CHAPTER 6

# The Design of Performance Pay in Education

**Derek Neal**
Department of Economics and the Committee on Education, University of Chicago and NBER

## Contents

## Abstract

This chapter analyzes the design of incentive schemes in education while reviewing empirical studies that evaluate performance pay programs for educators. Several themes emerge. First, education officials should not use one assessment system to create both educator performance metrics and measures of student achievement. To mitigate incentives for coaching, incentive systems should employ assessments that vary in both format and item content. Separate no-stakes assessments provide more reliable information about student achievement because they create no incentives for educators to take hidden actions that contaminate student test scores. Second, relative performance

*Handbook of the Economics of Education,* Volume 4
ISSN 0169-7218, DOI: 10.1016/B978-0-444-53444-6.00006-7

schemes are too rare in education. These schemes are more difficult to manipulate than systems built around psychometric or subjective performance standards. Third, assessment-based incentive schemes are mechanisms that complement rather than substitute for systems that promote parental choice, e.g. vouchers and charter schools.

## Keywords

Alignment
Standards
Relative Performance
Hidden Action

In most countries, the vast majority of elementary and secondary schools are publicly funded. Therefore, the study of the design of incentive systems employed in education is primarily the study of how governments structure the institutions and mechanisms used to procure a specific type of service. In democratic countries, the study of incentive design in education may also include explorations of agency problems that exist between voters and the elected officials that design these mechanisms.[1] However, in this chapter, I ignore these latter issues and focus on the design challenges that face a benevolent public education authority that seeks to maximize the welfare generated from public expenditures on education.

In broad terms, the personnel policies in most firms are designed, at least in part, to solve two problems. Firms must select and assign workers appropriately and then elicit efficient effort from them. If one treats schools as firms, the first problem involves the screening and assignment of teachers. Systems that create links between measures of teacher performance and the retention of teachers or the granting of tenure influence the distribution of talent among persons who occupy teaching jobs in equilibrium. Because the personnel policies employed in most public school systems create only weak links between teacher performance and retention or tenure decisions, scholars and reform advocates often cite existing tenure and retention policies as a potential source of inefficiency in personnel policies.[2] Nonetheless, systems that directly link retention and tenure decisions to measures of teacher performance have not yet been implemented on a large scale. Thus, my discussion below is restricted to simulation results that may help shed light on the potential effects of various reforms to tenure and retention policies. To date, there are no empirical studies that evaluate the effects of performance based promotion and retention systems.

A large section of this chapter examines the effects of performance pay systems that are designed to induce more effort from teachers. Although few incentive systems in education are currently employed as screening devices, many existing systems represent

[1] See Dixit (2002).

[2] See Ballou and Podgursky (1997) for an extensive treatment of the features of hiring and tenure processes in public schools that appear inefficient.

attempts to solve moral hazard problems between teachers and education authorities. Teachers typically work in a setting where the majority of their actions are hidden from their supervisors and where the contextual information that determines the efficient choice of actions at any point in time is also often hidden from their supervisors. In this setting, it is prohibitively expensive to write forcing contracts that specify desired actions for each potential classroom setting and then provide the monitoring required to make sure that these desired actions are taken.

Faced with these monitoring problems, education authorities may pursue one of two strategies. They can pay educators flat salaries and seek to shape the effort choices of educators through professional development and the processes used to screen and recruit teachers, or they can link incentive payments to some measure of educator performance. In recent decades, education authorities throughout the world have begun to experiment with the latter approach, and the existing literature contains many papers that evaluate the impacts of various teacher incentive programs. However, at this point, the literature contains few papers that formally explore the design of incentive systems for educators. I argue here that many performance pay schemes in education are poorly designed, and a careful review of the empirical literature on the results of various performance pay schemes reveals that poor design yields poor results in predictable ways.

Most performance pay schemes in education are constructed as contests against predetermined performance standards, in which teachers receive rewards if their measured performance exceeds a specified target. A large literature notes that such schemes are problematic when teachers can take actions that inflate the measured achievement of their students relative to their students' true skill levels, and I devote considerable attention to this issue. However, I also note that the tasks of choosing the psychometric performance standards used in such contests and maintaining the integrity of these standards over time are difficult ones. Variation in student assessment results reflects not only variation in educator performance but also variation in the backgrounds and aptitudes of students. Systems that do not correctly control for student characteristics in the creation of performance targets for educators create incentives for educators to avoid certain types of students or schools. In addition, no existing contest schemes contain procedures that adjust performance standards over time to reflect secular progress in available teaching methods. Finally, there is considerable suggestive evidence that performance standards can be compromised by testing agencies that make changes to assessment content or the scaling of assessments over time that compromise the meaning of psychometric scales.

