The Role of Production Uncertainty in Teacher Performance Pay: Theory and Experimental Evidence
Aaron Phipps *

Abstract

Teacher performance incentives have not had consistently positive effects in the U.S. 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. The result is a context in which individual teachers are uncertain about the net marginal productivity of inputs. The innovation of this paper is to develop a model in which uncertainty about the production process in student learning (“production uncertainty”) is incorporated explicitly in the model of behavior and to assess agent responses to different incentive schemes in a laboratory experiment.

*Phipps: University of Virginia. Special thanks to Sarah Turner, James Wyckoff, William Johnson, Charles Holt, Leora Friedberg, Brian Jacob, and Derek Neal for helpful comments and direction. Funding for this experiment was generously provided by the Steer Endowment Fund. The research reported here was supported by the Institute of Education Sciences, U.S. 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 U.S. Department of Education.
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 consist-
tent efficacy (Dee and Wyckoff, 2015). Recent large-scale federal initiatives in the United States,
including Race to the Top and the Teacher Incentive Fund have placed substantial emphasis on
the development of compensation mechanisms to link pay and performance, as 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 for 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. The result is
a context in which individual teachers are uncertain about the net marginal productivity of inputs.
The innovation of this paper is to develop a model in which uncertainty about the production
process in student learning (“production uncertainty”) is incorporated explicitly in the model of
behavior and to assess agent responses to different incentive schemes in a laboratory experiment.

In the model, the presence of production uncertainty reduces the effect of an outcome-based incen-
tive on a teacher’s overall effort level due to risk aversion, an effect I label “futility.” Furthermore,
an outcome-based incentive can induce production friction, which predicts that teachers will re-
distribute their effort to inputs with lower variance in marginal productivity, potentially reducing
average total productivity without necessarily decreasing overall effort. The initial empirical test
of these predictions is in a laboratory experiment, which is intended to provide a controlled setting
to isolate potential effects and should be seen as an opportunity to explore behavioral mecha-
nisms without the expense and implementation challenges of a field experiment or policy change.
Preliminary results of this experiment confirm the predictions of my model.

The experimental design itself is an innovation, allowing me to test many hypotheses of worker
behavior in complicated work settings with multiple inputs. The policy innovation of this paper is
to examine the saliency 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 in teacher incen-
tives 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. 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
that have been neglected in contract theory, such as health care.
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, and the existing empirical work on teacher performance incentives has a mixture of results that are not easily sorted by contract theory. This places my proposed research in the intersection of contract theory and education policy with contributions to both.

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, Barrow and Sander, 2007; Chetty, Friedman and Rockoff, 2014). There is room to improve the teaching labor force given the wide variance in the distribution of teacher effectiveness.\(^1\) 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.\(^2\) They 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 U.S. 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.

2.1 Teacher Performance Incentive Programs in the U.S. 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, U.S. 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 have responded by implementing teacher performance incentives that vary considerably on implementation details. While some districts pay bonuses to teachers if the entire school reaches certain goals, I will focus on incentives based on individual performance.\(^4\)

Empirical evidence of the effects of teacher-level performance incentives has been mixed. Dee

\(^1\)Hanushek (1992) finds that a teacher at the 95th percentile will get a gain of the equivalent of 1.5 academic years out of her students, while a teacher at the 5th percentile will get a gain of only 0.5.

\(^2\)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.

\(^4\)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).
and Keys (2004) found that an in-class evaluation based bonus built into the Tennessee STAR experiment, conducted in 1985, improved student math scores. The working paper Hudson (2010) and a technical report Mann, Leutscher and Reardon (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. 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.

Other programs have been found to have some positive effects, but they were small or not consistent across districts. Sojourner, Mykerezi and West (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 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).

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 et al., 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. Contract theory suggests that an output-based incentive, like those based on student test scores, should be sufficient to illicit increased effort from teachers, yet the theory cannot readily explain why some of these programs have worked while others have not.

### 2.2 Contract Theory in the Education Context

Contract theory has a rich set of models designed to address a variety of situations, not all of which are applicable in the teaching context (Prendergast, 1999). Among the key theoretical findings most salient to teaching that performance incentives should be based on measures of the final good, not individual inputs or multiple outputs (Holmstrom and Milgrom, 1991). Incentives have stronger effects as measurement noise decreases (Lazear, 1986), and they should be used when there is limited opportunity to cheat the measurement mechanism (Baker, 1992). Given advancements in standardized tests and their universal use, teacher value-added measures – estimates of how much a teacher has improved a student’s test scores – are a good candidate for performance incentives if standardized tests are assumed to capture enough important dimensions of the teaching production
function.

