlated with teacher improvement on that rubric (Kane et al., 2010). Our analysis enhances this literature by showing how teachers with varying degrees of skill employ the behaviors specified in the TLF evaluation rubric. We find that teachers consistently improve in areas of the rubric with more concrete language as well as other teaching domains.

Other studies highlight how rubric language may play a part in teacher improvement, but do not establish a causal link between rubric practices and teacher development. In Cincinnati, OH, teachers improved student achievement in years after receiving a detailed evaluation (Taylor and Tyler, 2012). Yet the authors do not address the specific mechanism that led to these continued student achievement gains. They hypothesize that the improvements were driven, at least in part, by the enhanced reflective discussions among teachers around a core set of practices introduced or solidified by the rubric, but it is not clear what role the rubric design played. In New Haven, CT, researchers used surveys to show that the process of designing the evaluation system had an effect related to teachers' information and awareness about the assessed practices (Donaldson and Papay, 2015). Our results provide empirical evidence that causally links teacher responses to particular evaluation rubric elements. Our results reinforce the notion that rubric

language matters, as teachers are clearly cognizant of what the evaluation rubric covers.

The effects of rubric language may also vary based on a teacher's skill and experience or the grade and subject she teaches. Hill and Grossman (2013) argues the evaluation rubric should be tailored based on grade and subject in order to support teacher improvement. The alignment (or lack thereof) between the observation rubric and the practices necessary for student achievement in a particular subject or grade may create tension for teachers regarding how to allocate time and resources (Cohen and Grossman, 2016; Hill and Grossman, 2013). Highly-skilled teachers may more flexibly alternate between particular practices as needed. We address part of this issue by examining how teachers of different skill and experience levels respond to the evaluation rubric, showing that higher skill teachers appear more adaptable to the rubric. Future empirical work can employ our analysis technique to examine the importance of designing evaluation rubrics around specific subjects and grades.

2.3.2 Evaluation Observers

While teachers are effective at improving their evaluation score as an observation becomes more likely, we find they are more effective at improving their score when evaluated by an ME over a principal. We cannot provide additional evidence indicating what causes this difference, but determining whether or not principals are ideal observers is an open question relevant to all evaluation systems. Despite robust observation protocols with typically four or five performance categories, principals still primarily rate teachers as effective (Kraft and Gilmour, 2016a; Grissom and Loeb, 2017). There is evidence that principals tend to inflate the distribution of teacher performance; that is, when the same lesson is scored by an external evaluator, the principal rating tends to be higher (Sartain et al., 2011). Compared to external or peer evaluators in the MET project, administrators rated teachers 0.1 points higher on a four-point Framework for Effective Teaching scale, a substantial difference given a compressed distribution across the board (Ho and Kane, 2013). Even at the practice level, very few teachers receive the second to lowest or lowest ratings. For example, just about 2.9 percent of teachers in Miami-Dade, FL, received these lowest category ratings on any instructional standard (Grissom and Loeb, 2017).

There are a variety of reasons why principal ratings may produce less variety than expected. Principals may have difficulty distinguishing between more

granular practices, or lack the instructional expertise necessary to observe certain skills (Kraft and Gilmour, 2016b). An exploratory factor analysis of both high and low stakes evaluation ratings on similar rubrics found that principals essentially rate teachers based on a singular perception of performance, which may be either by design, as teacher practices are highly correlated with one another, or because principals do not have the necessary instructional expertise to tease out performance by individual dimensions (Grissom and Loeb, 2017). In New York, more experienced teachers gave observation scores that correlated with student value-added more than observation scores given by novice principals, suggesting that this kind of evaluation expertise develops over time (Rockoff et al., 2012). Survey evidence shows that time constraints due to the additional work associated with assigning a low rating, motivating low-rated teachers, balancing roles as feedback provider and rater, and personal discomfort with assigning low categories led principals to artificially rate fewer teachers in the lowest category (Kraft and Gilmour, 2016b).

High stakes settings in which dismissal is tied to ratings exacerbate some of these tensions for principals. Principals are keenly aware of labor market implications; administrator interviews reflected a sentiment about who would replace the low performing teacher should they be "evaluated out" (Kraft and Gilmour, 2016b). In a recent study, low stakes principal evaluations were compared to high stakes official evaluation scores, and the proportion of teachers rated in the lowest two categories substantially increased in the low stakes category (Grissom and Loeb, 2017).[1] Principal perceptions of observation rubrics and the way they are used are certainly a factor worth considering when designing teacher evaluation systems. Our results contribute to the body of literature by highlighting differences in the malleability of evaluations conducted by principals and MEs in a high-stakes setting.

[1]Interestingly, the observation standards most correlated with value-added differed by high and low stakes contexts. In the high stakes context, Instructional Delivery and Engagement was most highly correlated with value-added measures, while a researcher-developed construct called Improving Critical Thinking was more associated with value-added measures in the low stakes context (Grissom and Loeb, 2017). Strangely, the low stakes rating was more consistently associated with value-added measures in math but not in reading, where with the exception of high school ELA, the high stakes ratings were more strongly associated with ELA value-added (Grissom and Loeb, 2017).

2.4 Model of Teacher Responses to Evaluations

In this study, we examine the extent to which multiple measure teacher evaluation systems result in teachers shifting their practice, particularly in response to classroom observation. Underlying this analysis is the assumption that teachers will respond to evaluation policy. Our empirical result relies on the belief that teachers will prepare more as an unannounced classroom evaluation become more likely. In what follows, we present a economic model of behavior to describe how teachers in a high-stakes environment would adjust their planning to fit the observation rubric.

Given the many measures of teacher quality in a program like IMPACT, teachers must choose among a variety of potential teaching styles and lesson structures in order to improve their final rating. The multi-tasking model originating from Holmstrom and Milgrom (1991) describes the process of allocating time and effort towards a variety of tasks, each of which is rewarded.[2] Let teachers have a set, call it K, of possible tasks or practices on which they can focus their time and attention, both in lesson reparation and in the classroom. A teacher will plan her lesson given the needs of her students, the standards on which they are assessed, and the curriculum and resources available to her. She then enacts that lesson using a breadth of instructional knowledge and skills, adjusting her plan based on student mastery of the material. Then let x_i be a teacher's allocation of time towards task i, and let $x = [x_1, x_2, \ldots, x_n]^T$ be a vector of all time allocation across tasks, where n is the size of the set K. Then x has some utility cost $c(x)$, which has increasing marginal costs for each input. The teacher receives wage $w(x)$ as a result of her score, which is determined solely by x given no measurement noise. Then with a standard exponential utility function with coefficient of risk aversion r, a teacher's utility is

$$U(x) = -\exp\{-r(w(x) - c(x))\}. \tag{2.1}$$

To maximize utility, the first-order conditions require that a teacher chooses x such that the net marginal benefit of each input is zero.

If the bonus system only rewards certain behaviors, say $K' \subset K$, then the marginal benefit of those tasks increases, leading to an increase in how much

[2] Holmstrom and Milgrom (1991) include measurement noise on each individual component. While it is likely that each TLF component is measured with noise, we ignore measurement noise to facilitate simplicity. The qualitative expected responses are not different.

time and effort a teacher allocates to those tasks. That is, x_i for $i \in K'$ will increase. If the elements in x are cost substitutes, there will also be a decrease in x_i for $i \notin K'$. That is, a teacher will favor elements on the rubric that are easy to adjust over elements that are difficult to adjust or do not earn rewards in the evaluation rubric, a key result of the Holmstrom-Milgrom model. This reflects the limited time available to teachers, where spending time on preparing one aspect of a lesson or on a specific approach in class naturally requires reducing time spent on another approach.

Because evaluations determine a teacher's potential bonus, a teacher will modify her choice of teaching practices, x, based on how likely she thinks an evaluation is. When classroom observations have to be conducted once within a pre-specified time frame, she can estimate the probability of being evaluated. Intuitively, as the end of the evaluation window approaches, it becomes more likely each day that a teacher who has not already had their observation will be evaluated.

To incorporate the probability of evaluation into the utility function, allow a teacher's utility at the end of the day to be different depending on whether she was evaluated or not. If she is not evaluated on a given day, her effort exerted on activities that only improved her in-class observation score will not contribute to her utility. In other words, if there is no in-class observation that day, effort towards improving her observation score are wasted. For days in which she is not evaluated, her utility is

$$U(x, m = 0) = -\exp\{-r(-c(x))\}, \tag{2.2}$$

where $m = 0$ indicates she was not evaluated. But if she is evaluated, her utility is as before. To combine these two possible outcomes each day, let p be the probability of an evaluation on the next day. Then in the evening, as a teacher prepares for her next day, her expected utility is

$$EU(x) = pU(x, m = 1) + (1 - p)U(x, m = 0). \tag{2.3}$$

Equation 2.3 shows the intuition that as the probability of evaluation increases, there are larger marginal returns to using evaluated practices, x_i for $i \in K'$. As a result, her use of evaluated practices should increase with the probability of an evaluation, leading to an increase in her evaluation score. This result drives our

empirical approach: as p increases, teachers will shift their preparation and time towards practices that will improve their evaluation score.

In theory, this incentive should lead to improved teacher practice, but there are ways in which teachers may respond contrary to the policy's intent. If there are a few elements of the rubric that emphasize easily observable practices like updating the daily objective board each morning, teachers may use valuable time completing this task rather than engaging in other more pedagogically important classroom preparation. Our model shows that a teacher will emphasize any practices that have low costs and still earn higher evaluation scores. If a teacher can earn a sufficiently high evaluation score by simply adjusting a single TLF component then the incentive has not achieved its purpose. We can observe if this is the case by considering which of the nine TLF components a teacher adjusts most. If she only adjusts one in response to a likely evaluation, the evaluation rubric may not adequately measure and encourage holistic teacher practice. More cynically, it is also possible that teachers will circumvent the purpose of evaluations altogether by preparing a perfect lesson-in-a-box, an issue we also address.

2.5 Data and Econometric Approach

Using administrative data from DCPS in the 2009-2010, 2010-2011, and 2011-2012 school years, we use the random timing of unannounced evaluations to identify how teachers modify their behavior in preparation for an evaluation. This is done by estimating the probability of being evaluated by either a principal or master educator on any day in each school. We can then estimate how teachers modify their behavior in preparation for an evaluation.

2.5.1 Description and Calculation of Treatment Measure

The probability of an evaluation is partly determined by how far into an evaluation window a teacher gets without receiving her evaluation (see also Phipps, 2017). Figure 2.1 illustrates the window of each of the five evaluations. The first principal evaluation occurs between mid-September and December 1st, the second between December 1st and March 1st, and the third between March 1st and the end of classes. The first master educator evaluation occurs between mid-September and February 1st, and the second is conducted before the end of classes.

We observe the date of each observation for each teacher, which we use to calculate how likely the remaining teachers are to be evaluated in each of the remaining days. Intuitively, a teacher knows that if tomorrow is the last day of an evaluation window and she has not been evaluated, she will be evaluated tomorrow. Two factors determine a teacher's estimate of the probability of being evaluated on any particular day: the number of teachers that remain to be evaluated and how many evaluations a teacher expects to be conducted. It is then straightforward to calculate evaluation probability if each remaining teacher has an equal probability.

To be more formal, let k be an evaluation indicator, where k is $P1$, $P2$, or $P3$ for the principal evaluations and $M1$ or $M2$ for master educator evaluations. Then let a teacher's estimate of the number of evaluations to be conducted on day t at school s be $\hat{N}^k_{ts}$. If R^k_{ts} is the number of teachers who still need evaluation k on day t at school s, then each remaining teacher's probability of being evaluated is

$$p^k_{ts} = \frac{\hat{N}^k_{ts}}{R^k_{ts}}. \tag{2.4}$$

We can determine how many teachers remain to be evaluated, R^k_{ts}, but estimating how many evaluations a teacher expects to be conducted, $\hat{N}^k_{ts}$, requires assumptions about a teacher's knowledge of when evaluators will conduct more evaluations. If a teacher knew exactly how many evaluations would be conducted on every day, then $\hat{N}^k_{ts} = N^k_{ts}$. This is a strong assumption that is unlikely to be true, especially if evaluations are not evenly distributed within a window.

Principals tend to bunch their evaluations near the last third of the time window, which means the expected number of evaluations changes over time. In the beginning of an observation window, teachers expect that principals will conduct few evaluations, but towards the end of the window, teachers expect more each day. On the other hand, master educators distribute their evaluations more evenly, so the expected number of evaluations remains constant. Figure 2.2 shows the overall distribution of evaluations across each window. While the master educators maintain a fairly uniform distribution, principals are very often conducting evaluations in the last third of the available time. The dip in evaluations in M2 around day 45 is a result of student testing days in April.

Instead of assuming teachers know exactly how many evaluations will be conducted on each day, we assume they are broadly aware of the trend. That is, we allow for a teacher to know that the number of evaluations conducted

by a principal will increase towards the end of the window. We also allow for a teacher to notice an increase in evaluations over the past few days. We can approximate this information by estimating the distribution of evaluations with a kernel density. The kernel smoothing approximates changes in the trend of daily evaluations that we expect teachers notice. Our results are not sensitive to this assumption.

For many days in the year, a teacher has the possibility of either a principal evaluation or a master educator evaluation (or both). The two events are independent and in rare cases both occur on the same day for a single teacher. To determine the probability of any evaluation, I use the sum of their individual probabilities. For example, if a teacher has not yet had either $P1$ or $M1$ evaluations, her probability of *any* evaluation the next day is $p_{ts} = p_{ts}^{P1} + p_{ts}^{M1}$, but if she had already received her $P1$ evaluation, her probability is just $p_{ts} = p_{ts}^{M1}$.[3] Then in our specification, we use p_{ts}.

Given an estimate of the probability of any evaluation for each day in each school, we know the probability of an evaluation on the day in which a teacher was, in fact, evaluated. We use P^k in capital letters without the subscript t, to indicate the probability of any evaluation on the day when evaluation k occurred. For a teacher receiving evaluation $P1$ on day t at school s, the treatment variable is $P^{P1} = p_{ts}^{P1} + p_{ts}^{M1}$, assuming she has not yet received her $M1$ evaluation.

2.5.2 Data Summary

Over the three years studied, DCPS has between 121 and 124 elementary, middle and high schools, with roughly 3,500 teachers each year. Table 2.1 summarizes school characteristics over the study period to show there were no meaningful changes to school or student composition. The average classroom size was constant at about 17.6 students, and the fraction of schools classified as high-poverty schools ranged between 0.75 and 0.77.

Table 2.2 provides detailed information on how teacher evaluation scores are distributed over evaluations and years. Because scores are bounded below at 1 and above at 4, we have included percentile measures at the 10th percentile and the 90th percentile instead of the minimum and maximum. Master educator scores are lower on average, but the difference is not statistically significant. With a median score around 3.75, these are not normally distributed. The cut-

[3] These additive probabilities are capped at 1, though the cap was rarely needed (it applied to 0.38 percent of all observations).

off for a Minimally Effective rating is 2.5, and slightly more than 10 percent of teachers receive that rating. The cutoff for Highly Effective is 3.5, and about 65 percent of teachers have overall evaluation ratings above this threshold. Most of the variation, however, occurs just at the Highly Effective threshold.

To understand the effect sizes, we have included a summary of each treatment variable, evaluation probabilities, for each year in Table 2.3. The average evaluation probability for principals is larger than for master educators, as expected given how clumped principal evaluations are at the end of each window. We include the minimum probability and the 90th percentile, since probabilities are bounded above at 1. The median treatment for principal evaluations is around 10 percent, meaning the median teacher had a 10 percent chance of being evaluated on the day of their principal evaluation.

There are some exogenous factors that will affect teacher observation scores that may have to do with school events. For example, the timing of an evaluation within the school year may also matter regardless of evaluation probability. It is possible that teachers are able to develop stronger classroom cultures as time progresses, for example, which could improve their evaluation scores. In our analysis, we control for seasonal changes in teacher observation scores by including the timing of an evaluation within each window.

Teachers also perform differently on evaluations after their students have completed standardized tests, which is consistent with other research on DCPS and suggests some teachers change their focus after tests are completed. The only evaluations affected by this post-test change are the second master educator and third principal evaluations. We control for this systematic response by including an indicator for whether the evaluation occurred before or after standardized testing for these estimates for evaluations $M2$ and $P3$.

2.5.3 Econometric Specification

The outcome variable of interest is the standardized evaluation score, Y_{ij}^{k}, for teacher i in year j on evaluation k, where k is $P1$, $P2$, or $P3$ for principal evaluations and $M1$ or $M2$ for master educator evaluations. Evaluation scores are standardized within year. We control for school-level characteristics using school fixed effects ϕ_s.

One complication in estimating treatment effects on teachers is how to appropriately control for experience (see Taylor and Tyler, 2012, for example). It is common in the literature to use a quadratic form, with experience capped at 15 or

20 years, or to use experience-level fixed-effects for each year of experience. Our results are unaffected by the specification of experience, but given the richness of our data, we have opted to use experience level fixed-effects. We use X_{it} to be a vector of experience level indicators. A set of controls N_{ij}^k is unique to each evaluation k, and includes an indicator for whether the evaluation occurs after standardized testing (for $M2$ and $P3$). We control for the timing of an evaluation within the evaluation window k with T^k. Let P^k be the probability of any evaluation on the day of evaluation k. We then estimate a teacher's evaluation score with:

$$Y_{ijs}^k = \beta_0 + \phi_s + \beta_{P^k} P_{ij}^k + \beta_X X_{ij} + \beta_{T^k} T^k + \beta_N N_{ij} + \varepsilon_{ij}. \tag{2.5}$$

The errors are clustered at both the school by year level and by teacher. Because our outcome variable is standardized at the year level, the measured effects of P^k are in units of standard deviations. Crucially, we assume that the errors are conditionally independent of any unobservable characteristics that correlate with a teacher's evaluation score.