Performance pay schemes built around subjective performance evaluation avoid the technical problems involved in setting statistical performance standards, but these systems have not worked well in practice. Two recent studies suggest that when one group of government educators evaluates the performance of another group of government educators subjectively, performance pay schemes may well morph into increases in base pay for teachers that are not accompanied by improvements in teacher performance.

Even if we assume that an education authority has access to a set of performance metrics that do isolate variation in educator performance relative to the current pedagogical frontier, a simple model of pay for performance contests shows that education authorities can waste resources by setting performance standards either too low or too high. Systems that set standards too low either pay more in prize money than is required given the effort they elicit or elicit less than efficient effort levels or both. Further, systems that set standards too high can be especially wasteful because some educators respond by keeping effort levels low and treating the incentive system as a lottery. Circumstantial evidence suggests that designers of recent systems have, in some cases, set performance standards well above efficient levels, and in other cases, set them far below efficient levels.

In contrast to performance pay schemes built around fixed performance standards, relative performance schemes often elicit more useful information for two reasons. First, the evolution of the distribution of measured performance among educators over time can provide information about how the education production frontier is evolving over time. Further, systems that involve competition among educators for a fixed pool of reward money cannot easily be manipulated into a means of raising base pay for an entire population of teachers who make no changes in their effort levels. Nonetheless, relative performance incentive schemes are rare in education and thus far have been mostly confined to the realm of short-term experiments. Although these experiments produced some encouraging results, there are no large scale relative pay for performance schemes in operation at this time.

Many accountability and performance pay systems employ test scores from assessment systems that produce information used not only to determine rewards and punishments for educators but also to inform the public about secular progress in student learning. As long as education authorities keep trying to accomplish both of these tasks with one set of assessments, they will continue to fail at both tasks. If the goal of assessing students is to measure trends in secular achievement, separate no-stakes assessments provide information that is not likely to be contaminated by hidden actions. However, when authorities use one set of assessment results for both incentive pay and student assessment, educators face incentives to take numerous hidden actions that simultaneously inflate their own measured performance and contaminate information about levels of student achievement.

If education authorities implement separate assessment systems for performance incentives and student assessment, they still face the possibility that educators will engage in wasteful hidden actions that manipulate the results of tests used to determine performance pay, but authorities can mitigate some of these concerns by linking performance pay to the results of assessments that contain no year to year overlap in item content or format. This design eliminates the incentive for teachers to engage in coaching behaviors that do not build lasting skills but simply prepare students for a particular set of questions or test formats. Although assessments without repeated items and common formats cannot be readily placed on a common psychometric scale, the ordinal content of these assessment results can be used to implement performance contest schemes that elicit efficient effort from teachers.

Throughout much of the chapter, I consider an education authority that employs many teachers in many different schools and seeks to design personnel policies that screen and motivate teachers. However, the final sections of the chapter consider the design of incentive systems that operate at the school level, and I discuss how governments can design systems that require schools to compete for public support. Seen through this lens, voucher systems, charter schools, and other systems that expand school choice are complements to and not substitutes for incentive systems built around assessment results.

## 1. SCREENING TEACHERS

A large empirical literature documents the fact that measured teacher productivity varies greatly among teachers, holding constant observed characteristics of their students and their school environments. However, it is difficult to use observed characteristics of candidate teachers to predict who will actually perform well in the classroom.[3] This later finding is consistent with two different views of the information structure in teacher labor markets. The first view contends that candidate teachers know whether or not they will be effective teachers, but they cannot directly reveal this information to prospective employers in a credible way. In this asymmetric information scenario, personnel policies must be designed in ways that induce teachers to reveal their ability type indirectly. A second view is that neither a new teacher or her principal knows how effective she will be and that both parties learn about her effectiveness as she gains experience. In this symmetric learning scenario, personnel policies dictate how teacher compensation evolves as new information about her productivity is revealed and also whether or not she will be allowed to continue teaching given her performance record.

For the purpose of this chapter, I adopt the second view and consider the design of policies that maximize the output of teachers employed at a point in time as well as the sequence of teachers who will occupy a given position over time. I return below to the question of how pay should vary with measured performance. For now, I focus on the issue of whether or not teachers should be allowed to continue teaching based on their past record.

Rockoff and Staiger (2010) make the first formal attempt to derive firing rules that maximize the steady-state output of teachers in a school district. They note that the

[3] See Aaronson, Barrow, and Sander (2003); Rockoff (2004); Rivkin, Hanushek, and Kain (2005); and Hanushek and Rivkin (2006).

measured productivity of teachers varies greatly among teachers and that existing research suggests that current hiring and screening procedures in public schools may do little to narrow the dispersion of productivity among new teachers.[4] They also note that the most important cost of replacing teachers after one or two years on the job is that new teachers typically perform significantly worse than teachers with one or two years of experience.