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 experiment, Bandiera, Barankay and Rasul (2007) shows 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 contract theory does not clearly delineate the mixed results.

The two main levers in contract theory that will modify the effect of an incentive are employee risk aversion and output measurement error. Increases in either dimension will reduce the expected effect of a performance incentive (Lazear, 1986). But teacher performance incentives do not vary along these dimensions: it is unlikely that teachers in Denver are more risk averse than teachers in Washington DC. Similarly, there is little difference in the measurement error of value-added scores among programs.

The consistent element of effective teacher performance incentives is the use of in-class evaluations. But given the established correlation between higher in-class evaluations and student test scores, contract theory would predict there is little advantage to making an incentive based both on in-class evaluation scores and value-added. Earning a bonus on student test scores should induce increased effort in practices measured by in-class evaluations if those practices improve test scores. Yet, in forthcoming work, Phipps uses the timing of unannounced in-class evaluations in Washington DC to show that in-class evaluations do in fact have an incentive effect even in the presence of a value-added incentive.

These observations are consistent with Prendergast’s concern that incentive contract theory is “not capturing all relevant features of compensation contracts” (Prendergast, 1999, p. 56). Incentive contract models determine optimal contracts based on output measurement noise and cost, employee risk aversion, and complementarity between inputs. I add a new dimension by relaxing a common assumption that teachers have a perfect knowledge of the teaching production process. This allows for differentiation between teacher incentive programs based on the uncertainty of the marginal productivity of different activities. In particular, it explains why in-class evaluations may have a stronger incentive effect.

### 2.3 Existing Incentive Program Design and Implementation

To illustrate the importance of in-class evaluations, I divide up the teacher incentive programs detailed above in Table 1 based on the use of incentives for in-class evaluation scores and any

---

5One possible difference between in-class evaluations and value-added measures is that there could be a notable difference in measurement noise. Because value-added measures draw on the test scores of 20-30 students, and the test scores are meant to capture the cumulative effects of a teacher throughout the year, it is unlikely that a couple hours of classroom observation will be less noisy. Empirically, there is no evidence that value-added measures for a teacher are meaningfully noisier than in-class observations, though this is inherently difficult to test.
measurable incentive effects. 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 1 highlights a positive relationship between a performance incentive’s effect and the use of in-class evaluations.

Table 1: Performance incentives appear more likely to improve student test scores when based, in part, on in-class evaluations.

<table>
<thead>
<tr>
<th>Differentiated In-Class Evaluations</th>
<th>Improved Student Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>3 0 0</td>
</tr>
<tr>
<td>Some</td>
<td>0 2 0</td>
</tr>
<tr>
<td>No</td>
<td>0 0 2</td>
</tr>
</tbody>
</table>

These programs were selected based on their emphasis on individual-level measures of teacher performance.

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. But Table 1 is suggestive. 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?

Contract theory would predict that a performance incentive based on measured output (student test scores) will be at least as effective as one based on measured inputs (teaching practices measured by in-class evaluations). The intuitive reasoning is that a teacher has asymmetric information about her students and her own skills. But in addition to the asymmetric information, the multiple areas scored in an in-class evaluation rubric allow the teacher to engage in multitasking, which in this context would imply she inefficiently distributes effort to the areas most likely to improve her in-class evaluation score, regardless of the effects on student test scores (Holmstrom and Milgrom, 1991). To explain this mismatch between contract theory and the empirical results in education performance incentives, I present an incentives contract model that explicitly allows for uncertainty in the production process. My model’s predictions are consistent with the empirical literature and can guide future field experiments and policy.

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.