The term N_{ij}^k also contains controls for classroom-specific characteristics and a teacher's ability by using the first principal evaluation as a control for the subsequent evaluations. Because $P1$ is announced, it is not affected by timing in the way subsequent evaluations are. This evaluation also represents a baseline measure for a teacher's ability under the best evaluation circumstances. For example, in estimating the effect of evaluation probability P^{P2} on the second principal evaluation score Y^{P2}, we use the scores from the first principal evaluation, Y^{P1}, as a control. In addition to looking at the effect of evaluation probability on unannounced evaluations, we also estimate the effect of evaluation probability for announced evaluations. In this case, we use the first master educator evaluation as a control.[4]

Every specification includes a control T_{ij}^k for the evaluation timing. Because the probability of being evaluated depends on how many other teachers have been evaluated before, and by how much time remains in the evaluation window, we want to separate out any effects caused by simply having an evaluation later in the year. If there was only a single possible evaluation without overlapping windows, evaluation probability would be defined by evaluation timing, though nonlinearly. But given the overlapping windows and the non-uniform distribution of evaluations, evaluation timing and evaluation probability are separately identified.

[4]In the 2009-2010 school-year, we also add the first master educator evaluation since it was announced.

2.5.4 Treatment Exogeneity

Our key identifying assumption is that the timing of evaluations is independent of teacher characteristics that would affect their evaluation score, conditional on observable characteristics. Principals may want to conduct evaluations of their worst teachers first in order to provide feedback earlier in the year, which would bias our results upwards since evaluation probability is low in the beginning of each window. If evaluators target weaker teachers early using information we cannot observe, our identification assumption is invalid.

To test for treatment selection bias, we regress characteristics that are observed by principals and master educators on our treatment variables, P^k. We estimate the following regression on the probability of each evaluation for each observable characteristic X_{ij} for teacher i in year j at school s:

$$P^k_{ijs} = \beta_0 + \phi_{sj} + \beta X_{ijs} + \varepsilon_{ij}\,. \tag{2.6}$$

Because there may be school-by-year systematic differences in the timing of evaluations, ϕ_{sj} is a school-by-year fixed effect for school s in year j. The observable characteristics we consider are teacher value-added scores in reading and math in the previous year, an indicator for first-year teachers, the final IMPACT rating a teacher received the previous year, and the final evaluation rating a teacher received in the previous year. Because teacher selection into schools is not random, we mean-center each observable characteristic around the school-by-year mean. That is, $X_{ijs} = x_{ijs} - \bar{x}_{js}$, where x_{ijs} is the raw, uncentered value for teacher i and $\bar{x}_{js}$ is the average at school s for year j.[5]

Table 2.4 shows the results of the exogeneity checks specified by Equation 2.6, with the treatment variables (P^k_{ijs}) across the columns and the observable characteristics evaluators may use to target teachers (X_{ij}) in each row. The cells are the coefficient β in Equation 2.6. The statistical significance shown has not been adjusted for multiple hypothesis testing.

The most important conclusion from Table 2.4 is that any potentially significant characteristics have effects in a direction that would bias our results downward. For example, if principals target their second evaluation towards teachers with a Highly Effective IMPACT rating from the previous year, then Highly Effective teachers will have lower evaluation probability, reducing any positive

[5] The results of our exogeneity checks are effectively unchanged when mean-centering is not used on the independent variables.

effect we observe from evaluation probability. No evaluations are timed relative to whether a teacher is under the threat of dismissal or not, nor are they correlated with a teacher's evaluation score being "Minimally Effective" or "Highly Effective" in the previous year.

Our exogeneity tests cannot be exhaustive since we are concerned with characteristics observed by evaluators but not observed in the data. However, our checks support our identifying assumption by showing that on a variety of observable characteristics we know correlate with teacher quality, evaluators are not systematically targeting weak teachers early in the window.

2.6 Results

Our main results are presented in Table 2.5 for announced evaluations and Table 2.6 for unannounced evaluations. For each evaluation, we have a simple overall treatment regression with the effect of evaluation probability (P^k) on the score for that evaluation measured in standard deviations. The coefficients on probability represent the effect of an increase in probability by one full unit. For example, the interpretation of the coefficient in column four row one is that a teacher who is certain she will be evaluated improves her score by 0.28 standard deviations on $P2$ over a teacher who does not expect to be evaluated.[6] Because the median evaluation probability for $P2$ is about 0.10, the effect size is about a 0.03 standard deviation increase for a teacher with the median probability of evaluation over a teacher with zero probability.

To see how teachers with different previous evaluation ratings respond differently to evaluation probability, we have broken out the effects by first-year status and then by the rating received on the previous year's in-class evaluations. For the first master educator evaluation in the 2009-2010 school-year, we do not have a prior year rating for teachers, so it is not possible to estimate the effect separately by prior-year rating. Veteran teachers are consistently able to improve their evaluations in response to increased evaluation probability. Among veteran teachers, one with a 10 percent greater chance of an evaluation will improve her score by 0.02 to 0.06 standard deviations, depending on the evaluator, as shown in row one columns 2, 5, 8, and 11.

[6] While we have assumed linear effects, our results are not qualitatively different when using a log specification.

Rows three and four in both Table 2.5 and Table 2.6 show the average effect of having been rated a minimally effective teacher in the previous year or a highly effective teacher in the previous year. As expected, teachers that were rated highly effective in the previous year have higher evaluation ratings this year when compared with a teacher rated effective in the previous year. The result is robust for both announced and unannounced evaluations.

For the announced evaluations, probability is never a significant predictor of evaluation score. The average effect of having a minimally effective rating last year is significant across all evaluations. The average effect of being a low-rated teacher is significantly different between announced and unannounced evaluations, which is consistent with our hypothesis that higher performing teachers are better able to prepare for evaluations.

Teachers improve their evaluation score for a more likely evaluation, especially for $P2$ and $M2$ (columns four through nine in Table 2.6). The effects are statistically significant. A teacher that improves her TLF score by 0.6 standard deviations improves her TLF score by roughly 0.4 points on a scale from one to four. The implication is that a teacher who is nearly certain of an evaluation improves her TLF score by roughly half a point relative to a teacher who is very unlikely to be evaluated. In this extreme case, the difference is large enough to bump a teacher into a different IMPACT rating, but the extreme case is rare.

The effects of evaluation probability are concentrated among veteran teachers, which is consistent with our hypothesis that veteran teachers are better positioned to showcase particular skills. When broken out by the previous year's evaluation rating, highly effective teachers appear very capable of improving their scores in response to increased evaluation probability, except for $P3$. The effects by previous-year rating are not significant for $M1$, though veteran teachers overall have a significant effect. The standard errors for effects on $M1$ are large. We suspect that evaluations timed close to the winter break would explain the additional variation. Though we control for the timing of an evaluation, our control assumes that time affects evaluation scores linearly (i.e. overall classroom behavior improves over time). Within the $M1$ window, there is likely large non-linear seasonal variation. Furthermore, for roughly 25 percent of teachers each year, the $M1$ evaluation is their first of the year.

It appears that highly effective teachers respond differently to $P3$ depending on whether or not it is the last evaluation. To test this notion, Table 2.7 divides up the effect of evaluation probability on $P3$ by whether or not it is last. While

last year's minimally effective teachers appear strongly responsive to evaluation probability regardless of whether or not $P3$ is last, last year's effective and highly effective teachers appear to have no response to evaluation probability if $P3$ is the last evaluation of the year. We next turn our attention to consider how teachers modify their behavior to improve their overall observation score.

With nine different separately scored components on each evaluation, our basic multi-tasking model would predict that teachers will shift their attention towards components that are easy to improve. The language describing how each component is scored varies in specificity, where some provide specific examples of teacher behaviors that will score better, while others have more general and vague language. Similarly, some practices are more difficult to adjust within a few days but rather require consistent development over weeks and months.

To identify which specific components teachers are able to adjust in response to the increased probability of an evaluation, we look at how veteran teachers improve each specific component. We use the same specification as before but with the dependent variable changed to each of the nine Teach components. We have also restricted our analysis to the average effect for veteran teachers without breaking the effects out by a teacher's previous rating.

As found in other studies (Adnot, 2016, see), these components are highly correlated, making our hypotheses on observation components highly dependent. We have adjusted the significance levels using a Bonferroni correction.[7] While Teach 2 through 4 are significant sometimes, Teach 8 is significant for all evaluations except $M1$. No components are negatively affected by teacher preparations for an evaluation. As with the overall results, we suspect that $M1$ is affected by non-linear seasonal changes in classroom behavior and the fact that it is the first evaluation for many teachers each year.

The fact that Teach 8 is consistently an area of improvement fits with our prior expectations. Teach 8 is meant to evaluate classroom routines, procedures, and behavior management. While routines and procedures in a classroom are built over time and must be in place for students to respond appropriately, it is possible to pay particular attention to this construct in planning for a given day to obtain a higher score. For example, a teacher who ordinarily allows students to work on a non-academic project after they've completed the lesson may plan to provide students with a more academic-focused activity to satisfy the

[7] For a hypothesis threshold α, the adjusted threshold for significance across m hypotheses is $\alpha^* = \frac{\alpha}{m}$.

"idleness" component. Similarly, a teacher may employ additional patience and de-escalation techniques when dealing with inappropriate or off-task student behavior to ensure it is efficiently addressed, or even spend additional time during the week preparing a challenging student for an evaluation. A teacher may also plan to pay additional attention to giving instructions before a class transition to ensure minimal prompting, whereas in the absence of an anticipated observation the teacher may have been comfortable relying on prompts to redirect student behavior. These behaviors are not easily adjusted on-the-fly, but with some forethought in lesson planning, they can be accomplished more easily.

2.7 Discussion

Our key finding is that teachers respond to the possibility of a high-stakes evaluation as intended by the IMPACT policy. We have shown that teachers are cognizant of elements on the TLF evaluation rubric, and that they take steps to improve their teaching practice as a result of unannounced evaluations. Unlike previous research, we reveal the specific behavioral adjustments teachers make throughout the year in response to unannounced evaluations. The value of our findings is to unpack the mechanism through which teachers achieve the results seen in Dee and Wyckoff (2015). Their analysis shows that teachers perform better on in-class evaluations in response to high-powered incentives but it is silent on how teachers achieve those gains. The key remaining question is, when teachers face large incentives, do they improve measured outcomes by engaging in behaviors that are contrary to the intent of the policy? Our analysis answers this question by mapping the incremental improvements teachers make as evaluations become more likely.

There are two general ways in which a teacher could improve her evaluation scores without truly improving her practice as intended by the policy. The first is having a perfect, pre-planned lesson (sometimes called a "lesson in a box") that she can easily swap in if an evaluator enters her classroom. The second would be to identify some TLF components that she can easily adjust, allowing her to continue with business as usual and only change a single component of her teaching practice as an evaluation becomes more likely. There may be enough low-hanging fruit that guarantees a good score without requiring any real teaching improvements. For example, if she could easily boost her score by adding items to the

whiteboard before class or by completing some checklist of classroom preparation items, she may ignore the other more impactful TLF components.

Our results show that teachers likely do not use a lesson-in-a-box. If they did, we would expect our results for unannounced evaluations to mirror our results for announced evaluations. A perfect lesson-in-a-box would inoculate a teacher against being unprepared, making her unresponsive to evaluation probability. Our results for announced evaluations support this notion: when teachers are prepared, increased evaluation probability does not lead to higher scores. For announced evaluations, the effect of evaluation probability is statistically insignificant and substantively small. Yet for unannounced evaluations, teachers meaningfully change their practice as an evaluation becomes more likely.

Our results also show that teachers improve their teaching practice across multiple desired dimensions in preparation for a possible evaluation. A concern with using behavior-based incentives is that, theoretically, an incentive based on multiple measures can lead to inefficient activities as teachers may focus excessively on one dimension (Holmstrom and Milgrom, 1991). Quality teaching requires skill across multiple dimensions, but if one measured practice is particularly easy to improve, teachers may become less effective along other important practices. Table 2.8 shows teachers consistently improve Teach 8 when expecting an evaluation, which we would expect. Teach 8 assesses specific classroom management techniques that may be easy to implement in preparation for an evaluation. However, the effect of evaluation probability on other Teach components is statistically significant and larger in magnitude. In particular, Teach 3, Teach 4, Teach 6 and Teach 7 all appear to be significantly affected by evaluation probability. We interpret this as evidence that teachers do not simply cherry pick a few easy practices for improving their evaluation score but rather make holistic improvements that affect their score across a variety of domains.

As further evidence that teacher responses to an increasingly likely evaluation are not excessively focused on a single dimension, we use a multivariate regression to test whether evaluation probability has significantly different effects across all nine components. In this test, we fail to reject the null hypothesis that there is any heterogeneity in effect size across Teach components. This highlights that the statistically significant effects seen for Teach 8 are driven by smaller variance in the coefficient, not by a significantly larger magnitude. Our results are consistent with those of Adnot (2016), which found that the majority of variation

in evaluation scores in the IMPACT program can be explained with a single factor that encompasses all nine of the evaluation components.

Our results have secondary implications about the choice of evaluator. The literature is increasingly concerned with who should conduct evaluations, particularly in high-stakes environments. Our main results in Table 2.6 show that principal evaluations are more inert than master educators. The overall effect for veteran teachers are statistically significantly different between $P2$ and $M2$, as well as between $P3$ and $M2$, while the difference between $P2$ and $P3$ is not. However, we are unable to determine why principal evaluation scores are less responsive to evaluation probability, but we consider two possible explanations.

The broader literature shows that principals compress the distribution of evaluation scores, suggesting they do not identify as much nuance in teaching or they are less willing to make errors with strong consequences for teachers. If this is the case, evaluation probability should have a smaller observed effect on scores when principals conduct the evaluation, as we find. As further evidence that principals compress the distribution of evaluation scores, we use Levene's test of homogeneity and confirm that there is a statistically significant difference between the variance of principal evaluations and the variance of master educator evaluations.

There is ample evidence that principals systematically evaluate their own teachers differently than teachers from another school, suggesting a separate hypothesis: teachers may incorporate additional information about students or the teacher in their evaluation. This could be principals considering teacher characteristics like collegiality, for example, or it is possible that principals are less susceptible to teachers showcasing their skills. While our results cannot distinguish between the two hypotheses, these results suggest that the compression of principal evaluation scores observed here (and elsewhere) may be a result of their evaluations being more inert or less manipulatable, which may be a desirable trait.

2.8 Conclusion

The rapid growth of multi-measure teacher incentive programs has progressed with little evidence on how these programs affect daily teacher practice. While seminal work has shown that these programs have the potential to improve student outcomes and increase teacher ratings, no work has revealed how those out-

comes are achieved. The intent of IMPACT is to provide guidance and encouragement for teachers to engage in teaching best practices in multiple domains. Our results demonstrate that teachers respond to the IMPACT evaluation program as intended by improving their teaching practice along multiple dimensions.

Our results highlight the need to ensure that specific rubric constructs and language encourages desired behavior. To this end, DCPS has continued to revise and improve their evaluation rubric, moving to a more conceptual framework in the most recent year (2016-2017). Our analysis suggests that other districts should follow suit, reflecting on the ways in which rubric constructs and language encourage the desired teacher response. The results presented also caution against ad hoc evaluation systems, which are unlikely to provide the needed guidance to change behavior in desired directions.

The difference between master educator evaluations and principal evaluations emphasizes the need to understand why principals as observers are different than outside observers, both in terms of the evaluator quality and in terms of how teachers respond to the evaluator. We are unable to clearly establish the cause of the difference in evaluation malleability between master educators and principals in DCPS. This is particularly salient as DCPS recently stopped using master educators in order to reallocate funds to a greatly expanded professional development program.

Multiple measure teacher evaluation systems have the potential to improve teacher output by prioritizing behaviors in a production process known to be difficult and uncertain. Teachers cannot always know how a particular approach or style will affect students relative to another approach. Phipps (2017) proposes a theoretical model that allows for production uncertainty among employees. His model shows how test-based incentives in such an uncertain context can lead to inefficient effort allocation. A possible solution is to use structured in-class evaluations to reduce teacher uncertainty about their daily practice. In this framework, evaluations play an important role in guiding teaching priorities and in rewarding effective teaching practices. Evaluations then function both as direction and as measurements of teacher quality. Our results demonstrate that multiple measure teacher evaluation systems can satisfy both roles.

Table 2.1: School-level summary statistics on enrollment, class size, and school poverty status.

	2009-2010	**2010-2011**	**2011-2012**
Number of Schools	124	121	123
Total Enrollment	44,035	45,004	45,013
Class Size			
Mean	17.42	17.72	17.65
Std Dev	4.32	4.29	4.12
Fraction of Schools High Poverty	0.771	0.750	0.772

Table 2.2: Summary information on teacher observation scores by year and observer.

	Observation Score on Scale of 1 to 4				
	Principal 1	**Principal 2**	**Principal 3**	**Master Educator 1**	**Master Educator 2**
2009-2010					
Mean	3.113	3.149	3.200	2.963	2.994
Std Dev	0.623	0.645	0.651	0.597	0.602
10th %ile	2.259	2.185	2.259	2.111	2.148
Median	3.185	3.259	3.333	3.074	3.074
90th %ile	3.852	3.889	3.889	3.667	3.704
2010-2011					
Mean	2.980	3.083	3.156	2.877	3.002
Std Dev	0.618	0.589	0.596	0.647	0.584
10th %ile	2.110	2.330	2.380	2.000	2.220
Median	3.000	3.110	3.220	3.000	3.000
90th %ile	3.750	3.780	3.880	3.670	3.670
2011-2012					
Mean	3.161	3.112	3.201	3.004	2.963
Std Dev	0.549	0.564	0.529	0.572	0.563
10th %ile	2.444	2.375	2.556	2.222	2.250
Median	3.222	3.222	3.250	3.000	3.000
90th %ile	3.778	3.778	3.778	3.667	3.625
Total					
Mean	3.085	3.115	3.186	2.948	2.987
Std Dev	0.602	0.602	0.596	0.608	0.584
10th %ile	2.259	2.296	2.380	2.110	2.220
Median	3.125	3.220	3.250	3.000	3.000
90th %ile	3.780	3.815	3.880	3.667	3.670

Table 2.3: Treatment Summary by Year

	Probability of any evaluation on day of evaluation.				
	Principal 1	**Principal 2**	**Principal 3**	**Master Educator 1**	**Master Educator 2**
2009-2010					
Mean	0.156	0.117	0.139	0.064	0.068
Std Dev	0.140	0.114	0.141	0.079	0.077
Min	0.004	0.003	0.002	0.002	0.002
Median	0.113	0.087	0.096	0.043	0.044
90th %ile	0.419	0.325	0.398	0.179	0.203
2010-2011					
Mean	0.143	0.186	0.129	0.059	0.065
Std Dev	0.127	0.189	0.127	0.061	0.068
Min	0.005	0.003	0.001	0.003	0.003
Median	0.108	0.124	0.089	0.040	0.041
90th %ile	0.368	0.619	0.348	0.173	0.196
2011-2012					
Mean	0.128	0.167	0.156	0.052	0.070
Std Dev	0.126	0.159	0.157	0.049	0.076
Min	0.002	0.003	0.001	0.002	0.004
Median	0.092	0.125	0.108	0.039	0.045
90th %ile	0.341	0.477	0.438	0.147	0.209
Total					
Mean	0.142	0.158	0.141	0.058	0.068
Std Dev	0.132	0.161	0.142	0.064	0.074
Min	0.002	0.003	0.001	0.002	0.002
Median	0.103	0.110	0.097	0.040	0.043
90th %ile	0.377	0.482	0.404	0.166	0.201

Notes: Treatment is the probability of being evaluated by either a principal or master educator on the day of an evaluation. Treatment levels are higher for principal evaluations because these are clustered in the last third of the evaluation window.