Using a set of assumptions about the reliability of measured teacher performance, the dispersion in teacher performance, the returns to early experience, and the rate of exogenous exits from teaching, Rockoff and Staiger derive optimal firing rules under various assumptions about how many years teachers teach before school districts make an up or down decision on their retention. They choose rules that maximize the steady-state average productivity per teacher, which is equivalent to maximizing the steady-state total output of the school system since they are holding constant the number of teaching positions in their hypothetical school system.

The policy recommendations that Rockoff and Staiger produce are quite different from current practice in modern school systems. They consider a number of different scenarios that involve different tenure clocks, variances of measurement error in teacher productivity, and hiring costs. However, they always conclude that school systems should dismiss at least two thirds of each new cohort of teachers during their first few years of experience.

The Rockoff and Staiger approach is based on a steady-state analysis that involves the following thought experiment: for any retention policy that describes when teachers will be retained or dismissed based on their history of measured performance, derive the steady-state distribution of teacher quality in the system. Then, choose the policy that maximizes average steady-state teacher quality.

It is not clear that this exercise is the most relevant for policy analysis. If a given school system adopted a Rockoff and Staiger retention policy today that applied to new hires but existing teachers continued to enjoy the same employment protection they enjoy now, it could easily take 20 years for the system to approach the steady-state that Rockoff and Staiger describe. A different and possibly more relevant approach is to consider the policy that maximizes the expected discounted value of teacher quality generated by the sequence of teachers who will occupy a position that is open today. Further, because it is standard in the literature to assume that individual teacher productivity is not influenced by the quality of her co-workers, the optimal rule for one position is the optimal rule for all positions.

Rockoff and Staiger note that most of the returns to experience among teachers come, on average, quite early in their careers. They conclude that the existing literature

[4] See Ballou and Podgursky (1997).

implies that the performance of first year teachers is on average roughly .47 standard deviations below the average quality of experienced teachers,[5] but teachers with one year of experience perform almost as well as more experienced teachers, and the returns to experience appear to be roughly zero after two years of experience. Here, I ignore the small returns to experience in year two and focus on the larger returns to experience in the first year of teaching. Under the assumption that the education authority is risk neutral and that the authority is maximizing the present discounted value of teacher quality measured in standard deviation units, the assumption that new teachers are, on average .47 standard deviations less productive than experienced teachers is equivalent to the assumption that the authority must pay a search cost of .47 to fire an experienced teacher and hire a new one. Thus, if one ignores any effects of experience after the first year, the retention policy problem facing an education authority can be described using a well-known model of job matching.

Let $\theta$ denote the true productivity of a given teacher. $\theta$ is not observed directly, but each period $t$ that a teacher works, the education authority observes a productivity signal, $x_t$. In year one, $x_1 = -.47 + \theta + \varepsilon_1$. For years $t > 1$, $x_t = \theta + \varepsilon_t$. Here, $\varepsilon_t$ represents measurement error or some other transitory component of measured productivity. For all $t = 1, 2, \ldots, \infty$, $\varepsilon_t$ is drawn identically and independently over time and teachers. The model is denominated in standard deviation units of teacher quality. Assume that $\theta \sim N(0, 1)$ and that $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)\ \forall t$. Let $m_t$ be the posterior belief about expected productivity of a given teacher based on the history of her measured performance, $(x_{t-1}, x_{t-2}, \ldots, x_1)$, and let $\rho_t$ equal the precision of the authority's beliefs about teacher quality at the beginning of year $t$ of her career. Teachers never die in this model, but there is an exogenous probability, $\delta$, that a teacher leaves teaching in a given period for reasons unrelated to her productivity. Finally, let $\beta$ be the authority's discount rate.

The timing of events is as follows: the authority hires a new teacher. At the end of the teacher's first period of work, the authority observes $x_1$ and forms $(m_2, \rho_2)$. The authority then either allows the teacher to work another period or fires the teacher and hires a new teacher. If the authority retains the teacher, the authority repeats the same review and retention decision process at the end of the teacher's second period of work using both signals, $(x_1, x_2)$, and the same process repeats in future periods. At the beginning of each period, the education authority is trying to maximize the expected present value of teacher productivity generated by the teachers who fill a particular slot. The Bellman equation that describes the problem facing the education authority is:

$$V(m_t, \rho_t) = max[V_0, m_t + \beta(1-\delta)E[V(m_{t+1}, \rho_{t+1}) \mid m_t] + \beta\delta V_0]$$

[5] According to Rockoff and Staiger (2010), a one standard deviation improvement in teacher quality is associated with roughly a .15 standard deviation increase in expected student achievement, and on average, the students of rookie teachers perform about .07 standard deviations below students of experienced teachers.