Because some of the outlined above examine multiple programs in a single paper, “Some Differentiation” means that some of the programs in the study had differentiation in their in-class evaluations. Similarly, “Mixed” incentive effects means that some of the programs evaluated within the study have measurable effects.
3 Theoretical Model

3.1 Employee Problem

In a manner similar to the Holmstrom-Milgrom model, consider how a teacher can distribute her total time $\tau$, 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.\(^8\)

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 $\hat{1} = [1, 1, \ldots, 1]^T$, where the superscript $T$ indicates the transpose. Then assuming she is risk averse, her utility can be expressed using the exponential utility function

$$U = -\exp\{-r(b_w(w_0 + w_1 f(x)) + b_l(1 - \hat{1}^T x))\}. \quad (1)$$

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.\(^9\)

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 a Taylor Expansion around some reference input value, $x^r$. In a two-input case, this will be

$$f(x_1, x_2) \approx f(x_1^r, x_2^r) + f_1(x_1^r, x_2^r)(x_1 - x_1^r) + f_2(x_1^r, x_2^r)(x_2 - x_2^r) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2. \quad (2)$$

The values of $\gamma_1$, and $\gamma_2$ are then assumed to be random variables. I assume they have a multivariate normal distribution with mean $\mu_\gamma$ and covariance matrix $\Sigma_\gamma$.\(^{10}\) If the variance in $\gamma$ 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.

---

\(^8\)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.

\(^9\)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.

\(^{10}\)Notice that this specification assumes that the inputs are perfect substitutes in the traditional sense, but the covariance between $\gamma_1$ and $\gamma_2$ allows for an idea of risk substitution. That is, if $\text{cov}(\gamma_1, \gamma_2) > 0$ then if $x_1$ has high marginal productivity, then $x_2$ is likely to also have high marginal productivity. Taking a higher order Taylor Expansion will allow for complementarity in a more traditional sense.
Substituting the approximated production process into the expected utility function yields

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

(3)

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

\[ EU = - \exp \left\{ -r \left( b_w w_0 + b_l (1 - x^T \hat{1}) + b_w w_1 x^T \mu_\gamma - \frac{1}{2} r (b_w w_1)^2 x^T \Sigma_\gamma x \right) \right\} . \]  

(4)

Then the optimal choice of inputs \( x^* \) can be written as

\[ x^* = \frac{2}{r (b_w w_1)^2} \Sigma_\gamma^{-1} (b_w w_1 \mu_\gamma - b_l \hat{1}) . \]  

(5)

To explore the properties of this model, consider the two-input case. Let \( \sigma_{11} \) and \( \sigma_{22} \) denote the variances of \( \gamma_1 \) and \( \gamma_2 \), and let \( \sigma_{12} \) be their covariance. I can allow 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

\[ x_1^* = \left( \frac{A}{r (b_w w_1)^2} \right) \left( \frac{1}{\sigma_{11} \sigma_{22} - \sigma_{12}^2} \right) - \frac{C}{(b_w w_1 \mu_2 - b_l c) \sigma_{12} + (b_w w_1 \mu_1 - b_l) \sigma_{22}} \]  

(6)

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 \( \sigma_{12} \) is positive, it provides a form of risk reduction. This makes sense intuitively since positive covariance increases the information about the distribution of the marginal productivities of each input.\(^\text{11}\) Term \( C \) is the employee’s response to inputs being risk substitutes or complements, depending on the sign of \( \sigma_{12} \). 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**

While there are many interesting details and extensions, 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

\[ x_1^* = \left( \frac{b_w w_1 \mu_1 - c b_l}{r (b_w w_1)^2} \right) \frac{2}{\sigma_{11}} , \quad x_2^* = \left( \frac{b_w w_1 \mu_2 - b_l}{r (b_w w_1)^2} \right) \frac{2}{\sigma_{22}} . \]  

(7)

\(^\text{11}\)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.
The two main predictions I test with my experiment are Futility and Friction, which I introduce with abbreviated proofs:

**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 7.

In the experiment, this will be tested by measuring participants’ overall effort in completing their tasks, and then varying the magnitudes of \( \sigma_{11} \) and \( \sigma_{22} \). In the teaching context, futility represents the cost of a performance incentive that is inherent to the production process and independent of the quality of the measurement process.

**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 approximated linearly, the maximum average productivity requires that \( \sum_{i=1}^{N} x_i = x_{\text{max}} \) where \( x_{\text{max}} = \{ x_i : \mu_i > \mu_j \forall j \neq i \} \). Under a performance incentive with production uncertainty, it immediately follows from Equation 7 that \( x_i > 0 \) for all \( i \), which implies that \( \sum_{i=1}^{N} x_i \neq x_{\text{max}} \), and therefore expected productivity could be improved by re-allocating the same total amount of effort.

I 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 \( \sigma_{11} \) and \( \sigma_{22} \).