Table 2.4: Checks for treatment exogeneity.

	P^{P1}	P^{P2}	P^{P3}	P^{M1}	P^{M2}
Previous Reading VA	0.0075	0.0128	0.0009	0.0004	-0.0041
	(0.010)	(0.011)	(0.015)	(0.005)	(0.006)
Previous Math VA	-0.0035	0.0043	-0.0161	-0.0036	-0.0006
	(0.009)	(0.011)	(0.011)	(0.004)	(0.005)
First-Year	-0.0014	0.0083	-0.0019	0.0028	0.0000
	(0.004)	(0.006)	(0.005)	(0.002)	(0.002)
IMPACT ME Last Year	0.0051	-0.0017	0.0003	-0.0013	0.0002
	(0.006)	(0.007)	(0.006)	(0.002)	(0.003)
IMPACT HE Last Year	-0.0047	-0.0149**	-0.0057	0.0004	-0.0025
	(0.005)	(0.007)	(0.006)	(0.002)	(0.004)
Eval ME Last Year	0.0056	0.0039	0.0071	0.0002	0.0043
	(0.006)	(0.007)	(0.006)	(0.002)	(0.004)
Eval HE Last Year	0.0020	0.0020	0.0016	-0.0011	0.0023
	(0.003)	(0.004)	(0.004)	(0.002)	(0.002)

Notes - Each cell represents the estimated correlation between the row label (observable characteristic) and the column label (treatment variable). Standard errors are in parentheses. All variables except First-Year are centered around the school mean to control for school fixed effects. Errors are clustered at the school by year level. Significance levels are not adjusted for multiple hypothesis testing.

*** Significant at the 1 percent level
** Significant at the 5 percent level
* Significant at the 10 percent level

Table 2.5: Effect of evaluation probability on announced evaluations.

	Announced				
	Principal Eval 1 All Years TLF Score in Std Devs			Master Educator Eval 1 2009-10 Only TLF Score in Std Devs	
Evaluation Probability (P^k)	0.086 (0.106)	0.170 (0.138)	-	-0.036 (0.317)	-0.045 (0.280)
P^k X First-Year Teacher	-	0.275 (0.293)	-	-	-0.212 (0.937)
Minimally Effective Last Year (ME)	-	-	-0.698*** (0.082)	-	-
Highly Effective Last Year (HE)	-	-	0.317*** (0.039)	-	-
P^k if ME Last Year	-	-	0.574 (0.417)	-	-
P^k if Effective Last Year	-	-	0.130 (0.133)	-	-
P^k if HE Last Year	-	-	-0.176 (0.155)	-	-
First-years Excluded			X		
N	9457	9457	5850	3035	3035

Notes - This table shows estimates of how a teacher's in-class evaluation score improves as an announced evaluation becomes more likely. We estimate evaluation probability using an approximation of how many teachers will be evaluated at each school on each day divided by the number of teachers remaining to be evaluated. Standard errors are shown in parentheses. The probability of an evaluation should not effect an announced evaluation score. Estimating effects by previous-year teacher rating is only possible for 2010-11 and 2011-12.

*** Significant at the 1 percent level
** Significant at the 5 percent level
* Significant at the 10 percent level

Table 2.6: Effect of evaluation probability on unannounced evaluations.

	Unannounced											
	Master Educator Eval 1 2010-11 and 2011-12 TLF Score in Std Devs			Principal Eval 2 All Years TLF Score in Std Devs			Master Educator Eval 2 All Years TLF Score in Std Devs			Principal Eval 3 All Years TLF Score in Std Devs		
Evaluation Probability (P^k)	0.221	0.527**	-	0.282***	0.280***	-	0.560***	0.582***	-	0.107	0.194**	
	(0.219)	(0.212)		(0.079)	(0.070)		(0.161)	(0.162)		(0.074)	(0.077)	
P^k X First-Year Teacher	-	-0.370	-	-	0.137	-	-	0.866*	-	-	0.223	
		(0.967)			(0.204)			(0.516)			(0.217)	
Minimally Effective Last Year (ME)	-	-	-0.238***	-	-	-0.365***	-	-	-0.323***	-	-	-0.649***
			(0.067)			(0.064)			(0.072)			(0.106)
Highly Effective Last Year (HE)	-	-	0.159***	-	-	0.102***	-	-	0.152***	-	-	0.062
			(0.038)			(0.030)			(0.036)			(0.054)
P^k if ME Last Year	-	-	0.353	-	-	0.528*	-	-	0.965	-	-	0.416**
			(0.787)			(0.289)			(0.632)			(0.202)
P^k if Effective Last Year	-	-	0.144	-	-	0.269***	-	-	0.481*	-	-	0.211**
			(0.298)			(0.093)			(0.248)			(0.097)
P^k if HE Last Year	-	-	0.027	-	-	0.336***	-	-	0.774**	-	-	-0.102
			(0.304)			(0.120)			(0.306)			(0.138)
First-years Excluded						X			X			X
N	6430	6430	5885	8624	8624	5500	9122	9122	5622	9124	9124	5562

Notes - This table shows estimates of how a teacher's in-class evaluation score improves as an unannounced evaluation becomes more likely. We estimate evaluation probability using an approximation of how many teachers will be evaluated at each school on each day divided by the number of teachers remaining to be evaluated. Standard errors are shown in parentheses. Estimating effects by previous-year teacher rating is only possible for 2010-11 and 2011-12. Results are arranged in the rough order in which evaluations are conducted each year.

*** Significant at the 1 percent level
** Significant at the 5 percent level
* Significant at the 10 percent level

Table 2.7: Effect of evaluation probability on third principal evaluation based on evaluation order.

	Principal Eval 3 All Years TLF Score in Std Devs	
	Not Last	Last
P^k if ME Last Year	2.227**	0.335
	(1.076)	(0.214)
P^k if Effective Last Year	1.011***	0.160
	(0.283)	(0.105)
P^k if HE Last Year	0.748*	-0.175
	(0.430)	(0.146)
First-years Excluded	X	X
N	5562	

Notes - This table shows estimates of how a teacher's third principal evaluation score improves as it becomes more likely, broken out by whether or not the evaluation is the last of the year. We estimate evaluation probability using an approximation of how many teachers will be evaluated at each school on each day divided by the number of teachers remaining to be evaluated. Standard errors are shown in parentheses.

*** Significant at the 1 percent level
** Significant at the 5 percent level
* Significant at the 10 percent level

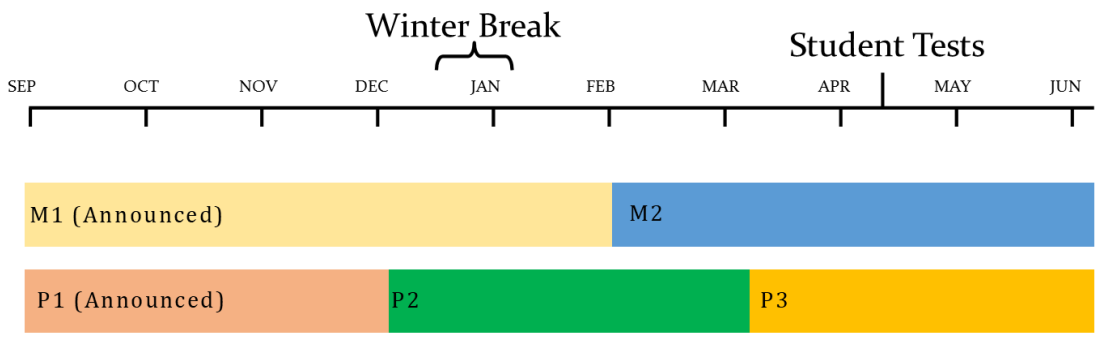
Table 2.8: Effect of evaluation probability on individual evaluation components.

	Effect of Eval Probability on Component Score (Std Devs)			
	M1	P2	M2	P3
Teach 1	0.001	0.108	0.362	0.287
	(0.257)	(0.101)	(0.182)	(0.219)
Teach 2	0.214	0.121	0.411	0.138
	(0.249)	(0.073)	(0.183)	(0.204)
Teach 3	0.048	0.492***	0.424	0.543
	(0.247)	(0.095)	(0.172)	(0.218)
Teach 4	0.076	0.187	0.518**	0.475
	(0.242)	(0.095)	(0.180)	(0.211)
Teach 5	0.238	0.054	0.270	0.362
	(0.250)	(0.097)	(0.183)	(0.175)
Teach 6	0.149	0.298**	0.261	0.529
	(0.454)	(0.099)	(0.256)	(0.282)
Teach 7	-0.017	0.257**	0.246	0.371
	(0.249)	(0.092)	(0.177)	(0.225)
Teach 8	0.040	0.398***	0.495**	0.492*
	(0.242)	(0.081)	(0.152)	(0.189)
Teach 9	0.100	0.086	0.141	0.026
	(0.214)	(0.074)	(0.169)	(0.219)
N	5960	7911	8332	8326

Notes - Teachers are scored along nine components (Teach 1 through Teach 9). All coefficients are in standard deviations. Evaluations conducted by the principal are $P2$ and $P3$, while evaluations conducted by the district employee are $M1$ and $M2$. Results for $M1$ are for the 2010-11 and 2011-12 years only. All significance levels have been adjusted using a Bonferroni correction factor. Sample excludes first-year teachers.

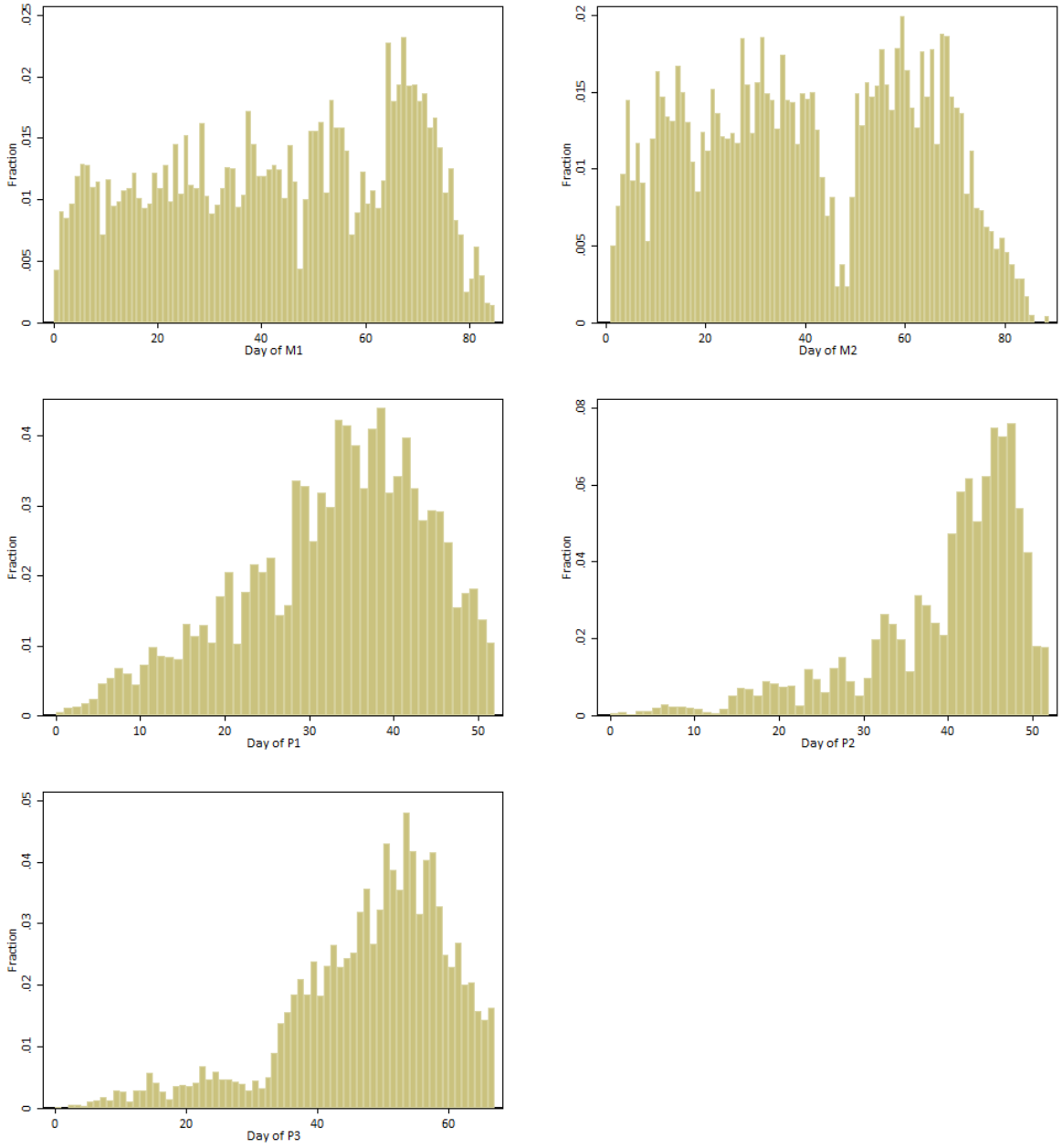
*** Significant at the 1 percent level
** Significant at the 5 percent level
* Significant at the 10 percent level

Figure 2.1: Depiction of each evaluation window in DCPS.



Notes - Each shaded area represents the time-frame in which a teacher must receive an evaluation. Teachers have 5 evaluations that occur in overlapping windows. P1, P2 and P3 are evaluations administered by the principal or vice principal. M1 and M2 are evaluations administered by a district employee called a master educator.

Figure 2.2: Histograms of evaluatoin timing within evaluation windows.



Notes - Days are measured as instruction days, which excludes in-service days, weekends, and holidays. Master educator evaluations, $M1$ and $M2$, are distributed uniformly across the window. Principal evaluations – $P1$, $P2$ and $P3$ – are often clustered near the end of each window.

Chapter 3

Teacher Performance Pay through the Lens of Production Uncertainty: Theory and Evidence from a Real-Effort Laboratory Experiment

Teacher performance incentives do not consistently improve student test outcomes in the US. A necessary theoretical and empirical problem is how to design an incentive to induce the optimal allocation of effort among multiple tasks, which is usually modeled by assuming agents know the production function. Unlike some production processes in which output relies solely on worker skill and effort, teaching is distinguished by its complexity and its dependence on the reciprocal effort of students. As a result, individual teachers are uncertain about the net marginal productivity of inputs. The innovation of this paper is to develop a model that explicitly incorporates uncertainty about the production process in student learning ("production uncertainty") in the model of behavior, and then to assess agent responses to a standard piece-rate performance pay scheme as the variance of an input's marginal payoff increases in a real-effort laboratory experiment.

3.1 Introduction

As it is well established that teacher experience and educational credentials are largely uncorrelated with teacher productivity, compensation based on these factors alone is a poor tool for attracting, motivating and retaining a strong teaching labor force (Rivkin et al., 2005). Yet, how to compensate teachers to reflect productivity differences and provide incentives that reward performance remains a practical challenge, as many alternative compensation schemes have not demonstrated consistent efficacy (Dee and Wyckoff, 2015). Recent large-scale federal initiatives in the United States, including Race to the Top and the Teacher Incentive Fund place substantial emphasis on the development of compensation mechanisms to link pay and performance. These programs have awarded a combined $6.4 billion since 2010 to 92 districts in 32 states for proposals to fund new teacher performance incentives or improve teacher accountability. Internationally, the United Kingdom, India, Chile, Mexico, Israel, Australia, and Portugal are considering or have implemented teacher incentive programs.

While large scale field experiments and policy innovations are surely needed in the teacher labor market, the design challenge resides in the more general space of models of incentive contracts (Lazear, 1986; Holmstrom and Milgrom, 1991; Prendergast, 1999). A necessary theoretical and empirical problem is how to design an incentive to induce the optimal allocation of effort among multiple tasks, which is usually modeled by assuming agents know the production function. Unlike some production processes in which output relies solely on worker skill and effort, teaching is distinguished by its complexity and its dependence on the reciprocal effort of students. As a result, teachers are uncertain about the net marginal productivity of teaching inputs. An innovation of this paper is to develop a model that explicitly incorporates uncertainty about the production process in student learning ("production uncertainty") in the model of behavior. More importantly, this paper develops empirical evidence in a laboratory experiment assessing agent responses to a piece-rate performance pay scheme as I vary the variance of input payoffs without changing their average payoff.

In the model, the presence of production uncertainty reduces the effect of an outcome-based incentive on a teacher's overall effort level due to risk aversion, an effect I label "futility." Furthermore, an outcome-based incentive can induce pro-

duction friction, which predicts that teachers will redistribute their effort to inputs with lower variance in marginal productivity, potentially reducing average total productivity without necessarily decreasing overall effort. For example, teachers may teach to the test instead of using more pedagogically sound techniques, resulting in lower average test scores (Blazar and Pollard, 2017). The initial empirical test of these predictions is in a laboratory experiment, which is intended to provide a controlled setting to isolate the potential effects of increasing production uncertainty and should be seen as an opportunity to explore behavioral mechanisms without the expense and implementation challenges of a field experiment or policy change. Initial results of this experiment confirm the predictions of my model, though additional study is necessary.