Here, $V(m_t, \rho_t)$ is the value of having a current teacher with $t-1$ periods of experience and a history of productivity such that $m_t = E(\theta \mid x_{t-1}, x_{t-2}, \ldots x_1)$, and $V_0 = V(0,1)$ is the expected value of hiring a new teacher.[6]

Many readers may recognize that I have characterized the education authority's problem using Jovanovic's (1979) model of job matching. Jovanovic describes how a worker optimally searches for a job when he believes that his potential match with each new job comes from the same distribution. I use the model to describe how an education authority optimally fills a vacancy when the authority believes that each new teacher is drawn from the same productivity distribution.[7] The Jovanovic model is well-known in labor economics, and it is well established that the optimal policy for the authority is to choose a set of cutoff values, $(r_1, r_2, r_3, \ldots)$, such that teachers are dismissed at the beginning of period $t$ if $m_t < r_t$. As long as one assumes that a teacher's actions only affect output in her own classroom, the authority can maximize the expected present value of total productivity in the school system by using this same policy to fill all teaching positions.

I have solved this model using the parameters for $\delta$ and $\sigma_\varepsilon^2$ that Rockoff and Staiger employ, and to simplify the numerical analysis, I assume no teacher works more than thirty years.[8] Given the exogenous quit rate of $\delta = .05$, this assumption has virtually no affect on the optimal cutoffs early in a teacher's career.

The Jovanovic approach differs conceptually from Rockoff and Staiger's steady-state analysis because it explicitly incorporates discounting and because it imposes no tenure clock. As Rockoff and Staiger acknowledge, policies that maximize steady-state payoffs do not properly discount the returns that occur in steady-state. The main cost of firing a teacher is the poor expected performance of the new replacement teacher. This cost is paid today. However, if we assume that existing teachers would continue to enjoy their current employment protections following any changes to the tenure system for new teachers, the benefits of a higher steady-state average teacher quality would come decades from now. Further, rules that force up or out tenure decisions early in a teacher's career raise optimal promotion standards because the education authority cannot correct the mistake of giving tenure to a candidate who is later revealed to be less than deserving.

Thus, it is not surprising that the Jovanovic simulations yield much more conservative dismissal policies than those produced by the Rockoff and Staiger simulations. Exact dismissal rates vary with parameter choices, but the typical set of rules implies that roughly fifty percent of new teachers should be dismissed after one year and small

[6] $\rho_t$ is only a function of $t$ because this is a normal learning problem.

[7] Jovanovic (1979) assumed that workers receive all the surplus for employer-employee matches. I am assuming that there is a fixed wage for teachers that the authority must pay to any teacher that fills a slot. Thus, the authority simply wants to maximize the expected present value of productivity generated by each teaching slot.

[8] $\delta = .05$ and $\sigma_\varepsilon^2 = 1.5$, which implies a reliability ratio of .4.

fractions of new teachers should be dismissed in years two through six of their tenure with roughly forty percent of new teachers never facing dismissal. Nonetheless, both simulations suggest more stringent firing rules than we currently observe in most public school systems.[9] Thus, it is important to consider whether or not these simulations form a solid basis for considering drastic changes in personnel practices.

Although both exercises provide interesting starting points for broader research on retention policy, both also share important shortcomings. First, if public schools adopt aggressive firing policies, schools may have to raise salaries to maintain the current quality of their applicant pool. It is not clear how elastic the quality constant supply of potential teachers is, but it is certainly a key consideration for any policy makers who contemplate following Rockoff and Staiger's advice. Second, the more important assumption built into both sets of simulations is that teacher productivity is a fixed trait that does not vary with teacher effort other than through mechanical learning by doing. The simulation exercises described here help us think about some of the costs of the current hiring and firing procedures in public schools, but those who take the resulting dismissal rules seriously as viable policy prescriptions are implicitly or explicitly embracing the view that differences in measured teacher productivity are entirely due to differences in teacher talent and not differences in teacher effort.

Given this starting point, the only way to deal with low performing teachers is to terminate them. Better incentive provision has no value. However, this view of personnel policy is rather extreme given the existing literature on incentives in professional labor markets, and it also reflects a false interpretation of some well known results from the empirical literature on teacher productivity.

The fact that teachers vary in terms of their measured productivity does not imply anything about whether or not most teachers provide socially efficient levels of effort given their talent or whether or not it is possible to improve the entire distribution of teacher productivity through the use of incentives. Further, while the evidence on heterogeneous teacher productivity surely reflects a degree of true talent heterogeneity among teachers, it may also reflect differences among teachers in their own personal effort norms. Given the absence of incentive pay and the level of job security protections in many public school settings, these differences in personal norms could be an important source of ex post differences in teacher performance.