**Policy implications of Futility and Friction**

My model’s predictions of futility and friction are unique in contract theory. While other models predict decreased effort when output measures are noisy, futility is independent of the measurement process and an integral part of the production function. This changes how it should be corrected. The implication is that incentive contracts are improved, all other things constant, by selecting a measurement with a clear production function. The other predicted effect, friction, illustrates an unexplored, potentially adverse effect of performance incentives. In particular, relative to an incentive based on in-class evaluations, test-based incentives may induce an inefficient allocation of effort such that average productivity will decrease even without reducing overall effort. These predictions also provide two clear tests of the validity of my model, which I will test directly in a lab setting.

### 3.2 Employer Problem

The employer problem does not directly relate to principals since they do not set wages. In the education setting, the applicability of the employer problem is to quantify the trade-off between incentives based on in-class observations and value-added. While value-added incentives are usually cheaper to implement than a rigorous in-class evaluation system, they may be significantly less effective in providing an incentive and professional differentiation, and may even have a negative effect on average productivity.
Consider an employer with a product that has a price $p$. A risk-neutral employer will maximize expected profits:

$$\pi(w_1) = \mu^T x^*(w_1) (p - w_1).$$

For now I assume $p = 1$. Then the optimal choice of $w_1$ is

$$w_1^* = \frac{2}{b_w \mu^T \Sigma^{-1} \mu / b_l + 1}.$$  \hspace{1cm} (9)

An increase in the marginal utility from wages ($b_w$) will unambiguously decrease the optimal wage chosen by an employer. An increase in the marginal utility from leisure, $b_l$, will cause $w_1^*$ to increase, as expected. Assuming $\sigma_{12} = 0$ and $\mu \gg 1$, an increase in either $\sigma_{11}$ or $\sigma_{22}$ will lead to an increase in the optimal wage. This occurs because an increase in uncertainty will decrease an employee’s total effort, and so a compensating increase in the wage is necessary.

4 Experimental Design

I have designed and programmed a novel experimental procedure that allows me to test contract incentive effects in a multi-input environment. Participants must allocate their constrained time towards real-effort tasks to earn financial rewards. This reduces the contract problem to a few salient variables, such as the variance in the marginal productivity of inputs or the effort required for different inputs. The lab has the advantage of treatment randomization and precise control over the variation in key parameters. This enables me to show that people distribute their effort as predicted by my model. This experimental procedure can be used to test many contract models where the production process has multiple inputs, an area that has remained largely unexplored in the experimental literature.

4.1 Experimental Procedures

I have programmed the experiment in Python and hosted it online using Google App Engine. 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, Proto and Sgroi (2015). Easy questions are three two-digit numbers, while hard questions are 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 simulates a form of leisure time.

The activity is broken into three main sections – Fixed Wage, Random Coefficient, and Constant

\hspace{1cm} $^{12}$ is a comparison of the average productivity of inputs to the marginal cost of inputs. It is possible to differentiate the cost of inputs by changing $\hat{1}$ to a vector of input costs. The intuition is the same: as long as the expected marginal productivity of an input is larger than the cost, increases in uncertainty will decrease employee effort.
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 randomly assigned to account for the possibility that participants apply information gained in the first section to the second (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 7.
4.2 Treatment Design

The treatment in this experiment is the level of uncertainty in the production function. A difficult experimental design question with a continuous treatment variable is how much it should vary between treatments and over what range. I designed the initial trials of my experiment to help identify upper and lower bounds for the treatment variables, which has helped design the necessary treatment cells.

Treatment Parameters

Using the trial rounds, I have identified three key dimensions for altering the parameters of each experiment. The first is the expected earnings rate ratio, \( \kappa \). As described in the theoretical section, when participants have perfect knowledge of the payoffs for each input, they should eventually shift all their effort to a single input that provides the highest net payoff. Given the hard and easy payoffs per question, \( p_h \) and \( p_e \), the participant needs to compare the earning rates per minute, \( \frac{p_h}{r_h} \) and \( \frac{p_e}{r_e} \), where \( r_h \) and \( r_e \) are the length of time needed to complete hard and easy questions. If \( \frac{p_h}{r_h} > \frac{p_e}{r_e} \), then the participant earns more money per minute by answering hard questions, and should therefore eventually commit all their time to hard questions. I define the earnings rate ratio as \( \kappa = \frac{p_h}{r_h} \frac{p_e}{r_e} \). If \( \kappa > 1 \), participants should do all hard questions.