The experimental design itself is an innovation, allowing me to directly test whether increases in production uncertainty induce inefficiency while holding everything else constant. The policy innovation of this paper is to examine the salience of production uncertainty in designing incentive contracts. For teachers, this will provide direction on whether incentives should be based on in-class evaluations or student test scores. This paper also contributes to the design of future field experiments on teacher incentives by identifying the relevant features that should be varied experimentally, an innovation given the tenuous theoretical basis for the design of most teacher performance incentives to date. In the past, researchers have expended tens of millions of dollars on field experiments that yielded no results without establishing a credible theoretical and empirical design that addresses the key complications of performance incentives in a context as complex as teaching. The theoretical innovation of this paper is important to contract design in industries with well-defined output measures, but it also creates a theoretical framework for modeling more complex professions, such as health care provision, that have been neglected in contract theory.

This paper tests whether or not agent effort allocation among tasks is affected by production uncertainty in a real-effort, multitask experiment, which provides an empirical test for the distinctive predictions of my model. The frontier of contract theory provides little guidance for fundamental teacher incentive design questions, such as whether or not to use classroom observations, test scores, or group-based incentives. The existing empirical work on teacher performance incentives has a mixture of results that are not easily understood through the lens of the existing personnel economics literature. The questions addressed

in this paper are at the intersection of contract theory and education policy with contributions to both.

3.2 Background

Good teachers have a meaningful effect on student outcomes immediately and later in life, making teachers a possible public lever for increasing the human capital of a nation (Rockoff, 2004; Rivkin et al., 2005; Kane and Staiger, 2008; Aaronson et al., 2007; Chetty et al., 2014). Because of the wide variation in teacher effectiveness, policies that improve the lower tail of the teaching labor force would yield meaningful improvements in overall student development.[1] To this end, school districts attempt to promote more teacher training and experience by using what is called a "steps and lanes" system (or "single salary"), which makes salaries depend only on education, certification, and teaching experience. But the available evidence establishes that these factors do not translate into improved student outcomes (Rivkin et al., 2005). Such pay systems are discouraging to young but effective teachers looking to distinguish themselves in their career.[2] Steps and lanes systems also lack any mechanism to motivate the creation and maintenance of reliable measures of teacher quality, which leads to unfocused and ineffective professional development.[3] An alternative payment scheme would seek to identify effective teachers based on their performance and reward them accordingly, a general policy idea I call performance incentives. A handful of school districts in the US have attempted to create performance incentives throughout the 20th century, but few of these programs survived for more than a couple years, at least until recently.

[1] Hanushek (1992) finds that a teacher at the 95th percentile will improve student outcomes by the equivalent of 1.5 academic years while a teacher at the 5th percentile improves outcomes by only 0.5 years.

[2] Hoxby and Leigh (2004) estimate that the share of teachers in the highest aptitude category fell from 5% to 1% from 1963 to 2000, and they estimate that 80% of this decline can be attributed to high quality teachers being pushed out of the profession because of the lack of pay differentiation.

[3] Weisberg et al. (2009) examines the lack of differentiated teacher evaluation systems across the country and then argues that professional development programs consistently have no measurable effect on student outcomes as a result of being unfocused.

3.2.1 Teacher Performance Incentive Programs in the US and Their Effects

Rewarding teachers for performance (also called "merit pay" or "performance pay") is an old, recurring idea since at least the early 20th century (Murnane and Cohen, 1986). The last 17 years have seen a strong resurgence of school and teacher accountability, starting with the No Child Left Behind Act on through the Teachers Incentive Fund and the Every Student Succeeds Act. In response, school districts continue to implement a variety of incentive and accountability initiatives, which I will briefly summarize.

Some districts pay bonuses to teachers based on group performance, requiring the entire school to reach a graduation threshold or all teachers in a specific grade range improve student test scores sufficiently.[4] These programs introduce a new set of theoretical design questions that complicate any effort to pinpoint crucial teacher incentive elements. Some of these concerns are discussed and evaluated empirically in Imberman and Lovenheim (2015). Moving forward, I will focus on teacher incentives based on individual performance.

Empirical evidence of the effects of teacher-level performance incentives is mixed. Dee and Keys (2004) found that an in-class evaluation based bonus implemented at the same time as the Tennessee STAR classroom size experiment, conducted in 1985, improved student math scores. The working paper Hudson (2010) and a technical report Mann et al. (2013) use propensity score matching to show that schools that implement the Teacher Advancement Program (TAP) had higher math scores than the schools they are matched with, though the evidence is less convincing. Dee and Wyckoff (2015) use discontinuities in bonus thresholds in the Washington DC IMPACT program (started in 2009 and ongoing) to show that teachers close to receiving large pay increases will improve their students' test scores more in the following year than teachers just above the threshold.

Studies of other programs find some positive effects, but they were small or not consistent across districts. Sojourner et al. (2014) find some small positive effects in the Minnesota Quality Compensation (Q-Comp) program, which started in 2005 and is ongoing. The statewide program is notable because it allows districts to design their own evaluation and professional development programs

[4]Notable examples of school-level incentive programs are the Kentucky Instructional Results Information System (KIRIS) (Koretz and Barron, 1998), the North Carolina ABC program (Vigdor, 2008), Chicago's modified implementation of the Teacher Advancement Program (Glazerman and Seifullah, 2012), and the New York City School-Wide Bonus Program (SWBP) experiment analyzed in Fryer (2013).

within loose guidelines. As for the results of the many programs funded by the Teacher Incentive Fund (TIF), the Department of Education's Institute of Education Sciences produced a report that shows small positive effects after three years, although there is variation in some key components of program implementation that makes the average results difficult to interpret (Wellington et al., 2016).

Other well-publicized, sizable programs showed little or no effects on student test scores. The Tennessee Project on Incentives in Teaching (POINT) was a three-year experiment started in 2006. While selection into the experiment was voluntary, assignment to treatment was randomized. Treated teachers would receive bonuses based solely on the test score improvements of their students. There were no significant positive effects from the incentive (Springer, 2010). The Denver Professional Compensation program (ProComp), started in 2007, created several routes for teachers to receive bonuses, but by far the largest bonus was awarded to teachers with large gains in student test scores. A report from the University of Colorado, Boulder finds that this incentive had no positive effects on student test scores (Briggs et al., 2014).

In all, this body of empirical results is far from a conclusive endorsement of teacher performance incentives. Each of these programs, except the Tennessee STAR experiment, uses student test score improvements as part of their incentive, and the incentive sizes are mostly comparable. The consistent element of effective programs is their use of in-class evaluations.

To illustrate the importance of in-class evaluations, I divide up the teacher incentive programs detailed above in Table 3.1 based on the use of incentives for in-class evaluation scores and any measurable incentive effects.[5] All of these incentive programs have teacher-level measurements and incentives with minimal school-level incentives. Six of the seven programs use value-added measures. While not a causal argument, Table 3.1 highlights a positive relationship between a performance incentive's effect and the use of in-class evaluations.

Of course, no two teacher performance incentive programs are directly comparable. There are many local factors and implementation decisions that will change the outcomes of a program. However, given the available evidence, there

[5]To assess the quality of in-class evaluations, I consider how much they differentiate among teachers. Weisberg et al. (2009) documents a known issue with many in-class evaluation programs. The majority of these systems to date have produced essentially no differentiation among teachers. For the purposes of this back-of-the-envelope calculation, I've defined differentiation to mean that fewer than 90% of teachers receive the same in-class evaluation score available. This rule is only necessary to distinguish the Denver ProComp program, in which over 99 percent of teachers receive a passing score.

is a strong case that in-class evaluations are a key element of effective teacher incentives. But if teachers can predictably improve student test scores by engaging in behaviors measured by in-class evaluations, why are they not using these behaviors in the absence of in-class evaluations to earn bonuses based on student test scores? To illustrate the gap between these empirical results and existing theory, I provide a brief overview of the relevant contract theory.

3.2.2 Incentive Contract Theory and Evidence in the Education Context

Economists have created a rich set of models depicting incentive contracts that address a variety of situations, not all of which are applicable in the teaching context (Prendergast, 1999). There are a few key theoretical findings relevant to teaching. Performance incentives should be based on measures of the final good, not individual inputs or multiple outputs (Holmstrom and Milgrom, 1991). In the teaching context, this implies that if policy makers ultimately want to improve student test scores, incentives should be based on student test scores. Another theoretical prediction is that incentives have stronger effects as measurement noise decreases (Lazear, 1986), which suggests test-based incentives will have limited effects if test scores have a lot of unexplained variation. Baker (1992) provides a concise model illustrating the importance of creating an incentive that is difficult to cheat, a prediction empirically validated by teachers cheating on student tests in Chicago (Jacob and Levitt, 2003). Neal (2011) adds additional insights in the context of teaching using a simplified two-input principal-agent model. He shows that if the teacher's input costs are separable, it is always optimal to have some performance-based pay. If input costs are not separable, it is possible that a test-based incentive will induce teachers to substitute towards teaching practices that promote student test scores without improving the true desired outcome, illustrating the fundamental importance of ensuring student tests accurately capture desired student outcomes. Secondly, his model demonstrates that setting teacher performance standards too high will have essentially no effect on teacher behavior.

There is strong evidence that performance incentives improve employee effort. For example, Lazear (2000) finds that a piece-rate wage in the Safelite Glass Corporation led to significant improvements in output. In a firm-level experiment, Bandiera et al. (2007) shows that managers receiving a performance incentive in-

crease the productivity of their team. In education, the successful application of performance incentives in some school districts suggest that teaching can be an appropriate context for performance incentives, but contract theory does not provide insight as to what distinguishes teaching from other industries, and it cannot identify the common traits of effective teacher incentives. The consistent element of effective teacher performance incentives is the use of in-class evaluations, contrary to existing theory. If improving student test scores is the goal of policy makers, paying teachers for improving test scores should be at least as effective as paying teachers for improvements on in-class evaluations. Principal-agent models determine optimal contracts based on output measurement noise and cost, employee risk aversion, and complementarity between inputs. Yet none of these dimensions accurately predict an effective incentive.

It is possible that the mixed results of teacher incentives are due to variation in implementation, not design. To that end, ten school districts agreed to roll out their incentive program in stages as part of their TIF agreement. In the first stage, a random subset of schools would implement the new program, while the other schools would not change their payment structure. Wellington et al. (2016) evaluate these ten programs along six dimensions, including differentiation of bonuses between teachers, the difficulty of earning a bonus, bonus timing, and teachers' understanding of the program and their own eligibility to receive a bonus. None of these implementation characteristics correlated with incentive effectiveness.

I argue that there is a more fundamental issue that has gone unexamined in both the empirical and theoretical literature: the production function for improving student test scores is full of uncertainty, dulling the effectiveness of a test-based incentive and inducing inefficient time allocation. Fryer (2013) briefly discusses the possibility that teachers do not know the education production function in the context of a large, group incentive in New York City. In his context, teachers are unresponsive to the group incentive, which he argues is likely due to the complexity of the incentive program, not teacher ignorance of the teaching production function. However, the model I present creates a clear distinction between teachers having an inaccurate understanding of the teaching production function as opposed to having an imprecise understanding. While both are problematic, Fryer is largely arguing against the notion that teachers do not know what good teaching looks like or which inputs lead to improved outcomes (inaccuracy). Given the known correlation between effective teaching practices as

measured through in-class observations and student outcomes (Kane and Staiger, 2012), I agree with Fryer that it seems unlikely that teachers are unaware of what constitutes effective teaching. But I show, in theory and in this experiment, that risk-averse teachers with a large test-based incentive may favor inefficient inputs even if their understanding of the production function is accurate.

To advance this theory, I add a new dimension to the principal-agent problem by allowing teachers to have uncertainty about the marginal productivity of each input, which I call production uncertainty. I can then divide programs into two types: incentives based on student test score improvements and those based on classroom observations of teacher performance. The key difference between the two incentives is that, from a teacher's perspective, the production function for improving student test scores is considerably more uncertain than the production function for improving in-class observation scores. The result is that incentives based on in-class evaluations may be significantly more effective at improving student test outcomes than incentives based on the outcomes themselves. The purpose of this experiment, then, is to test the prediction that uncertainty about the production function induces inefficient behavior, even when participants have an accurate understanding of the average effect of each input.

My experiment builds on the well-established experimental literature using real-effort tasks to examine participant responses to incentives (for example, see Van Dijk et al., 2001; Oswald et al., 2015). An innovation of this experiment is to create two real-effort tasks of varying difficulty that will have differing payoffs based on the participant's skill. A similar setup could be used to test predictions of the Holmstrom-Milgrom multi-tasking model, yet I am unaware of any such experiment. There are, however, experiments testing participant behavior under group-based and individual-based incentives in a multi-tasking environment, though the tasks vary only in how they affect individual earnings and group earnings (Fehr and Schmidt, 2004; Hoppe and Kusterer, 2011; Oosterbeek et al., 2011; Lynn Hannan et al., 2013). The other innovation of my experiment is to simulate production uncertainty on multiple dimensions. Zubanov (2013) tests the effects of inducing uncertainty about the marginal payoff of a single input, "multiplicative noise," but the predictions of his model and the outcomes of his experiment are indistinguishable from adding measurement noise in a standard piece-rate payment scheme. By examining uncertainty along multiple dimensions, my experiment illustrates how participants, without reducing overall effort, become less

productive as a result of differences in their uncertainty about the payoffs of two inputs.

3.3 Theoretical Model

In a manner similar to the Holmstrom-Milgrom model, consider how a teacher can distribute her total time τ, which includes both time in class and time spent preparing for class, among N different tasks. Call this time allocation choice x, which is an N-by-one vector such that $\sum_{i=1}^{N} x_i = \tau$. Example tasks could include showing the class a movie, conducting an experiment with the class, or lecturing with slides.[6]

Then suppose the teacher receives a wage that is based on her students' test scores, which have production function $f(x)$. The wage rule is $w(x) = w_0 + w_1 f(x)$, where w_0 is a guaranteed salary and w_1 is a piece-rate performance incentive. For notational simplicity, I use $\dot{1} = [1, 1, \ldots, 1]^T$, where the superscript T indicates the transpose. Then assuming she is risk averse, her utility can be expressed using the exponential utility function

$$U = -\exp\{-r\left(b_w(w_0 + w_1 f(x)) + b_l(1 - \dot{1}^T x)\right)\}. \tag{3.1}$$

The parameter r indicates the teacher's coefficient of risk aversion. The parameters b_w and b_l weight the utility gains from wages and leisure time.[7]

My innovation is to relax the assumption that a teacher knows the production process $f(x)$. A teacher is likely uncertain about the marginal effect of one teaching rubric or approach relative to another. Furthermore, because education is a user-input production process, the effort of students and their parents appears as a random variable to the teacher. I model this by allowing the marginal value of each input to be a random variable.

To model production uncertainty, I first assume teachers linearly approximate the production process. This can be accomplished mathematically by taking

[6]This model can easily be adapted to allow for effort intensity as well as time, where leisure is some combination of quality and time, and each task is a combination of intensity of effort and time. The cost function can also be made non-linear, but this is ignored for simplicity at the moment.

[7]At this point, the model can easily be adapted to the Holmstrom-Milgrom model by making $f(x)$ a vector of measured outputs, each with some measurement error. I omit measurement error for simplicity, but its inclusion does not fundamentally change the predictions of my model.

a first-order, Taylor Expansion around some reference input value x^r:

$$f(x) = f(x^r) + [(x - x^r)^T \nabla f(x^r)] + \xi(x - x^r) \tag{3.2}$$

where $\nabla f(x^r)$ is the gradient of $f(x)$ and $\xi(x - x^r)$ is a remainder function that is increasing in the distance between x and the reference input level, x^r. Then let γ be a vector of random variables with mean $\nabla f(x^r)$ and a multivariate normal distribution with covariance matrix Σ_γ. Then a rewriting the teacher's approximation of the production function:

$$f(x) \approx f(x^r) + \gamma^t x, \tag{3.3}$$

where I have omitted the remainder function.[8] If the variance in γ is zero, the teacher's problem becomes fairly trivial, and she will devote all her non-leisure time to the input with the highest net marginal productivity.

Substituting the approximated production process into the expected utility function yields

$$EU = E\left[-\exp\left\{-r\left(b_w(w_0 + w_1\gamma^T x) + b_l(1 - \dot{1}^T x)\right)\right\}\right] \tag{3.4}$$

Because γ^T has a multivariate normal distribution, I use the moment generating function to simplify the expected utility:

$$EU = -\exp\left\{-r\left(b_w w_0 + b_l(1 - x^T\dot{1}) + b_w w_1 x^T \mu_\gamma - \frac{1}{2}r(b_w w_1)^2 x^T \Sigma_\gamma x\right)\right\}. \tag{3.5}$$

The optimal choice of inputs x^* is then

$$x^* = \frac{2}{r(b_w w_1)^2}\Sigma_\gamma^{-1}\left(b_w w_1 \mu_\gamma - b_l\dot{1}\right). \tag{3.6}$$

To explore the properties of this model, consider the two-input case. Let σ_{11} and σ_{22} denote the variances of γ_1 and γ_2, and let σ_{12} be their covariance. I can allow

[8] If the remainder function is instead considered a random variable with mean increasing in $x - x^r$, the final effect on the model's predictions is similar to adding measurement noise, except the noise's negative effect depends on a teacher's reference point x^r. Notice also that this specification assumes that the inputs are perfect substitutes in the traditional sense, but the covariance between γ_i and γ_j allows for an idea of risk substitution. That is, if $cov(\gamma_i, \gamma_j) > 0$ then if x_i has high marginal productivity, then x_j is likely to also have high marginal productivity. Taking a higher order Taylor Expansion will allow for complementarity in a more traditional sense.

for differentiated marginal costs of x_1 and x_2 by allowing c to be the marginal cost of x_2 relative to x_1. Then the optimal choice of effort for input 1, x_1^*, is

$$x_1^* = \overbrace{\left(\frac{2}{r(b_w w_1)^2}\right)}^{A} \overbrace{\left(\frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\right)}^{B} \left(- \overbrace{(b_w w_1 \mu_2 - b_l c)\sigma_{12}}^{C} + \overbrace{(b_w w_1 \mu_1 - b_l)\sigma_{22}}^{D} \right). \tag{3.7}$$

The terms labeled A and B are multiplicative constants that will affect all inputs. Term A shows that a teacher will reduce effort in all tasks as her risk aversion r increases. There is also an income effect where increases in b_w or w_1 will decrease effort in all tasks, though this is partially offset by a substitution effect in terms C and D.