The distinction between talent heterogeneity and norm heterogeneity is important when one is trying to forecast the expected benefits from better incentive provision

[9] Work by Adams (1997) implies that total separations among young teachers are likely around half the levels of dismissals implied by the rules generated by the Jovanovic simulations. Thus, even if one assumes that all current teachers who quit are being forced out, the implied dismissal rates in the data are quite different than those implied by either set of simulations.

for teachers. If bad teachers are simply teachers who are not able to learn how to teach well, then better performance pay schemes should yield negligible improvements in the distribution of teacher performance. On the other hand, if bad teachers are teachers who are not motivated to take the steps required to teach well, then improvements in the design of incentives may generate significant improvements in the distribution of teacher performance without significant increases in total teacher compensation.

Finally, the types of firing rules discussed here can never operate only as screening mechanisms. Policies that link retention decisions to measures of teacher performance should induce more effort from teachers,[10] and if differences in effort norms are important ex ante, the introduction of these policies should alter the ex post distribution of teacher productivity. In fact, it seems reasonable to conjecture that, if a school system announced even the Jovanovic style dismissal rules that I describe above, the administrators of this system would observe that the threat of dismissal alters the distribution of teacher productivity by compressing differences in teacher effort levels among teachers who share the same talent level. In this scenario, the dismissal rules announced ex ante would no longer be optimal ex post because key parameters in the simulation would be influenced by the change in policy.[11]

There is little evidence that existing hiring procedures in public schools work well as mechanisms for identifying candidates who will perform well in the classroom. Further, many public school teachers receive tenure in almost a perfunctory manner quite early in their careers.[12] These observations give credence to the notion that better screening and retention polices could yield large gains in teacher productivity. However, the combination of perfunctory tenure, civil service employment protections, and civil service salary schedules also suggest that the dead weight loss associated with inefficient effort allocation among existing teachers is a first order concern as well, regardless of whether or not one contends that many existing teachers should not be allowed to continue teaching.

## 2. MORAL HAZARD

The literature on the use of assessment based incentive schemes in education often draws a distinction between accountability systems and performance pay systems. Assessment-based accountability systems are promoted as vehicles for holding public schools accountable

[10] The analyses presented here rest on the assumption that teachers earn more than they could in other jobs requiring the same effort levels. If no teachers are earning rents, then it is hard to imagine how any changes in personnel policies could improve teacher performance without spending more money on teacher salaries.

[11] Technically, the simulations that I conducted and that Rockoff and Staiger conducted suffer from the same problem because estimates of the variance of teacher value-added are taken from the existing stock of current teachers. However, the Rockoff and Staiger agenda is motivated by the view that there is now a weak correlation at best between being a poor performing teacher and a teacher that leaves teaching.

[12] See Ballou and Podgursky (1997).

for their use of public funds. These systems define achievement standards for students and then measure the performance of schools using metrics that describe the degree of discrepancy between the standards set by the accountability systems and the achievement of the student populations in various schools. Further, these systems often also include a set of sanctions that school administrators and teachers face if their students fail to meet the performance targets set by the accountability system.

In sum, accountability systems typically seek to accomplish two tasks. They attempt to measure the performance of schools relative to a set of public standards in a manner that is consistent over schools and over time. Further, they create incentives for educators to provide effective instruction for their students. Thus, the paradigm that dominates accountability system design involves a two-step procedure. First, measure performance relative to public standards. Then, reward or punish schools based on success or failure to meet these standards.

Because accountability systems typically contain rewards and sanctions that are either not spelled out in detail or less than credible because they cannot be enforced ex post,[13] the primary function of most accountability systems is performance measurement. In contrast, performance pay systems are more explicitly focused on incentive provision and often contain precise mappings between the performance of students and the compensation and employment status of educators.

In this chapter, I focus most of my attention on performance pay systems for several reasons. To begin, the purpose of this chapter is to explore theory and evidence concerning the design of incentive systems for educators, and performance pay systems are explicit incentive schemes. Further, one of my main conclusions will be that accountability systems should not be used as incentive systems. Systems that serve as mechanisms for providing public information about the achievement of students and the performance of schools relative to public education standards should not contain rewards or sanctions that provide incentives for educators.

Donald Campbell (1976) offered the following summary of common patterns he observed in case studies of the use of performance metrics for incentive provision in government agencies,

> *"I come to the following pessimistic laws (at least for the U.S. scene): The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."*
>
> **Campbell, (1976)**

[13] One of the most infamous examples of a system that contains incredible threats of sanctions is the No Child Left Behind Act of 2001 (NCLB). Neal (2010) provided a discussion of how the school re-organization threats attached to the 100% proficiency requirement create confusion concerning how the law will be enforced in future years when this requirement becomes binding and tens of thousands of schools have failed to meet it.

A key component of social decision making in Campbell's analyses is resource allocation among government workers or their units. Thus, one way to understand Campbell's Law is that, when government tries to pursue two missions, e.g. incentive provision and performance measurement, with one system, it fails at both missions. Campbell offered this observation as an empirical law. I will use a simple model of worker responses to incentive schemes to explain why Campbell observed what he observed. In section 2.7, I will discuss how education officials can improve both performance measurement and incentive provision by developing separate systems that address these goals independently.