Because a participant’s question speed varies between questions, if \( \kappa \) is close to 1, it can be difficult for a participant to clearly identify which input is most efficient on average, which would make them less likely to commit to a single input in the constant coefficients treatment. There is also reason to believe that hard questions may carry additional mental costs that extend beyond the time they require to complete. If \( \kappa > 1 \), participants may have a different response to the random coefficients case than if \( \kappa < 1 \). Based on the trial rounds already run, I can set the parameters of the experiment such that \( \kappa = 1 \) for the average participant, but this isn’t necessarily desirable. For this reason, I need to vary the treatment parameters such that \( E[\kappa] \) is \( 1 - \sigma \), 1, and \( 1 + \sigma \), where \( \sigma \) is the standard deviation of the distribution of \( \kappa \) across participants.

Another important parameter is the size of the variance in payoffs for both easy and hard questions, represented as \( \sigma^2_e \) and \( \sigma^2_h \). I have two levels for each of these parameters, low and high. The last parameter of importance is the difference between payoff variance for easy and hard questions, \( \delta = \sigma^2_h - \sigma^2_e \). The effect of \( \delta \) depends on the difference between the average prices \( p_h \) and \( p_e \). The different treatment sizes are \( \frac{1}{2}(p_h - p_e) \), \( p_h - p_e \), and 0.

Risk Aversion in Teachers

As discussed in the background section, one common explanation in contract theory for the mixed results of teacher performance incentives is that teachers may be abnormally risk averse. To test for differences in risk aversion between education majors and other students, I am using both the Holt-Laury risk aversion test (Holt and Laury, 2002) and an improved version being developed by Holt. My sampling strategy will recruit roughly half of the participants from the Curry School of Education at the University of Virginia, and the other half from the general undergraduate student body at the University of Virginia. Most of the students that register to participate from the general student body are in pre-law, pre-business, or economics majors.

\( ^{13} \)These are analogous to \( \sigma_{11} \) and \( \sigma_{22} \) in the theory section.
5 Preliminary Empirical Results

There are three main predictions of the production uncertainty model that I test with this experiment. (1) Under constant coefficient, employees should devote all their working time to the input with the highest net marginal benefit. (2) Swapping the variance between easy and hard questions should lead to changes in the amount of time devoted to easy questions. In particular, going from high-variance in hard payoffs to high-variance in easy payoffs should lead to a decrease in the number of easy questions attempted and completed, even when the average output is optimized by focusing on a single input (friction). And (3) Relative to a constant coefficients payoff, participants in random coefficient treatments will attempt and complete fewer questions of either type (futility).

Sequencing Effects

Because each participant will be exposed to both constant and random coefficient treatments, it is possible that their second session is affected by their first session. These are potential sequencing effects. One way this can occur is if participants gain important information about their abilities in the first session. Much of this is mitigated by using the Fixed Payments session first, since participants experimented with both easy and hard questions and were able to evaluate their own abilities, albeit briefly. Another possible source of sequencing effects is that participants may get tired by the end of the experiment. Finally, it is possible that in going from a constant coefficients session to a random coefficients session, participants assume that the mean payoff is the same as the constant coefficient from the previous session.

Participants were randomly assigned to a treatment sequence. To check for sequential learning and for exhaustion, Table 3 reports the average number of Easy and Hard questions answered correctly per round by session. Because there are so few observations, the standard errors are very large. There is also large variation between subjects and within sessions. The average number of successful questions between sessions remains near constant around 13 easy questions and 1.5 hard questions per round, suggesting there were not strong sequencing effects from learning about ability or from exhaustion.

Distribution of Ability

While hard questions had exactly twice as many numbers to be added, the difficulty of retaining that many numbers at once makes it unlikely that hard questions were exactly twice as difficult as easy questions. Table 4 shows statistics about the distribution of completion times for easy and hard questions per participant in seconds. These are calculated by taking the average completion time for each correctly answered question across sessions one and two (but not the fixed payment session). The participant with the fastest average easy question completion time could successfully finish an easy question in 5.55 seconds, while the slowest averaged 19.6 seconds. Hard questions varied from 17.6 seconds to 35.6 seconds.