Term B can be thought of as an uncertainty premium multiplier. If σ_{12} is positive, it provides a form of risk reduction. This is somewhat intuitive since positive covariance increases the information about the distribution of the marginal productivities of each input.[9] Term C is the employee's response to inputs being risk substitutes or complements, depending on the sign of σ_{12}. Assuming input 2 has higher net marginal utility (i.e. $(b_w w_1 \mu_2 - b_l c) > 0$), then the employee will substitute effort towards input 2, other things constant. Intuitively, if two inputs have highly positively correlated marginal productivities, there is less risk associated with putting more effort into just one input, hence the inputs act as substitutes in risk diversification.

Testable Predictions from Production Uncertainty

I now focus on the two simple predictions that I will test directly in my lab experiment. To illustrate these, I further simplify the model by assuming $\sigma_{12} = 0$, which allows me to rewrite the optimal input choices as

$$x_1^* = \left(\frac{b_w w_1 \mu_1 - c b_l}{r(b_w w_1)^2}\right)\frac{2}{\sigma_{11}}, \qquad x_2^* = \left(\frac{b_w w_1 \mu_2 - b_l}{r(b_w w_1)^2}\right)\frac{2}{\sigma_{22}}. \tag{3.8}$$

The two main predictions I test with my experiment are Futility and Friction, which I introduce with abbreviated proofs:

[9] The interpretation of Term B is aided by a more generalized interpretation of the determinant of a matrix. In short, the determinant can be thought of as a volume measure of the covariance matrix in n-dimensional space. If the covariance occupies a greater volume (higher determinant), this will decrease overall effort.

Proposition 1 (Futility)**.** In the absence of covariance, an increase in the variance of the marginal productivity of either input will reduce overall effort, $\sum_{i=1}^{N} x_i$.

Proof. Follows immediately from Equation 3.8. □

In the experiment, this will be tested by measuring participants' overall effort in completing their tasks, and then varying the magnitudes of σ_{11} and σ_{22}. I measure participant effort by observing their speed and effectiveness in answering questions, as well as observing their decision to use up all the available time in a round.

Proposition 2 (Friction)**.** A performance incentive with production uncertainty will induce employees to distribute their effort among tasks such that the expected productivity is strictly less than the maximum expected productivity with the same total effort.

Proof. Because the production function is linear, the maximum average productivity requires that $\sum_{i=1}^{N} x_i = x_{max}$ where $x_{max} = \{x_i : \mu_i > \mu_j \, \forall j \neq i\}$. Under a performance incentive with production uncertainty, it immediately follows from Equation 3.8 that $x_i > 0$ for all i, which implies that $\sum_{i=1}^{N} x_i \neq x_{max}$, and therefore expected productivity could be improved by re-allocating the same total amount of effort. □

I can test friction in my experiment by observing how participants re-allocate their overall effort away from the task with the highest expected marginal productivity as I change σ_{11} and σ_{22}. Importantly, going from zero production uncertainty to any production uncertainty is bound to have a negative effect on how participants allocate their effort. The key prediction of production uncertainty is that changing the variance of an input, without changing its average productivity, will induce participants to use that input less.

Production uncertainty provides new insights into the salient issues of incentive design, particularly in complex production environments like teaching. For comparison, the multi-task model in Neal (2011) summarizes the standard theoretical concerns of teacher performance incentives well. In his model, teachers increase human capital and are paid based on their score in a performance review, say teacher value-added. Factors that will reduce the effect of a performance incentive on both teacher value-added include noisy test scores, setting the requirements to receive a bonus too high, or making bonuses too small. Yet all available evidence suggests that these characteristics do not predict an effective program as measured by improvements in student test scores. Instead, test

scores improve when an incentive uses in-class observations. To place this in the context of Neal's model, assume for the moment that test scores perfectly capture increases in human capital. Then the question I am posing is, should incentives be based on in-class evaluations or test scores? Neal's model unequivocally concludes that any test-based incentive is efficient, while there are many ways in which an incentive based on in-class observations could be inefficient and have no measured effect on test outcomes. In short, the incentive based on in-class observations could never be more effective at improving test scores. Production uncertainty bridges this gap between theory and the empirical literature.

Production uncertainty is an inherent trait of the production function. In the teaching context, even if it were possible to measure with absolute accuracy how much a student has learned throughout the year, the process of actually imparting that knowledge still has uncertainty. My model illustrates unexplored, potentially adverse effects of performance incentives. In particular, relative to an incentive based on in-class evaluations, test-based incentives may induce an inefficient allocation of effort, decreasing average productivity without even reducing overall effort. To illustrate, consider a teacher with two possible inputs for teaching math: she can use rote memorization ("teach to the test") or use a more pedagogically sound approach that employs counting cubes. While a teacher's training and well-known best practices suggest the counting cubes are a better approach on average, its effect may have more variance than just rote memorization. As a result, the teacher may favor rote memorization because, in certainty equivalent terms, its marginal benefit is higher. Again, existing theory only predicts that teachers switch to rote memorization *if it improves test scores*, while production uncertainty shows it is possible that teachers switch even when it may be less effective at improving test scores, on average. Indeed, recent research finds that teachers' attempts to teach to the test result in lower test scores on average (Blazar and Pollard, 2017; Hill et al., 2015; Blazar, 2015).

3.4 Experimental Design

The goal of the experiment is to simulate production uncertainty in a multitasking environment. Participants must allocate their time between two tasks of varying difficulty and financial payoffs. This reduces the contract problem to a key salient variable: the variance in the marginal productivity of inputs. The lab

has the advantage of treatment randomization and precise control over the degree of production uncertainty.

To place the experiment in the context of the model, I compare how efficiently participants allocate their time between easy and hard addition questions (x_1^* and x_2^* in Equation 3.8) while holding the average marginal value of each input constant (μ_1 and μ_2) and varying only the uncertainty about each input's marginal value (σ_{11} and σ_{22}). The model predicts that as uncertainty about the payoff of easy questions (σ_{11}) increases, time spent on easy questions (x_1^*) will decrease, all else constant. Such a shift reduces a participant's average payoff even though the average marginal value of an easy question remains unchanged.

If there were absolutely no uncertainty about the payoff of each input ($\sigma_{11} = \sigma_{22} = 0$), participants would dedicate all their time to the input with the highest payoff per minute. However, in a real-effort experiment, I can never truly make the uncertainty about the marginal payoffs disappear entirely since there will always be some variation in how long a question will take, even among questions of the same difficulty rating. That is, even two easy questions will have variation in how easy they turn out to be for the participant, which will subsequently create uncertainty about the payoff per minute of an easy question. In the control group, I minimize the payoff uncertainty by guaranteeing a fixed payoff for each question type. Then, to induce additional uncertainty in the treatment group, I randomize the payoff for each question type between rounds.

3.4.1 Experimental Procedures

Participants use a web browser to engage in the tasks while a proctor controls the flow of the experiment in an administrative dashboard. The basic innovation is to present participants with a choice of two possible tasks (inputs). In the allowed time, they attempt to successfully complete either task as many times as possible. The two tasks have different difficulty and financial payoffs.

Multi-task Session Description

Participants answer easy or hard addition problems, similar to those used in Niederle and Vesterlund (2007) and Oswald et al. (2015). Easy questions require participants to add up three two-digit numbers, while hard questions require adding six two-digit numbers. There is no penalty for wrong answers and participants can end a round at any time. Participants are told they can quietly visit

other websites while they wait for the next section to begin, which is intended to simulate a leisure outside option. Participants can also leave the experiment if they finish early.

The activity is broken into three main sections – Fixed Wage, Random Coefficient, and Constant Coefficient – each consisting of a set of rounds. Sections have different payment schemes, round lengths, and number of rounds. During each round, participants can see their time remaining and how many questions of each type they have attempted. They do not see their results until the end of the round where a summary table displays all available information on questions attempted, time to completion, and payoff per question.

Part 1: Tutorial

To ensure that all participants understand the mechanics of the activity, they are led on an interactive guided tour. The tour individually highlights each element of the display and requires participants select questions and answer them for practice.

Part 2: Fixed Wage Session

As an introductory session, participants complete a Fixed Wage session. Here participants are told they will earn a fixed wage regardless of their performance. They are also told that, in this session, answering easy and hard questions will be valuable to the researchers and that hard questions are even more valuable. This is intended to provide non-monetary incentives to do well. This also is intended to imitate the vague notion that harder questions are more productive than easy questions.

Part 3: Random Coefficients and Constant Coefficients Sessions

In the final two sessions, participants are subject to two different incentive schemes that imitate employment contracts with and without production uncertainty. In order to compare an individual's performance between both contract types, all participants receive both treatments. The order in which the treatments are administered is randomized accross participants to account for the possibility that participants apply information gained in the first section to the second, a possibility called sequence effects.

In these rounds, participants are paid per successful completion of easy and hard questions. In addition, participants are paid a small amount for each minute remaining on the clock at the end of a round. This fixes a monetary value to leisure time and acts to assure participants that ending a round early is acceptable. The ability to visit other websites helps make leisure time less boring.

In the Constant Coefficients session, the value of easy and hard questions is displayed prominently at the top of the screen. This simulates an input-based incentive where the marginal payoff is known precisely. Participants should optimize their earnings by identifying which input has the highest marginal productivity and dedicating all their time to this task. Specifically, participants identify their earnings per minute for easy and hard tasks, and then dedicate all their time to the task with the highest earnings per minute.

In the Random Coefficients session, the payoff amount per question type is not known until the end of each round when it is randomly drawn. This simulates an output-based incentive where the marginal payoff of each input is known only after the output is measured. Because the distribution of payoffs remains constant across rounds, participants can optimally mix their inputs between easy and hard tasks as described in Equation 3.8.

3.4.2 Treatment Design

The treatment in this experiment is the variability of the payoff per minute of easy and hard questions, which simulates production uncertainty. The input that participants choose can best be thought of as the amount of time spent on easy and hard questions, x_1 and x_2 in Equation 3.8. The marginal benefit of each input, μ_1 and μ_2, is the payoff per minute of easy and hard questions. The payoff per minute depends both on how fast a participant is as well as the financial payoff set by the experiment.

To further understand how the treatment is designed, it is helpful to be precise about how the marginal payoff per minute of each input varies by individual and round. To that end, let $p_j + \gamma_{jk}$ be the random payoff for completing a question of difficulty $j \in (1, 2)$ in round k (remember that participants observe the price at the end of each round), where p_j is the average payoff and γ_{jk} is an error term. In the Control Session, $\gamma_{jk} = 0$ for all rounds. Importantly, there is variation between individuals in how long they need to successfully complete easy and hard questions. Let $r_{ij} + \varepsilon_{ijq}$ be the amount of time individual i needs to complete question q of difficulty type j. The term r_{ij} is the average amount of

time participant i needs to complete a question of type j, and ε_{ijq} is a mean-zero noise term capturing an individual's variation in the time required to complete questions of the same difficulty rating. Then the marginal payoff of answering question q of difficulty j for individual i is

$$\mu_{ijkq} = \frac{p_j + \gamma_{jk}}{r_{ij} + \varepsilon_{ijq}}. \tag{3.9}$$

For each individual, I can observe r_{ij} and estimate the average payoff: $\mu_{ij} = \frac{p_j}{r_j}$. By assuming that the distribution of ε_{ijq} remains unchanged for an individual between treatment and control, μ_{ij} remains unchanged. Furthermore, the variance of the marginal benefit, $Var(\mu_{ijq})$, only changes for an individual i if $Var(\gamma_{jk})$ changes. That is, $Var(r_{ij} + \varepsilon_{ijq})$ is the same in both treatment and control rounds, and so all changes to $Var(\mu_{ijq})$ come through changes to $Var(p_j + \gamma_{jk})$, providing my identification strategy. Referring back to Equation 3.8, I am changing σ_{11} and σ_{22} while holding everything else constant.

If the average marginal payoffs for time spent on different question types are roughly equal ($\mu_{i1} \approx \mu_{i2}$), the variation in question completion times makes participants mix their question types more. This is a straightforward prediction of the production uncertainty model. The numerator in Equation 3.8 is the net average payoff for each input. If the net payoff for x_1 is extremely high, participants should dedicate nearly all their time to easy questions. As the net payoffs become indistinguishable, participants should mix their inputs. Indeed, if $\sigma_{11} = \sigma_{22}$ and the net payoffs are equal, then participants should split their time evenly between easy and hard questions.

Designing treatment rounds requires choosing a payoff distribution, $p_j + \gamma_{jk}$. The goal is to set the average payouts p_1 and p_2 such that for some individuals, the average payoff per minute is greater for easy questions ($\mu_{i1} > \mu_{i2}$), while for other participants, the average payoff per minute is greater for hard questions ($\mu_{i1} < \mu_{i2}$). Otherwise all participants would essentially choose the same input. In other words, if easy questions are overwhelmingly more profitable, there will be few or no participants completing hard questions, and I would have insufficient information to determine if participants even know how to choose the right input when production uncertainty is minimal; my control group would provide little information. This is complicated by the fact that μ_{ij} is individual specific. While the identification of my treatment comes from changes in $Var(p_j + \gamma_{jk})$, I use different average payoffs for easy and hard questions, p_1 and p_2, between par-

ticipants in order to force a more representative sample. However, I always hold p_1 and p_2 constant between treatment and control for each individual participant.

As described before, participants complete a Control Session and a Treatment Session. Each session is composed of multiple rounds. A round is usually five minutes long, and in that time, participants complete as many easy or hard questions as they like. During the Treatment Session, participants learn the actual payoff at the end of each round. This means that, during the Treatment Session, the participant receives one draw of the payoff for each round. To be more precise, the payoff for a round k is $p_j + \gamma_{jk}$. For all rounds in a Control Session, the payoff is simply p_j. During the Treatment Session, the round payoffs, $p_j + \gamma_{jk}$, are drawn from a uniform distribution.[10] A uniform distribution has two distinct advantages. It allows me to avoid possibly negative payoffs and its parameters and functionality are easier for participants to understand.

The goal of this experiment is to test how changes in payoff variation induce inefficient behavior, all else constant. Equation 3.8 predicts that individuals will decrease an input as its payoff variance increases. One simple way to test this prediction would be to increase the payoff variance for one input while holding the other constant. In this experiment, the Treatment Session could consist of randomly drawing payoffs for one of the inputs, but having no randomness in the other. However, moving from no randomness in payoffs to randomness in only a single payoff may have unintended psychological effects. Subjects may try to infer what the experimenter wants them to do based on which input payoff was randomized. Instead, I test the basic hypothesis that increasing payoff uncertainty reduces efficiency by increasing the payoff variance for both inputs simultaneously, but vary the relative magnitude of those increases. In order to identify whether a greater increase in variance for an input leads to a greater decrease in its use, I have two treatment arms. In one, the easy questions have the a greater variance ("Easy High Variance"), and in the other, hard questions have the high variance ("Hard High Variance").

Table 3.2 lists the distribution parameters for treatment rounds by treatment arm. The variables a and b are the lower and upper bounds of the payoff distribution. Values are in cents. Note again that the average payoff for a subject, p_j, was unchanged between Treatment Session and Control Session, but p_j was different between subjects. However, the variance of payoffs in the Treatment Session is

[10] While the theoretical model derived earlier assumes uncertainty has a normal distribution, the same general conclusions are provable using a uniform distribution.

the same across all participants in a treatment arm. Given that the treatment is to change the *variance* of payoffs, this preserves the magnitude of the treatment across all participants in a treatment arm.

3.5 Empirical Results

The production uncertainty model predicts that as the payoff variance of an input increases, subjects will reduce their use of that input. If subjects shift their effort to a less effective input, they exhibit what I have called friction. If they simply reduce overall effort, they exhibit futility. In the experiment, subjects may manifest both friction and futility. Friction is shifting time away from the most effective question type to the less effective, whereas futility appears as spending more time on leisure (ending a round early) or reductions in question speed and success rate.

I observe each participant under no payoff variance (control) and under payoff variance (treatment), allowing me to hold the average payoff per minute for each question type constant between treatment and control sessions for an individual. While random assignment to treatment and control would make this unnecessary on average, having all participants complete both treatment and control sessions is a useful tool for utilizing my limited sample size. As a result, I can modify the payoff variance while holding everything else constant for each individual, isolating the effects of production uncertainty. However, the crucial assumption is that nothing systematically changes in a participant's average response time, r_{ij} in Equation 3.9, between the first and second sessions. On average, I overcome this possible identification threat by randomly assigning whether the Control Session is first or second. But given my limited sample size, I still provide evidence supporting the validity of my assumption that participants do not change their response rates between the first and second sessions, except in response to treatment.

There are five basic measures I use to evaluate subject behavior. The first is the payoff ratio κ_i, which is the ratio of the average payoff per minute of easy questions to the average payoff per minute of hard questions for an individual i. Using the notation from Equation 3.9, this is simply

$$\kappa_i = \frac{(price\ per\ easy) \times (easy\ questions/minute)}{(price\ per\ hard) \times (hard\ questions/minute)} = \frac{\mu_{i1}}{\mu_{i2}}. \tag{3.10}$$

If $\kappa_i > 1$, easy questions are that individual's most efficient input. The next measure is how much time, on average, a participant spends on easy questions as a fraction of total round time:

$$y_i = \frac{\textit{Time on Easy Questions}}{\textit{Time on Hard Questions} + \textit{Time on Easy Questions}} = \frac{x_1^*}{x_2^* + x_1^*} \tag{3.11}$$

In general, subjects for whom easy questions are the most efficient input should have more time spent on easy questions. That is, there should be a positive relationship between the payoff ratio κ_i and the percent of time spent on easy questions y_i. The second measure I use is how much time, as a percent, a participant spends on their most efficient input, which I call round efficiency. If hard questions are the most efficient input for a participant, then round efficiency is the percent time spent on hard questions. This allows me to compare the effects of production uncertainty across all participants, regardless of which input is the most efficient. To measure participant effort per question, I look at how long an average question takes to successfully complete and the question success rate. The question success rate is simply the number of correctly answered questions divided by the number of attempted questions of each type.

In what follows, I first provide a summary of subject behavior. This is useful for understanding the variation in participant skill and payoffs. I then examine the possibility that subjects systematically change their response times and question effectiveness from the first session to the second, showing that there are no sequencing effects. In the last two sections, I show that subjects do exhibit friction by shifting their time away from the most efficient input as the payoff variance increases, but there is little evidence of futility. It does not appear that participants increase their leisure time by ending rounds early as payoff variance increases, nor do they change their response time or question success rate.