## 2.1. A Simple Treatment of the Multi-Tasking Problem

The multi-tasking model of Holmstrom and Milgrom (1991) is the tool that economists most often employ to organize their thoughts about the responses of teachers to various merit pay schemes. Here, I use a special case of their model to build intuition concerning the forces that shape the optimal design of incentives in education.[14]

Consider an education authority that hires one teacher to teach one student. The teacher allocates her effort among two different tasks. Let $t_1$ be the time that the teacher devotes to task one, and let $t_2$ denote the time she devotes to task two. The human capital production technology is such that

$$h = f_1 t_1 + f_2 t_2 + e$$

where $(h - e)$ is the human capital acquired by the student as a result of the teacher's efforts. Here, $h$ is an addition to the value of a skill set, and it is measured in dollars. $f_1$ and $f_2$ are constants, and $e$ is a random shock to the learning process that captures factors beyond the teacher's control that affect the student's rate of learning. The authority cannot observe $h, t_1$, or $t_2$. However, the authority can observe a statistical measure of teacher performance, $p$, where

$$p = g_1 t_1 + g_2 t_2 + v$$

$g_1$ and $g_2$ are constants, and $v$ is a random shock that influences measured performance. Here, we assume that both $e$ and $v$ are shocks drawn independently from distributions with mean zero that do not depend on the actions of the teacher, $(t_1, t_2)$. We also assume that the teacher's utility function can be described by

$$U = X - C(t_1, t_2)$$

[14] Here, I follow the Gibbons (2010) exposition of the model. See Baker (2002) for a related treatment.

where $X$ is the teacher's expected income and $C(t_1, t_2)$ describes the teacher's cost of effort for any pair $(t_1, t_2)$. Now, suppose the education authority seeks to design an optimal compensation contract of the form

$$w = s + bp$$

where $s$ is a base salary, and $b$ is a bonus rate associated with the performance measure $p$. The base salary $s$ is not interesting for our purposes because it is only a mechanism for dividing surplus between the teacher and the authority. Given any choice of $b$, one can choose a base salary large enough to elicit the teacher's participation given some outside utility option $U_0$.

The optimal bonus rate $b$ is the solution to the following problem:

$$\max_{b} \quad f_1 t_1(b) + f_2 t_2(b) - C(t_1(b), t_2(b)) \quad s.t.$$

$$[t_1(b), t_2(b)] = \arg\max_{t_1, t_2} \quad s + b(g_1 t_1 + g_2 t_2) - C(t_1, t_2)$$

In words, the optimal bonus rate maximizes the difference between the expected value of the human capital created by the teacher's action and the cost of the teacher's actions taking into account that the teacher's response to any bonus rate, $b$, will be to chose actions that maximize her utility given $b$. Assume the following cost function for teacher effort,

$$C(t_1, t_2) = .5(t_1 - \bar{t}_1)^2 + .5(t_2)^2$$

where $\bar{t}_1$ is a norm for time devoted to effective instruction. The education authority may have established this norm through previous expenditures devoted to screening potential teachers or professional development activities. The key is that, for the purposes of this analysis, the norm is fixed and not affected by the incentive scheme. The education authority chooses the optimal incentive structure taking $\bar{t}_1$ as given.

Given this simple setup, a few calculations reveal that the optimal bonus rate is

$$b^* = \frac{f_1 g_1 + f_2 g_2}{g_1^2 + g_2^2} = \frac{\|\mathbf{f}\|}{\|\mathbf{g}\|} \cos\theta$$

where $\theta$ is the angle between the vectors $(f_1, f_2)$ and $(g_1, g_2)$. See Figure 6.1.[15]

By assuming that workers and firms are risk neutral and that costs are quadratic, I have made sure that the formula for $b$ is simple. Nonetheless, this formula highlights two factors that shape the optimal strength of incentives in more general settings. To begin, cos $\theta$ is an alignment factor. If the vectors are orthogonal, e.g. $(f_1 = 0, f_2 > 0)$ and

[15] The points [(0, 0), $(f_1, f_2)$, $(g_1, g_2)$] form a triangle that can be split into two right triangles. Based on the right triangle that includes the origin, it is easy to show that $\cos\theta = \frac{f_1 g_1 + f_2 g_2}{\|\mathbf{f}\| \|\mathbf{g}\|}$.