If participants accurately assess their ability, they will eventually focus on a single input type in the constant coefficients session. To evaluate how many participants should do all hard or all questions, Table 4 also includes the completion time ratio for an individual. Because hard questions are worth twice as much as easy questions in the constant coefficients session, a participant should dedicate all their time to hard questions if their Hard/Easy Ratio is less than 2 (assuming a negligible difference in the intensity cost between easy and hard questions). Only one participant averaged a Hard/Easy
Ratio of less than 2. The highest ratio was 3.93.

Treatment Comparison

To get a first comparison of the treatments, Table 5 reports mean summary statistics and their standard deviations. Easy/Round and Hard/Round are the average number of easy and hard questions correctly answered per round across all participants within a treatment. While this appears to have increased from fixed payment sessions to the treatment sessions, there is no meaningful difference between random and constant coefficient treatments.

Second/Correct is the length of time an average correct answer took across all rounds in a treatment. This is a measure of effort intensity. There is an increase in intensity for easy questions as participants move from the fixed payment session to the treatment sessions, but again there is little difference between the two treatment sessions.

The production value is calculated by applying the “true” marginal value of each input to the value of correctly answered questions. In this experiment, the true value is determined by the constant coefficients, which means an easy question is worth $0.07 and a hard question is worth $0.14. Again there is an increase from the fixed payment sessions to the treatment sessions, but little identifiable difference between random and constant coefficient treatments. Finally, the success rate also increases as participants move to random and constant coefficient treatments, but there is still little difference between the two treatments. This is not surprising since these averages do not account for changes in behavior across rounds and the differences in abilities between participants.

Testing Predictions

The first prediction is that, under constant coefficients, participants will eventually identify the input with the highest net marginal benefit and devote all their working time to that input. Figure 1 plots the percent of round time each participant dedicates to hard questions by round. In the constant coefficients rounds, participants have completely segregated their effort by the fifth round of eight. As noted, one individual had a hard/easy completion time ratio that was less than two, which would suggest that for them, hard questions had a higher net marginal benefit. This one individual is labeled as “High efficiency individual.” All participants eventually spent all of their time on easy questions in the constant coefficients round except the one individual for whom it was more efficient to dedicate all their time to hard questions. While only graphical evidence, this suggests participants can quickly identify the output-maximizing input.

In the random coefficients sessions, participants did not appear to converge on a single optimizing output in all cases. This is expected as participants adjust to their new beliefs about the distribution of payoffs.

To analyze the differences between the random and constant coefficient treatments, it is necessary to consider within-session sequencing effects and individual fixed effects. Table 6 reports a simple regression of the following specification:

$$Y_{i,r} = \beta_0 + \beta_1 D_{i,r} + \phi_i + \gamma_r + \epsilon_{i,r}$$ (10)

where $i$ and $r$ indicate individual and round. $D_{i,r}$ is a dummy variable indicating if the variance on hard question payoffs was high and $\phi_i$ and $\gamma_r$ are individual and round fixed effects.
The results indicate that there is a statistically significant relationship between the hard question payoff variance and the number of easy questions completed in a round, which confirms the second prediction of the production function uncertainty model. This latter outcome leads directly to the third prediction, since for nearly all participants, increasing the number of hard questions completed would decrease the average true value of production. This would also explain why, on average across random coefficient treatments, there was no observed difference in true production value between random coefficient treatments and constant coefficient treatments. Half of the random coefficient treatment group had an increased incentive to only answer easy questions because of their increased variance in the payoff for hard questions.

6 Conclusion

Production function uncertainty has several valuable predictions. In a version where agents are able to update their prior beliefs about the distribution of the marginal productivity of inputs, the model would predict 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 in $x$ whose variance in marginal productivity is nonzero increases, an agent’s overall effort will continue to decrease. Taken to infinity, this describes an infinite inputs dilemma, in which agents become paralyzed by too many options.\textsuperscript{14} This would imply that it is possible to overload teachers with too much training or too many curriculum options (Weisberg et al., 2009). There are also many behavioral economics applications to consumer choices in the health insurance market.

This paper demonstrates preliminary results in a laboratory setting to support the basic behavioral responses to an uncertain production function. Though limited in observations, the data already gathered illustrate that agents will allocate their effort as predicted when given constant and random marginal productivity. Under constant coefficients, agents eventually devote all their time to the input with the highest net marginal output. Under random coefficients, agents mix their inputs much more, and on average they respond to the variance of the marginal productivity of an input by shifting away as variance increases. This shift in inputs decreases average productivity when employees shift away from the input with the highest average net marginal productivity, even without decreasing effort.