3.5.1 Data Summary

In all, 28 participants completed at least a full Treatment Session or a full Control Session. The average total time for participants that completed both a treatment and control session was a little more than 1 hour and 15 minutes. The average earnings was $28.

Table 3.3 provides basic summary statistics of the average participant in terms of how many questions they can answer per round, how long an average question takes to complete, and how frequently participants answered questions

correctly. While hard questions required participants to add twice as many numbers as easy questions, completing a hard question took 2.34 times longer than completing an easy question for the average participant. Given the average payoff of a hard question is only twice that of an easy question, the average participant would maximize profit by completing only easy questions. There is, however, considerable variation in question response rates. The average participant took 11.3 seconds to complete an easy question, while some participants could successfully average 5.5 seconds per easy question. The range in times for completing hard questions was considerably larger, with some participants taking only 16.5 seconds and others using 42.9. The key takeaway is that there is sufficient variation in skill to ensure there are participants for whom hard questions are the optimal input. In fact, across all sessions, hard questions were the optimal choice for 8 of the 28 participants (29 percent).

3.5.2 Sequencing Effects

Participants may need to experience a few rounds to understand their own ability. To foster this learning, all participants first completed a round with fixed payoffs, regardless of performance. However, participants may still need additional experience under higher stakes before they understand their own ability. As a result, a participant may do poorly in the first few rounds when they are paid per question. Participants may also fatigue more by the end of their second session. The net result is that the sequence of sessions may lead to systematic differences unrelated to treatment status.

To circumvent sequencing bias, the sequence of treatment and control is randomized across participants. Even if there are systematic differences that depend on the order of sessions, these effects will be randomly distributed between treatment and control sessions. Regardless, I check for the existence of sequencing effects by comparing average session efficiency between the first session and the second. Table 3.4 shows the results of a permutation test in which I compare the average efficiency in the Control Session if it was first to the average efficiency in the Control Session if it was second. I find that subjects who completed their Control Session first had, on average, five percent less time dedicated to their efficient input. The permutation test provides a p-value of 0.73, which clearly fails to reject the hypothesis that there is no difference between having your Control Session first and having it second. Similarly, in Table 3.4, I find that participants who received their treatment round first spent two percent more round time on

their efficient input than participants who received treatment in their second session. With a p-value of 0.86, again there appears to be no meaningful difference between having the Treatment Session first or second.

It is valuable to observe the same individual in both treatment and control in order to check for potential differences in an individual's effort when production uncertainty is introduced, but large sequencing effects could be problematic with such a small sample size. On average, because the sequence of treatment and control is random, there is no identification threat. Perhaps most importantly, ruling out sequencing effects allows me to assume that each session is an independent draw, even though each participant completed two sessions. This significantly improves the experiment's power.

3.5.3 Testing for Friction

Friction, which is described in Proposition 2, predicts production uncertainty induces inefficient deviations from the optimal allocation of inputs. The key question, then, is how do participants change their input allocation? To understand the optimal allocation of inputs in terms of average productivity, consider the scenario in which there is no production uncertainty. If there is no production uncertainty, the covariance matrix Σ_γ is the zero matrix, and the inputs for the utility function in Equation 3.5 have constant marginal returns. The optimal choice of inputs for a participant is to dedicate all time to the single input with the highest net marginal return. Figure 3.1 depicts the optimal time allocation for participants as a function of their payoff ratio. On the x-axis is the payoff ratio κ and on the y-axis is the fraction of time spent on easy questions. If the ratio is less than one $\kappa < 1$, the participant should spend all his time on hard questions. On the other hand, if the ratio is greater than one $\kappa > 1$, all his time should be spent on easy questions.

Proposition 2 shows that with production uncertainty, participants should immediately begin to mix their inputs, decreasing their average productivity. To see how this would effect a subject's time allocation as a function of κ, based on Equation 3.8 the fraction of time spent on an easy question y is

$$y = \frac{x_1^*}{x_1^* + x_2^*} = \frac{\frac{\mu_1}{\sigma_{11}}}{\frac{\mu_1}{\sigma_{11}} + \frac{\mu_2}{\sigma_{22}}} = \frac{\kappa}{\kappa + \frac{\sigma_{11}}{\sigma_{22}}}, \tag{3.12}$$

where I have assumed leisure time has zero payoff and there are no meaningful cost differences between time spent on easy and hard questions. As κ increases,

a participant should spend more time on easy questions, but the rate at which he switches to easy questions depends on the ratio of the payoff variances, $\frac{\sigma_{11}}{\sigma_{22}}$. Using κ removes the need to consider differences in participant risk aversion and reduces participant responses to a question of payoff ratios and the ratio of payoff variances.

To visualize how participants should respond as $\frac{\sigma_{11}}{\sigma_{22}}$ changes, Figure 3.2 charts the hypothetical relationship for different levels of the variance ratio. The bounds on the payoff ratio κ were chosen to reflect the bounds observed in the data. Figure 3.2 demonstrates graphically the two aspects of friction that I will examine empirically: flattening out participant responses to κ and a downward shift as easy question payoff variance increases relative to hard questions.

Compared to the maximally efficient time allocation in Figure 3.1, participant input allocation is considerably flatter with production uncertainty, demonstrating that the utility-maximizing choice is to mix inputs. Mixing inputs should reduce a participant's average productivity. Notice also that as the variance ratio deviates from $\frac{\sigma_{11}}{\sigma_{22}} = 1$, the relationship flattens out even more. As the variance of the payoff for one input overwhelms the other, participants will choose the low-variance input always (except at extreme values of κ not graphed).

Another key takeaway from Figure 3.2 is that as easy question payoff variance σ_{11} increases relative to hard questions σ_{22}, the curve shifts downward. If easy question payoff variance increases relative to hard question payoff variance, participants will avoid easy questions even when, on average, they may be more profitable per minute.

Graphical Evidence of Friction

The Control Session minimizes a participant's uncertainty about the payoff per minute of each question type by fixing question prices ahead of time. Figure 3.3 plots participant time spent on easy questions during the Control Session by their payoff ratio. Ideally, participant behavior would match Figure 3.1.

The line drawn is a probit regression of the time spent on easy questions as a function of the the payoff ratio. While participants would ideally allocate their time as in Figure 3.1, they are prone to make errors. The probit regression reflects the probability of picking the right input given a payoff ratio κ. The specification is to estimate the probability that an individual i chooses an easy question, which is

$$Pr(\text{Easy Question})_i = \Phi(\beta_0 + \beta_1 \kappa_i)\,, \tag{3.13}$$

where $\Phi(\cdot)$ is the standard normal distribution function. Based on the result, participants clearly make errors in their time allocation. However, in general, individuals with particularly high κ are much more likely to allocate their time entirely to easy questions.

In calculating participant time allocation, I have included *all* rounds within the Control Session. Alternatively, I could drop the first few rounds of the Control Session since participants may still be learning which input is most productive given their payoffs. By keeping all rounds, I have essentially allowed the learning process to be factored into the overall participant error in their allocation choices.

In the Treatment Session, participants experience a considerable amount of production uncertainty. Based on the hypothetical responses in Figure 3.2, there are two key changes in participant response predicted by the model. First, participants should have a much flatter response curve than during the Control Session. Second, the participant response curve should shift down during the treatment arm in which easy questions have the higher payoff variance relative to the other treatment arm.

Figure 3.4 plots participant time spent on easy questions by payoff ratio for the two treatment arms. The line represents a regression of the log of payoff ratio on the fraction of time spent on easy questions. Using the log allows for a concave response curve. Because the anticipated response is not expected to be binary, I have not used a probit regression. However, both probit and linear regressions yield qualitatively similar results.

There are two key features of the participant reponse curve during the Treatment Session in Figure 3.4. The first is that the fitted curve is much flatter than the response curve during the Control Session in Figure 3.3. Intuitively, increasing the payoff uncertainty for all question types should induce participants to mix their inputs more than when there is very little uncertainty. Instead of simply making errors about the efficient input, participants are now managing their risk by mixing between inputs, matching the utility-optimizing allocation depicted in Figure 3.2 fairly closely on average.

The second important feature of Figure 3.4 is the general shift towards hard questions as the easy question payoff variance increases, represented as the shift from the solid red line to the dotted blue line. On average, subjects in the "Easy High Variance" treatment reduced their fraction of time spent on easy questions by 0.09. The p-value of the permutation test was 0.38 (see Table 3.7), making

the difference not statistically significant, which is not surprising given the small sample size.

Quantitative Evidence

The two ways in which Friction will induce inefficiency are, first, to make participants less responsive to differences in the marginal benefit of each input (flattening), and second, to induce shifts towards the low-variance input regardless of its marginal benefit. Having presented some graphical evidence of these two effects, I now present quantitative evidence.

One way to quantitatively test how much the response curve flattens between control and treatment is to use a linear regression on participant responses that allows for a slope change between treatment and control. The specification for individual i is

$$y_i = \beta_0 + \beta_1 \kappa_i + \beta_2 D_i + \beta_{12} \kappa_i \times D_i + \varepsilon_i \tag{3.14}$$

where D_{is} is a dummy for treatment status in session s. The coefficient of interest is β_{12}, which measures the change in slope from the Control Session to the Treatment Session.

Table 3.5 shows the results of estimating Equation 3.14 as Model (1). Model (2) adds an additional indicator variable to distinguish between the two treatment arms. The first row of Table 3.5 shows the strong positive relationship between payoff ratio κ and the fraction of time spent on easy questions during the Control Session. Row two shows there is no statistically significant shift in the average fraction of time spent on easy questions for the Treatment Session. The coefficient of interest is in the third row, which shows a significant reduction in the slope during treatment rounds, which is statistically significant and negative. Treatment reduces the slope by nearly 0.4. In other words, during the Control Session, increasing the payoff ratio by 0.10 would mean a participant increases the fraction of his time spent on easy questions by 0.055. A similar increase in the payoff ratio in the Treatment Session leads to only an increase of roughly 0.015. This result confirms the basic prediction that participants will become less responsive to the differences in the average payoff between the two inputs after adding production uncertainty.

Another way to quantify the general flattening of participant responses to the payoff ratio is to consider how much time they spend on their efficient input.

Without production uncertainty, participants should spend a greater proportion of their time on the efficient input (as in Figure 3.3). The flattening out of the response curve is a general reduction in efficiency, since subjects with $\kappa < 1$ will inefficiently spend more time on easy questions during the Treatment Session, and subjects with $\kappa > 1$ will inefficiently spend less time on easy questions.

In Table 3.6, I evaluate how participants reduce efficiency under production uncertainty. In the first row, I compare the average fraction of time spent on the efficient input during Treatment and Control Sessions. As shown, I find that production uncertainty significantly reduces allocation efficiency. Participants in the Treatment Session spend a full 21.5 percent less time on their efficient input. Using a permutation test, I find that the p-value is 0.007, making the difference statistically significant. In row two, I attempt to isolate how individuals with a payoff ratio greater than one $\kappa > 1$ reduce their time spent on easy questions. For this subsample, I find that they reduce the fraction of time spent on easy questions by 0.20, which is statistically significant. Visually, this represents how participants to the right of $\kappa = 1$ all reduce their use of easy questions by 20 percent, on average. Similarly, participants with a payoff ratio less than one $\kappa < 1$ inefficiently increase their use of easy questions by an average of 17 percent, which is also statistically significant.

One concern with randomizing question payoffs as I have in the experiment is that it may simply induce random behavior. In this case, the results I have previously quantified may be the result of participants guessing which input to use. To counter this notion, production uncertainty uniquely predicts a reduction in the use of easy questions as their payoff variance increases relative to the payoff variance of hard questions, all else constant. This result is represented by the downward shift of the response curve illustrated in the hypothetical case in Figure 3.2.

To quantitatively test the significance of the downward shift observed between treatment arms in Figure 3.4, I first compare the average fraction of time spent on easy questions between the two treatment arms. Table 3.7 shows that, under the treatment arm with high easy payoff variance, participants reduce the fraction of time spent on easy questions by 0.09. The difference, however, is not statistically significant in a permutation test.

Table 3.7 also breaks out a comparison of treatment arms by whether a participant is above or below a payoff ratio of one. Looking only at participants *for whom easy questions are the most efficient*, simply increasing the variance of easy

question payoffs reduced their use of easy questions by 10 percent. However, for participants with a payoff ratio less than one, the results are inconclusive. Given there were only six participants with payoff ratio below one in the Treatment Sessions, this is unsurprising.

The results shown in Table 3.7 do not conclusively show the predicted downward shift, but the coefficients have the right sign. However, more observations are needed to identify whether there really is a downward shift. Given the errors even in the control group, the subtle downward shift from increasing the relative payoff variance of easy questions will be difficult to measure without increasing the experiment's power.

In all, there is clear evidence that introducing production uncertainty significantly reduced participant efficiency. The evidence on the effect of increasing payoff variance without changing the payoff means is less clear. Even still, the results are generally consistent with the predictions of the production uncertainty model. Among participants whose optimal input is easy questions, there appears to be a reduction in time spent on easy questions as the payoff variance of easy questions increases. These results underscore the need to conduct additional rounds of the experiment with some modifications to the parameters to ensure more participants whose optimal input is hard questions.

3.5.4 Testing for Futility

While friction addresses the allocative inefficiencies induced by production uncertainty, futility addresses reductions in effort. Futility can be measured by evaluating whether or not participants reduced their effort under treatment. There are three ways I test to see if participants reduce their effort during the Treatment Session. The first is to test if participants reduce their question completion speed, and the second is to test if they reduce their question success rate. These two measures look at how participants reduce their overall effort along the intensive margin. The third measure is to look at how participants reduce their effort along the extensive margin by ending their rounds early.

Measuring Futility along Intensive Margin

Figures 3.5a through 3.5d plot effort measures for each participant's control rounds against their treatment rounds with a 45 degree line included. For both easy and hard questions, I plot participant effort between Treatment and Control Sessions

as measured by question success rate and question completion time. If participants systematically reduced their effort intensity during the Treatment Session, there would be a disproportionate number of observations below the 45 degree line on each panel shown.

Figures 3.5a and 3.5b show participant success rates. In theory, participants may reduce the care with which they answer questions, being discouraged by the uncertainty during the Treatment Session. However, the graphical evidence to that effect is underwhelming, and a statistical test comparing success rates confirms there is no statistically significant difference. Figures 3.5c and 3.5d plot the average time to complete easy and hard questions for each participant. Participants may dedicate less attention to answering questions because of the induced production uncertainty, and would then reduce their response times during the Treatment Session. But again, the graphical evidence does not confirm this hypothesis, and a quantitative analysis comparing question completion times between Treatment and Control Sessions finds no statistically significant difference.

Measuring Futility along Extensive Margin

Futility could also manifest as reductions in participant willingness to complete rounds. Participants may favor other uses of their time and try to leave early. They may also prefer to browse the Internet. The simplest measure of the extensive margin of effort is to evaluate whether or not participants leave rounds earlier during the Treatment Session. I find no statistically significant or empirically meaningful difference between Treatment and Control Sessions. However, this is not entirely surprising. While the experiment script was careful to make it clear that participants could end rounds early with no penalty whatsoever, and that doing so would not negatively affect the experiment, only 6 percent of participants ended a session with more than three minutes remaining. The vast majority of participants used up all their time, suggesting that either the payoff was sufficiently high to overcome potential futility effects, or participants had an additional sense of obligation to the experimenter.

3.6 Conclusion

The ideas contained within the production uncertainty model create a lens through which policy makers and researchers can understand the sometimes confusing results of teacher performance incentives. Because teaching is a complex production

process, the uncertainties inherent in the production function create an environment in which output-based incentives are likely to have little or no effect. Even worse, production uncertainty demonstrates how a test-based incentive may have zero or even negative effects on test scores, without requiring teachers to reduce their effort.

This paper demonstrates the results of a laboratory experiment that support the basic predicted behavioral responses to an uncertain production function. The results illustrate that while agents make some errors allocating their effort when there is no induced production uncertainty, on average they favor the most efficient input. Under production uncertainty, participants mix their inputs, as predicted by the production uncertainty model. While this response is utility optimizing for risk-averse agents, it is considerably less efficient on average. Furthermore, as the variance of an input increases, participants shift away from that input even though its average payoff remains unchanged, though the shift is not statistically significant.

The experiment uses the same payment scheme between treatment and control sessions but introduces simulated production uncertainty during treatment. In reality, when designing an incentive, policy makers are not deciding whether or not to make the payoff uncertain. Instead, the analogy from the experiment to actual policy is to consider the Treatment Session an outcome-based incentive like teacher value-added, and then consider the Control Session an input-based incentive, such as in-class teacher observations. In this analogy, policy makers do not need to have a better understanding of the teaching production function – participants know the average payoff during treatment rounds since they know the distribution of payoffs. The inefficiency during the Treatment Session is not from a lack of knowledge, but rather the result of risk management on the participant's part. Similarly, teachers may introduce inefficient behavior, even as measured by test outcomes, if their incentive is based on student test scores.

One important concern with this experimental approach is how quickly participants learn their own production function. If participants need a lot of time to learn how quickly and accurately they can answer easy questions relative to hard questions, it becomes considerably more difficult to do this when the payoffs are randomized as well. This could make participants behave randomly during the Treatment Session. I have tried to address this in several ways. First, by randomly assigning the order of treatment and control, I am able to see how much participants are learning about their own ability in the first session, regardless

of treatment status. I find no meaningful difference in subject behavior during the Control Session by whether it was first or second, and similarly for the Treatment Session. Second, I have used all the rounds within a session to evaluate participant behavior, which effectively incorporates all the error associated with learning during the Control Session. As a result, the Control Session creates a counterfactual that incorporates the allocative errors participants make as they work through their rounds. Finally, I have also identified a unique theoretical prediction of my production uncertainty model that does not depend on comparing between treatment and control, but rather compares between two different treatment arms. Unfortunately, the test is underpowered, but preliminary results are consistent with the prediction that increasing the easy question payoff variance relative to hard questions will induce participants to dedicate less time to easy questions. Another way to address this problem in future experiments is to allow participants longer sessions.