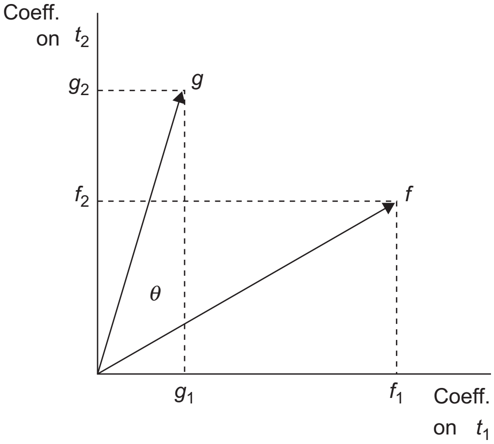

**Figure 6.1** Alignment ard Incentives.

$(g_1 > 0, g_2 = 0)$, then $\cos\theta = \cos 90 = 0$, and $b^* = 0$. If the performance measure is aligned perfectly with true output, the two vectors fall on the same ray from the origin, $\cos\theta = \cos 0 = 1$. The ratio preceding $\cos\theta$ is a scale factor. Note that when the performance metric is perfectly aligned, the ratio of the two vector lengths simply transforms the units of the performance metric into dollars, i.e. the monetary value of the human capital created by the teacher's actions.

## 2.2. Is Incentive Pay Efficient in Education?

I return specifically to the topics of alignment and scale below, but I first want to discuss what this model teaches concerning whether or not the presence of at least some form of incentive pay is optimal, i.e. $b^* > 0$. This version of the multi-tasking model implies that positive incentives are optimal, $b^* > 0$, as long as $f_1g_1 + f_2g_2 \neq 0$. To see this, note that in cases where $b^* < 0$, the authority can always implement $b^* > 0$ by defining a new performance metric $p' = -p$.

The condition, $f_1g_1 + f_2g_2 > 0$, always holds if at least three of the four constants, $(f_1, f_2, g_1, g_2)$, are strictly positive and none are negative, i.e. if one task contributes to both output and the performance measure, the other task contributes to one or both, and neither task is detrimental to either real output or the performance measure, then some positive level of incentive pay is optimal. Define $t_1$ as time spent teaching the curriculum using teaching techniques deemed to be best practice, and then note that $t_2$ may be given many different labels, e.g. coaching students regarding test-taking strategies, changing student answers before assessments are graded, etc. Most discussions of educators' responses to high-stakes testing regimes implicitly assume that $f_1 > 0$, $g_1 > 0$, $g_2 > 0$, and thus, if the gaming activities, $t_2$, do not harm students, $f_2 \geq 0$, then optimal policy in this framework must involve some positive level of performance pay, $b^* > 0$.

On its surface, the condition $f_2 \geq 0$ seems like a fairly weak requirement, and thus the formula for $b^*$ above seems to indicate that some positive level of incentive pay is always efficient. However, this feature of $b^*$ is not a robust feature of the multi-tasking model because it hinges on separability in the cost function.

Since education requires both teacher and student time and since students have finite energy and attention, cost functions of the following form may be just as interesting to consider:

$$C(t_1, t_2) = .5(t_1 + t_2 - \bar{t})^2$$

Here, $t_1$ and $t_2$ are perfect substitutes, and $\bar{t}$ is a norm for total effort that influences teacher costs. I assume that in the absence of any incentive pay, teachers choose $t_1 = \bar{t}$ and $t_2 = 0$.[16] Given this setting, if the education authority chooses $b > 0$, then teachers choose $t_1 = 0$ as long as $g_2 > g_1$ and there are many combinations of $f_1, f_2$, $g_2$, and $\bar{t}$ such that the authority should optimally choose to have no incentive pay and accept $f_1\bar{t}$ as classroom output. When $f_1 > f_2$ and $f_1\bar{t}$ represents baseline output, any incentive scheme that causes teachers to substitute small amounts of $t_2$ for $t_1$ lowers output while holding costs constant.

Nonetheless, if $\bar{t}$ is low enough, incentive pay may still increase total surplus. Since $t_1$ cannot be negative, the global benefits of increasing $t_2$ well beyond $\bar{t}$ may compensate for the loss of $t_1 = \bar{t}$. Thus, whether or not $b^* = 0$ hinges on the technology of classroom instruction and the norm, $\bar{t}$, that exists in a school system.

A key consideration in the literature on responses to incentive schemes in education has been the precise nature of the activities, $t_2$, induced by the incentive scheme and the relative values of $f_2$ and $g_2$ given the maintained assumption that $g_2 > g_1$. However, it is equally important to consider whether or not $t_2$ actions represent an increase in total teacher effort or simply substitution away from $t_1$. In the latter scenario, effort norms and the nature of teacher cost functions are key. Schemes that induce teachers to devote more effort to coaching and test preparation may improve overall performance among teachers who were devoting little time to effective teaching ex ante. However, these same schemes are harmful when they induce effective teachers to replace effective instruction with significant amounts of coaching.