\textsuperscript{14}This can be applied to health insurance choices and default options for pension plans, where the many options leads to inaction.
References


Table 2: Random Coefficient Distributions

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy, low variance</td>
<td>0.04</td>
<td>0.10</td>
<td>0.07</td>
<td>0.0003</td>
</tr>
<tr>
<td>Hard, high variance</td>
<td>0.07</td>
<td>0.21</td>
<td>0.14</td>
<td>0.0016</td>
</tr>
<tr>
<td>Easy, high variance</td>
<td>0.01</td>
<td>0.13</td>
<td>0.07</td>
<td>0.0012</td>
</tr>
<tr>
<td>Hard, low variance</td>
<td>0.11</td>
<td>0.17</td>
<td>0.14</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

Table 3: Success Rate by Session

<table>
<thead>
<tr>
<th></th>
<th>Fixed Payment</th>
<th>Session 1</th>
<th>Session 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy/Round</td>
<td>5.80 (2.44)</td>
<td>12.13 (9.28)</td>
<td>13.75 (9.03)</td>
</tr>
<tr>
<td>Hard/Round</td>
<td>2.30 (1.84)</td>
<td>1.60 (2.54)</td>
<td>1.50 (2.59)</td>
</tr>
</tbody>
</table>

Table 4: Summary Statistics of Question Completion Time in Seconds

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>5.55</td>
<td>8.56</td>
<td>19.57</td>
<td>10.11</td>
</tr>
<tr>
<td>Hard</td>
<td>17.62</td>
<td>23.38</td>
<td>35.63</td>
<td>24.85</td>
</tr>
<tr>
<td>Hard/Easy Ratio</td>
<td>1.82</td>
<td>2.59</td>
<td>3.93</td>
<td>2.72</td>
</tr>
</tbody>
</table>
Table 5: Summary of effort measures by treatment group

<table>
<thead>
<tr>
<th></th>
<th>Fixed</th>
<th>Random</th>
<th>Constant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy/Round</td>
<td>0.690</td>
<td>0.853</td>
<td>0.789</td>
</tr>
<tr>
<td></td>
<td>(0.32)</td>
<td>(0.19)</td>
<td>(0.26)</td>
</tr>
<tr>
<td>Hard/Round</td>
<td>0.560</td>
<td>0.742</td>
<td>0.815</td>
</tr>
<tr>
<td></td>
<td>(0.46)</td>
<td>(0.33)</td>
<td>(0.30)</td>
</tr>
<tr>
<td>Seconds/Correct Easy</td>
<td>12.698</td>
<td>9.760</td>
<td>9.493</td>
</tr>
<tr>
<td></td>
<td>(4.75)</td>
<td>(3.95)</td>
<td>(5.87)</td>
</tr>
<tr>
<td></td>
<td>(7.16)</td>
<td>(9.69)</td>
<td>(8.06)</td>
</tr>
<tr>
<td>Production Value</td>
<td>0.404</td>
<td>1.157</td>
<td>1.132</td>
</tr>
<tr>
<td></td>
<td>(0.26)</td>
<td>(0.43)</td>
<td>(0.56)</td>
</tr>
<tr>
<td>Easy Success Rate</td>
<td>0.690</td>
<td>0.853</td>
<td>0.789</td>
</tr>
<tr>
<td></td>
<td>(0.32)</td>
<td>(0.19)</td>
<td>(0.26)</td>
</tr>
<tr>
<td>Hard Success Rate</td>
<td>0.560</td>
<td>0.742</td>
<td>0.815</td>
</tr>
<tr>
<td></td>
<td>(0.46)</td>
<td>(0.33)</td>
<td>(0.30)</td>
</tr>
</tbody>
</table>

Table 6: Effect of high hard payoff variance on completed easy questions

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>7.14</td>
<td>2.72</td>
<td>2.63</td>
<td>0.012</td>
</tr>
<tr>
<td>High Hard Payoff Variance</td>
<td>10.75</td>
<td>3.26</td>
<td>3.30</td>
<td>0.002</td>
</tr>
<tr>
<td>Round Number</td>
<td>1.00</td>
<td>0.41</td>
<td>2.42</td>
<td>0.020</td>
</tr>
</tbody>
</table>
Figure 1: Percent of time allocated to hard questions by round and by treatment type.