In future experiments, there are three important improvements to make. First, I need to make hard questions have a relatively higher average payoff for a larger portion of participants. Second, I need to experiment with different methods of providing better performance information for participants. While participants receive detailed information at the end of each round in a session, it is not clear that they are using this information effectively to make their allocation choices, even during the Control Session. Finally, I need to make the relative variance of easy question payoffs more dramatic between treatment arms.

Production function uncertainty has several future extensions and generalizable results that I plan to explore in future work. For example, expanding the model to allow teachers to update their knowledge of the production function predicts an additional futility effect if the variance of the prior distribution does not shrink sufficiently. It can also be shown that as the number of possible inputs increases, an agent's overall effort will decrease in a form of analysis paralysis. Taken to infinity, this describes an infinite inputs dilemma in which agents become paralyzed by too many options, requiring some form of rational inattention. There are also other applications in consumer choice theory that can be used to explain unexpected consumer behavior when choosing health insurance plans and pension plans.

Table 3.1: Teacher performance incentives in the US by effectiveness and use of in-class evaluations.

Differentiated In-Class Evaluations	**Improved Student Scores**		
	Yes	Mixed	No
Yes	3	0	0
Some	0	2	0
No	0	0	2

Notes - There is an apparent correlation between effective teacher incentive programs and the use of in-class evaluations. This is not a comprehensive review of teacher incentive programs in the US. The programs in this table were selected based on their emphasis on individual-level measures of teacher performance. Many other programs use large school-level bonuses, which I have excluded. The analyses in the Yes/Yes cell are Dee and Wyckoff (2015); Dee and Keys (2004); Hudson (2010); in the Mixed/Some cell are Sojourner et al. (2014); Wellington et al. (2016); in the No/No cell are Briggs et al. (2014); Springer (2010).

Table 3.2: Experiment parameters for two treatment arms.

	a	b	Mean	Variance
Easy High Variance				
Easy Payoff	[1, 2]	[11, 12]	[6, 7]	8.3
Hard Payoff	[10.5, 11.5]	[15.5, 16.5]	[13, 14]	2.1
Hard High Variance				
Easy Payoff	[3.5, 4.5]	[8.5, 9.5]	[6, 7]	2.1
Hard Payoff	[8, 9]	[18, 19]	[13, 14]	8.3

Notes - For each round in a Treatment Session, question payoffs are randomly drawn from a uniform distribution with start and end values, a and b, as shown. Within a treatment arm, the distance between a and b is always preserved, guaranteeing each subject in a treatment arm had an identical variance in payoffs. Changing payoff means induces additional variation between participants, but does not change the treatment parameter, which is payoff variance. The mean payoff for the Treatment Session was always the same as the mean payoff for a participants's Control Session.

Table 3.3: Summary statistics of subject performance measures.

	Mean	Median	St. Dev.	Min	Max
Correct Easy	8.00	6.84	5.46	1.00	26.07
Correct Hard	4.25	3.74	2.02	1.33	8.74
Seconds Per Easy	11.28	10.89	3.34	5.55	19.51
Seconds Per Hard	26.44	27.34	6.90	16.49	42.91
Success Rate - Easy	0.88	0.89	0.08	0.67	1.00
Success Rate - Hard	0.86	0.86	0.09	0.70	1.00

Notes - The averages shown are calculated at the individual level. The question completion times shown are the average time in seconds required to *correctly* answer a question. There was a considerably large range of average completion times for hard questions, creating variation between participants as to which input was most profitable.

Table 3.4: Efficiency differences by session order.

	$E[Y\|\text{First}] - E[Y\|\text{Second}]$	P-value	N
Control	-0.0501	0.73	42
Treatment	0.0216	0.86	42

Notes - P-values are calculated using a permutation test. The differences shown are the average efficiency difference between subjects who received the Control Session first and those who received it second. Similarly for the treatment row. Efficiency is the fraction of time a participant spends on their most profitable input. There are no observable differences based on session order.

Table 3.5: Correlation between percent time spent on easy questions and Easy/Hard payoff ratio.

	Fraction of Time Spent on Easy Questions	
	(1)	(2)
Payoff Ratio	0.549^{***}	0.549^{***}
	(0.121)	(0.122)
Treatment Round	0.393	0.406
	(0.256)	(0.260)
Payoff $\times$ Treatment Round	-0.361^{*}	-0.395^{*}
	(0.201)	(0.213)
Include Treatment Arm Indicator		X
N	54	54

Notes - Coefficients are from estimating the regression in Equation 3.14. The relationship between the Easy/Hard question payoff ratio and the amount of time spent on easy questions is strong in the control group (row one). During the treatment, participants are more inefficient and the relationship becomes much weaker (row three), which is demonstrated by the flat line in Figure 3.4. When an indicator is added to allow for differences between treatment arms, the slope change between treatment and control is even greater.

Table 3.6: Nonparametric tests for inefficient time allocation under production uncertainty.

	Difference	P-value	N
$E[Y\|\text{Treatment}] - E[Y\|\text{Control}]$			
% Time on Efficient Input	-0.215	0.007	54
% Time on Easy $\mid \kappa > 1$	-0.201	0.023	38
% Time on Easy $\mid \kappa < 1$	0.168	0.068	16

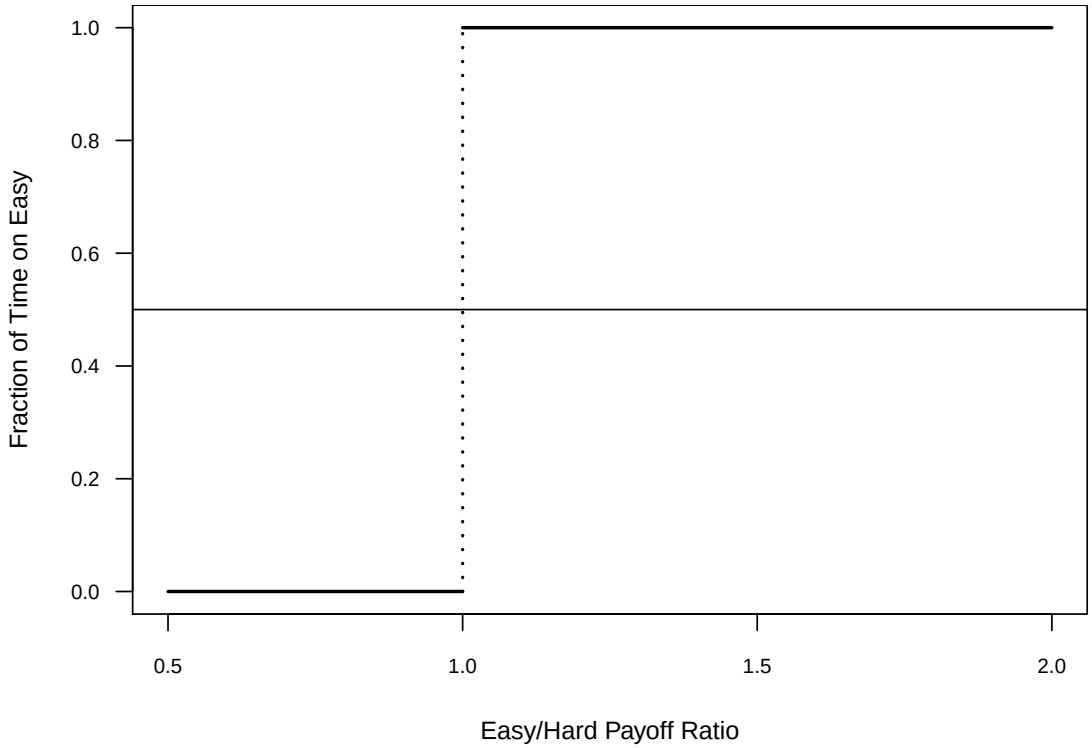
Notes - The p-values shown are generated using permutation tests. Participants significantly reduced their efficiency during the Treatment Session (row one). Participants for whom easy questions were the most profitable ($\kappa > 1$) inefficiently reduce their time on easy questions. Similarly, participants for whom hard questions were the most profitable ($\kappa < 1$) inefficiently increased the time spent on easy questions.

Table 3.7: Nonparametric tests for reduction in time spent on easy questions due to increased relative payoff variance between treatment arms.

	Difference	P-value	N
$E[Y\|\text{Easy Variance High}] - E[Y\|\text{Hard Variance High}]$			
% Time on Easy Input	-0.091	0.384	28
% Time on Easy $\mid \kappa > 1$	-0.105	0.204	22
% Time on Easy $\mid \kappa < 1$	0.010	0.533	6

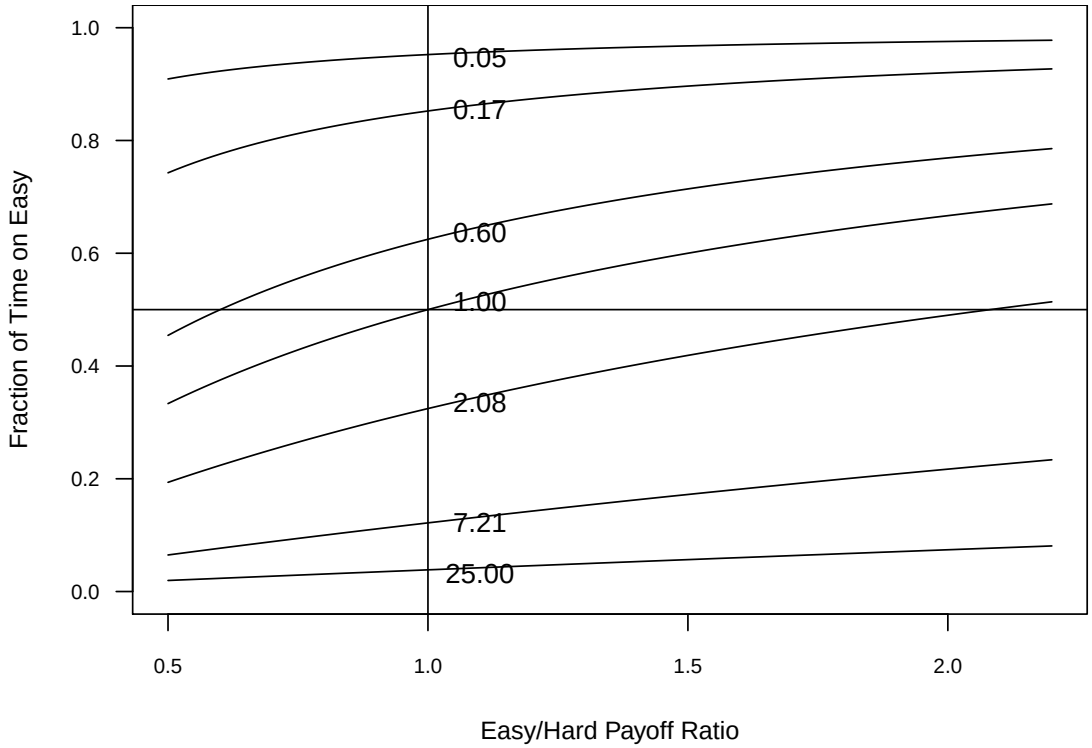
Notes - P-values shown are generated using permutation tests. The theoretical model predicts that the "Easy Variance High" treatment arm should induce less time spent on easy questions for all participants (first row). However, the effect was strongest among participants for whom easy questions *were the most profitable*. Sample sizes for comparing between treatment arms are greatly reduced in large part because individual ability largely determines if the payoff ratio κ is greater or less than one.

Figure 3.1: Hypothetical subject time allocation if there is no uncertainty in the payoff per minute of each input.



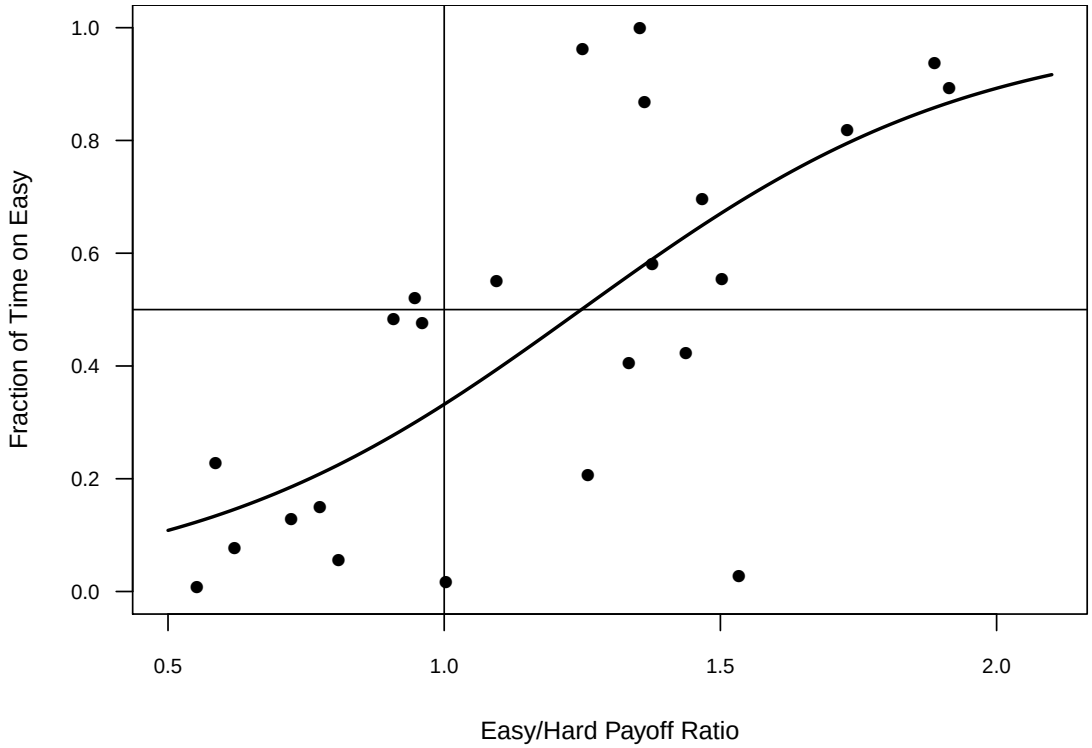
Notes - The payoff ratio κ is calculated as the ratio of $(\textit{easy price}) \times (\textit{easy questions/minute})$ to $(\textit{hard price}) \times (\textit{hard questions/minute})$. In the hypothetical case in which there is no uncertainty about the payoff per minute, participants with $\kappa < 1$ would dedicate all their time to hard questions, and participants with $\kappa > 1$ would dedicate all their time to easy questions.

Figure 3.2: Hypothetical subject time allocation with production uncertainty by payoff ratio κ across different variance ratios $\frac{\sigma_{11}}{\sigma_{22}}$.



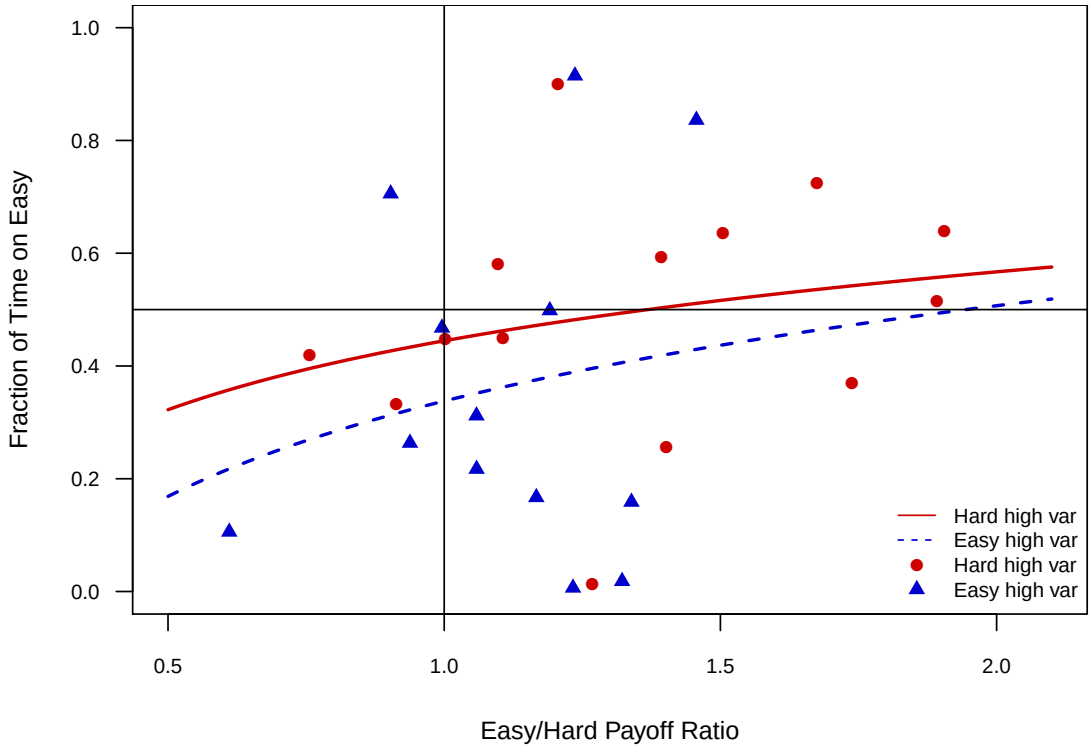
Notes - The lines shown are the utility-maximizing values of time allocation as predicted in Equation 3.8, not the efficiency maximizing allocation (which are represented in Figure 3.1). The values printed under each line are the payoff variance ratios, $\frac{\sigma_{11}}{\sigma_{22}}$. The Easy/Hard payoff ratio κ is calculated as the ratio of $(\textit{easy price}) \times (\textit{easy questions/minute})$ to $(\textit{hard price}) \times (\textit{hard questions/minute})$. The fraction of time spent on easy questions as a function of κ is $y = \frac{\kappa}{\kappa + \frac{\sigma_{11}}{\sigma_{22}}}$, from Equation 3.8.

Figure 3.3: Subject time allocation in Control Session by Easy/Hard payoff ratio κ.



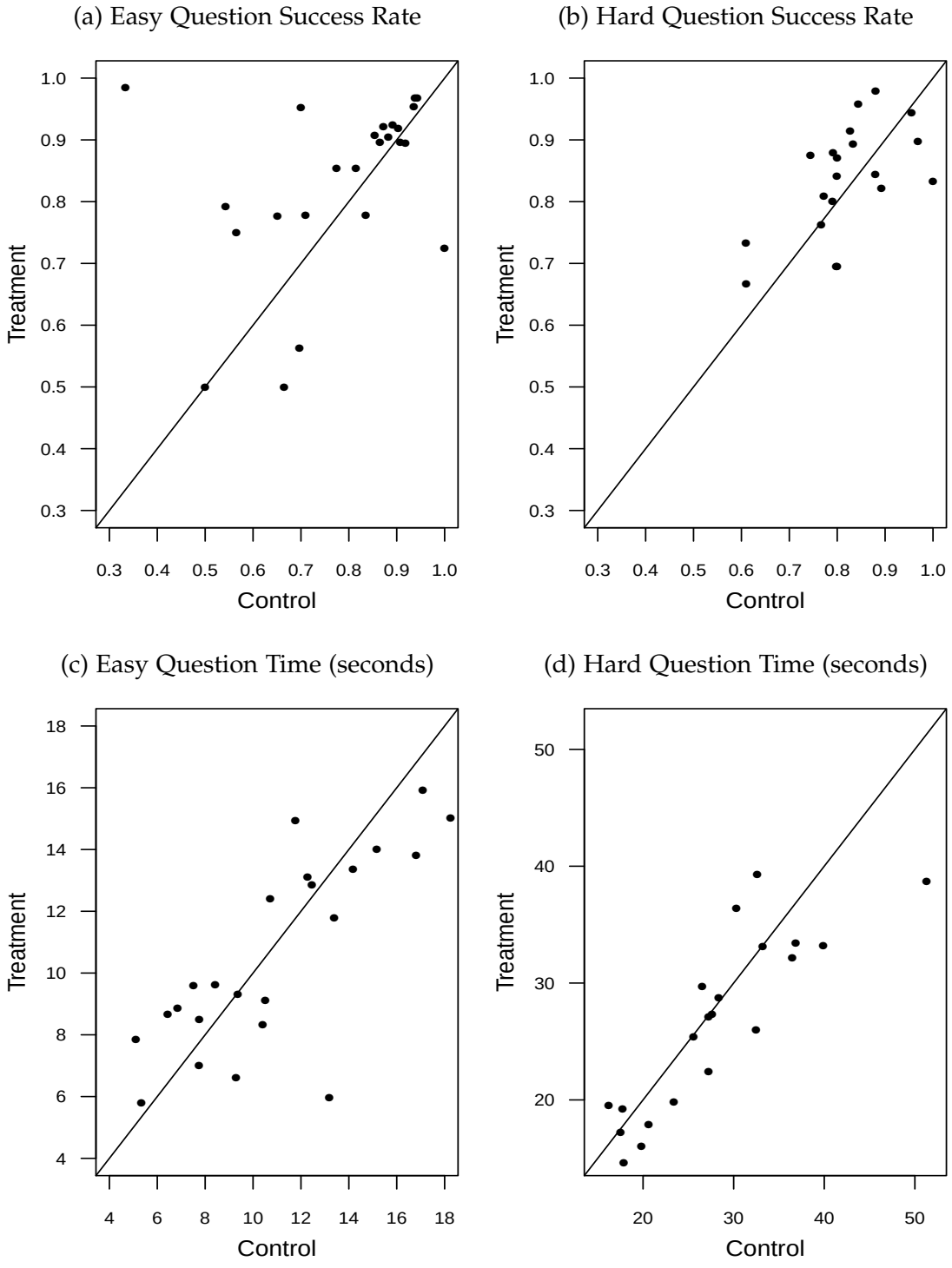
Notes - As the payoff ratio increases, subjects should dedicate all their time to easy questions. However, individuals make errors in their time allocation, which are shown as deviations from the optimal allocation depicted in Figure 3.1. The payoff ratio is calculated as the ratio of $(\textit{easy price}) \times (\textit{easy questions/minute})$ to $(\textit{hard price}) \times (\textit{hard questions/minute})$

Figure 3.4: Subject time allocation in Treatment Session by Easy/Hard payoff ratio κ.



Notes - As the payoff ratio increases, subjects should dedicate all their time to easy questions. In contrast to Figure 3.3, subjects were far less efficient under production uncertainty. Figure 3.1 shows the optimal time allocation from an efficiency perspective, while Figure 3.2 shows the utility maximizing time allocation. A key feature is that the relationship has a mean shift downward when easy questions have higher payoff variance, even without changing the average payoff. Even among participants for whom easy questions are the most efficient input, increasing the variance of easy questions induces inefficient time allocation towards hard questions. The payoff ratio is calculated as the ratio of $(\textit{easy price}) \times (\textit{easy questions/minute})$ to $(\textit{hard price}) \times (\textit{hard questions/minute})$

Figure 3.5: Comparisons of subject effort between control and treatment rounds.



Notes - Question success rate is the number of correctly answered questions divided by the number of attempted questions. Question time is the average time to *successfully* complete a question. Participants do not appear to have any changes in their question success rates or question completion times between control and treatment rounds.

Bibliography

Aaronson, D., L. Barrow, and W. Sander (2007): "Teachers and Student Achievement in the Chicago Public High Schools," *Journal of Labor Economics*, 25, 95–135.

Adnot, M. (2016): "Teacher Evaluation, Instructional Practice and Student Achievement: Evidence from the District of Columbia Public Schools and the Measures of Effective Teaching Project, PhD Thesis," *University of Virginia*.

Adnot, M., T. Dee, V. Katz, and J. Wyckoff (2017): "Teacher Turnover, Teacher Quality, and Student Achievement in DCPS," *Educational Evaluation and Policy Analysis*, 39, 54–76.

Akerlof, G. A. and R. E. Kranton (2005): "Identity and the Economics of Organizations," *Journal of Economic Perspectives*, 19, 9–32.

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

Bandiera, O., I. Barankay, and I. Rasul (2007): "Incentives for managers and inequality among workers: Evidence from a firm-level experiment," *Quarterly Journal of Economics*, 729–773.

Benabou, R. (2016): "Bonus Culture: Competitive Pay, Screening, and Multitasking," *The Journal of Political Economy*, 124, 305–370.

Blazar, D. (2015): "Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement," *Economics of Education Review*, 48, 16–29.

Blazar, D. and M. A. Kraft (2016): "Teacher and Teaching Effects on Students Attitudes and Behaviors," *Educational Evaluation and Policy Analysis*, 39.

Blazar, D. and C. Pollard (2017): "Does Test Preparation Mean Low-Quality Instruction?" *Educational Researcher*, 20, 1–14.

Briggs, D., E. DiazBbilello, A. Maul, M. Turner, and C. Bibilos (2014): "Denver ProComp Evaluation Report: 2010-2012," Tech. rep., Colorado Assessment Design Research and Evaluation Center, University of Colorado, Boulder.

Chetty, R., J. Friedman, and J. Rockoff (2014): "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 104, 2633–2679.

Chetty, R., A. Looney, and K. Kroft (2009): "Salience and Taxation: Theory and Evidence," *American Economic Review*, 99, 1145–1177.

Cohen, J. and D. Goldhaber (2016): "Observations on evaluating teacher performance," *Improving teacher evaluation systems: Making the most of multiple measures*, 8–21.

Cohen, J. and P. Grossman (2016): "Respecting complexity in measures of teaching: Keeping students and schools in focus," *Teaching and Teacher Education*, 55, 308–317.

Cronbach, L. J. (1972): "Theory of generalizability for scores and profiles," in *The dependability of behavioral measurements*, ed. by L. J. Cronbach, G. C. Gleser, H. Nanda, and N. Rajaratnam, New York: John Wiley and Sons, 161–188, first ed.

Dee, T. S. and B. J. Keys (2004): "Does merit pay reward good teachers? Evidence from a randomized experiment," *Journal of Policy Analysis and Management*, 23, 471–488.

Dee, T. S. and J. Wyckoff (2015): "Incentives, Selection, and Teacher Performance: Evidence from IMPACT," *Journal of Policy Analysis and Management*, 34, 1–31.

Dixit, A. (2002): "Incentives and Organizations in the Public Sector: An Interpretative Review," *Journal of Human Resources*, 37, 696–727.

Dobbie, W. and R. G. Fryer (2013): "Getting Beneath the Veil of Effective Schools : Evidence From New York City," *American Economic Journal: Applied Economics*, 5, 28–60.

Donaldson, M. L. and J. P. Papay (2015): "An Idea Whose Time Had Come: Negotiating Teacher Evaluation Reform in New Haven, Connecticut," *American Journal of Education*, 122, 39–70.

Fehr, E. and K. M. Schmidt (2004): "Fairness and incentives in a multi-task principal-agent model," *Scandinavian Journal of Economics*, 106, 453–474.

Flodgren, G., M. P. Eccles, S. Shepperd, A. Scott, E. Parmelli, and F. R. Beyer (2011): "An overview of reviews evaluating the effectiveness of financial incentives in changing healthcare professional behaviours and patient outcomes," *The Cochrane Library*.

Francois, P. (2000): "Public service motivation as an argument for government provision," *Journal of Public Economics*, 78, 275–299.

Fryer, R. G. (2013): "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools," *Journal of Labor Economics*, 31, 373–407.

Gibbs, M., K. Merchant, W. Van der Steded, and M. E. Vargus (2004): "Determinants and Effects of Subjectivity in Incentives," *The Accounting Review*, 79, 409–436.

Glazerman, S. and A. Seifullah (2012): "An Evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after Four Years," Tech. rep., Mathematica Policy Research, Inc.

Goodman, J. (2015): "In Defense of Snow Days," *Education Next*, 15.

Grissom, J. A. and S. Loeb (2017): "Assessing Principals' Assessments: Subjective Evaluations of Teacher Effectiveness in Low- and High-Stakes Environments," *Education Finance and Policy*, forthcoming.

Hamre, B. K. and R. C. Pianta (2001): "Early teacher–child relationships and the trajectory of children's school outcomes through eighth grade," *Child development*, 72, 625–638.

Hanushek, E. (1992): "The Trade-Off between Child Quality and Quantity," *The Journal of Political Economy*, 100, 84–117.

——— (2007): "The Single Salary Schedule and Other Issues of Teacher Pay," *Peabody Journal of Education*, 82, 574–586.

Hanushek, E. and S. G. Rivkin (2010): "Generalizations about using value-added measures of teacher quality," *American Economic Review*, 100, 267–271.

Hill, C. J., H. S. Bloom, A. R. Black, and M. W. Lipsey (2008): "Empirical Benchmarks for Interpretting Effect Sizes in Research," *Child Development Perspectives*, 2, 172–177.

Hill, H. C., D. Blazar, and K. Lynch (2015): "Resources for Teaching," *AERA Open*, 1.

Hill, H. C. and P. Grossman (2013): "Learning from Teacher Observations: Challenges and Opportunities Posed by New Teacher Evaluation Systems," *Harvard Educational Review*, 83, 371–384.

Ho, A. D. and T. J. Kane (2013): "The Reliability of Classroom Observations by School Personnel. Research Paper. MET Project." *Bill & Melinda Gates Foundation*.

Holmstrom, B. (1982): "Moral hazard in teams," *The Bell Journal of Economics*, 324–340.

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

Hoppe, E. I. and D. J. Kusterer (2011): "Conflicting tasks and moral hazard: Theory and experimental evidence," *European Economic Review*, 55, 1094–1108.

Hoxby, C. M. and A. Leigh (2004): "Pulled away or pushed out? Explaining the decline of teacher aptitude in the United States," *American Economic Review*, 94, 236–240.

Hudson, S. (2010): "The effects of performance-based teacher pay on student achievement," *SIEPR Discussion Papers*.

Imberman, S. A. and M. F. Lovenheim (2015): "Incentive Strength and Teacher Productivity: Evidence from a Group-Based Teacher Incentive Pay System," *Review of Economics and Statistics*, 97.

Jacob, B. A. and S. D. Levitt (2003): "Catching cheating teachers: The results of an unusual experiment in implementing theory," Tech. rep.

Kandel, E. and E. P. Lazear (1992): "Peer pressure and partnerships," *Journal of political Economy*, 100, 801–817.

Kane, T. and D. Staiger (2008): "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," *National Bureau of Economic Research*, 51.

Kane, T. J., D. F. McCaffrey, T. Miller, and D. O. Staiger (2013): "Have we identified effective teachers? Validating measures of effective teaching using random assignment," Tech. rep., MET Project, Seattle.

Kane, T. J. and D. O. Staiger (2012): "Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains," Tech. rep., MET Project, Seattle.

Kane, T. J., E. S. Taylor, J. H. Tyler, and A. L. Wooten (2010): "Identifying Effective Classroom Practices Using Student Achievement Data," *The Journal of Human Resources*, 6, 587–615.

Koretz, D. and S. Barron (1998): "The Validity of Gains in Scores on the Kentucky Instructional Results Information System," Tech. rep., RAND.

Kraft, M. A. and A. F. Gilmour (2016a): "Can Principals Promote Teacher Development as Evaluators? A Case Study of Principals Views and Experiences," *Educational Administration Quarterly*.

——— (2016b): "Revisiting the Widget Effect: Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness," *Brown University Working Paper*.

Kristensen, S. R., R. Meacock, A. J. Turner, R. Boaden, R. McDonald, M. Roland, and M. Sutton (2014): "Long-term effect of hospital pay for performance on mortality in England," *New England Journal of Medicine*, 371, 540–548.

Ladd, H. F. (1999): "The Dallas school accountability and incentive program: an evaluation of its impacts on student outcomes," *Economics of Education Review*, 18, 1–16.

Lazear, E. (1986): "Salaries and Piece Rates," *The Journal of Business*, 59, 405–431.

——— (2000): "Performance pay and productivity," *American Economic Review*, 90, 1346–1361.

Lazear, E. and S. Rosen (1981): "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy*, 89, 841–864.

Lazear, E. P. and P. Oyer (2012): "Personnel Economics," *The Handbook of Organizational Economics*, 479–517.

Levin, J. (2003): "Relational Incentive Contracts," *American Economic Review*, 93, 835–857.

Lynn Hannan, R., G. P. McPhee, A. H. Newman, and I. D. Tafkov (2013): "The effect of relative performance information on performance and effort allocation in a multi-task environment," *Accounting Review*, 88, 553–575.

MacLeod, W. B. (2003): "Optimal Contracting with Subjective Evaluation," *American Economic Review*, 93, 216–240.

Mann, D., T. Leutscher, and R. M. Reardon (2013): "Findings from a two-year examination of teacher engagement in TAP schools across Louisiana," Tech. Rep. September, Interactive Inc., Ashland.

Meyer, P. (2016): "Reliability of and Validity Evidence for Teaching Learning Framework Scores for the District of Columbia Public School System," Tech. rep., University of Virginia, Charlottesville.

Mirrlees, J. A. (1971): "An exploration in the theory of optimum income taxation," *The review of economic studies*, 38, 175–208.

Murnane, R. and D. Cohen (1986): "Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and Few Survive." *Harvard Educational Review*, 56, 1–17.

Neal, D. (2011): "The Design of Performance Pay in Education," in *Handbook of the Economics of Education*, ed. by E. Hanushek, Elsevier Science, chap. 6, 499–548, volume 4a ed.

Niederle, M. and L. Vesterlund (2007): "Do Women Shy Away From Competition? Do Men Compete Too Much?" *The Quarterly Journal of Economics*, 122, 1067–1101.

Oosterbeek, H., R. Sloof, and J. Sonnemans (2011): "Rent-seeking versus productive activities in a multi-task experiment," *European Economic Review*, 55, 630–643.

Oswald, A., E. Proto, and D. Sgroi (2015): "Happiness and productivity," *Journal of Labor Economics*, 33, 789.

Papay, J. P. (2012): "Refocusing the Debate: Assessing the Purposes and Tools of Teacher Evaluation," *Harvard Educational Review*, 82, 123–141.

Phipps, A. (2017): "Personnel Contracts under Production Uncertainty: Theory and Evidence from Teacher Performance Incentives," *University of Virginia Working Paper*.

Phipps, A. and E. Wiseman (2017): "Teacher Improvements in Windows of High-stakes Observation," *University of Virginia Working Paper*.

Prendergast, C. (1999): "The Provision of Incentives in Firms," *Journal of Economic Literature*, 37, 7–63.

Rivkin, S. G., E. Hanushek, and J. Kain (2005): "Teachers, schools, and academic achievement," *Econometrica*, 73, 417–458.

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

Rockoff, J. E. (2004): "The Impact of Individual Teachers on Student Achievement:Evidence from Panel Data," *The American Economic Review*, 94, 247–52.

Sartain, L., S. R. Stoelinga, and E. R. Brown (2011): "Rethinking teacher evaluation in chicago: lessons learned from classroom observations, principal-teacher conferences, and district implementation. research report," Tech. rep., Consortium on Chicago School Research, Chicago.

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

Springer, M. (2010): "Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching," Tech. rep., National Center On Performance Incentives at Vanderbilt Peabody College.

Steinberg, M. P. and M. L. Donaldson (2016): "The New Educational Accountability: Understanding the Landscape of Teacher Evaluation in the Post-NCLB Era," *Education Finance and Policy*, 11, 340–359.

Steinberg, M. P. and L. Sartain (2015): "Does Teacher Evaluation Improve School Performance? Experimental Evidence from Chicago's Excellence in Teaching Project," *Education Finance and Policy*, 10.

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

Van Dijk, F., J. Sonnemans, and F. Van Winden (2001): "Incentive systems in a real effort experiment," *European Economic Review*, 45, 187–214.

Van Herck, P., D. De Smedt, L. Annemans, R. Remmen, M. B. Rosenthal, and W. Sermeus (2010): "Systematic review: effects, design choices, and context of pay-for-performance in health care," *BMC health services research*, 10, 247.

Vigdor, J. L. (2008): "Teacher salary bonuses in North Carolina," *Vanderbilt Peabody College Working Papers*.

Weisberg, D., S. Sexton, J. Mulhern, and D. Keeling (2009): "The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness," Tech. rep., The New Teacher Project.

Wellington, A., H. Chiang, H. Kristin, C. Speroni, M. Herrmann, P. Burkander, and E. Warner (2016): "Evaluation of the Teacher Incentive Fund: Implementation and Early Impacts of Pay-for-Performance After Three Years," Tech. Rep. August, Institute of Education Sciences.

Zubanov, N. (2013): "Risk aversion and effort under an incentive pay scheme with multiplicative noise: Theory and experimental evidence," *Evidence-based HRM: a Global Forum for Empirical Scholarship*, 3, 130–144.