Now that I have presented a basic model that can serve as a guide to interpreting empirical results on incentive schemes in education, I draw your attention to Table 6.1. This table contains a summary of pay for performance schemes that are either ongoing

[16] Teachers are indifferent among any combinations $(t_1, t_2)$ such that $t_1 + t_2 = \bar{t}$, but here I assume that they use the students' best interests as a tie breaker.

Table 6.1 Existing Evidence on Performance Pay Systems and Their Effects

| Program | Place/Time | Description | Study | Results |
|---|---|---|---|---|
| Career Ladder Evaluation System | Tennessee, 1985–1989 | 5-stage career ladder for teachers that awarded bonuses after reaching the third stage. Bonuses ranged from $1000 for the third certification level to $7000 for the fifth certification level. | Dee and Keys (2004) | Math scores increased by 3%, reading scores by 2%, but only increases in math were statistically significant. Teachers on the lower 3 rungs were more effective at promoting math achievement, and teachers at the higher rungs were more effective at promoting reading achievement. |
| CIS | Kenya, 1997–1998 | School based program that awarded bonuses to schools for either being the top-scoring school or for showing the most improvement. Bonuses were divided equally among all teachers in a school, who were working with grades 4–8. | Glewwe, Ilias, and Kremer (2010) | The program increased government exam participation. It did not increase scores in the first year, but treatment scores rose by .14 SDs relative to controls in the second year. However, this improvement did not persist after the completion of the program, and there were no improvements on parallel low stakes NGO exams. |
| ABC | North Carolina 1996–Present | School based program that awards bonuses to all teachers if school-wide scores meet statistical target. $1500 maximum bonus. Part of the state accountability system. | Vigdor (2009) | Large Gains in Math and Reading Proficiency on the State Test. NAEP trends suggest that reading gains are suspect, but math gains may reflect real improvement. |
| DSAIP | Dallas, 1991–1995 | Schools were ranked based on gains in student learning. Approximately the top 20% of schools received awards for | Ladd (1999) | Pass rates on standardized tests of Reading and Math increased significantly, but only for white and Hispanic students. |

*Continued*

**Table 6.1** Existing Evidence on Performance Pay Systems and Their Effects—continued

| Program | Place/Time | Description | Study | Results |
|---|---|---|---|---|
| | | each member of their staff. Principals and teachers received $1000 bonuses, and other staff received $500. | | Black students did not exhibit significant gains relative to other cities. The dropout rate decreased more in Dallas relative to other cities from 1991 to 1994. |
| KIRIS | Kentucky, 1992–1996 | Schools could earn bonus money if they achieved growth targets for school-wide performance on assessments as well as other objectives. | Koretz and Barron (1998), Koretz (2002) | Scores on KIRIS assessments rose dramatically in all subjects, but Kentucky students showed modest gains or no improvement on many comparable NAEP or ACT tests. |
| Teachers' Incentive Intervention | Israel, 1995–1997 | Schools were ranked based on their relative performance adjusted for student background characteristics. Credits hours, matriculation exam pass rates, and dropout rates served as performance criteria. The top 1/3 of schools received awards. 75% of the award went to bonuses for teachers, 25% of the award went to facilities improvements. | Lavy (2002) | Clear evidence of improved outcomes on most dimensions with larger impacts observed in religious schools. Matriculation certificates did not increase in secular schools, but average test scores increased in both secular and religious schools. |
| PRP | England, 1999–Present | Teachers submit applications for bonus pay and provide documentation of better than average performance in promoting student achievement. Teachers who are promoted become eligible for future raises if they meet documented criteria. | Atkinson et al. (2009) | No clear evidence of improvement. Given one strategy that sought to adjust for experience differences between treatment and controls, English and Science teachers showed modest improvement. Math teachers did not show improvement. |

*Continued*

**Table 6.1** Existing Evidence on Performance Pay Systems and Their Effects—continued

| Program | Place/Time | Description | Study | Results |
|---|---|---|---|---|
| TAP 1999- | 17 states, 227 schools | Statistical VAM method produces teacher performance indices of 1 to 5. Teachers with scores of 3 or greater earn a bonus that increases with their score. | Hudson (2010) | Introduction of TAP raises math achievement relative to samples in a synthetic control group by .15 SDs. Reading impacts positive but smaller and imprecisely estimated. |
| Israel (experiment) | Israel, 2000–2001 | A rank-order tournament among teachers of each subject, with fixed rewards of several levels. Teachers were ranked based on how many students passed the matriculation exam, as well as the average scores of their students. | Lavy (2009) | There were overall improvements in pass rates in Math and English due to an overall change in teaching methods, increased after school teaching, and increased responsiveness to student needs among teachers. Increased exam participation rates also played a role in test score gains. |
| Andhra Pradesh (Randomized Evaluation Study) | India, 2005–2007 | 100 schools got group bonuses based on school performance, and 100 got individual bonuses based on teacher performance. Bonuses were awarded based on how much the percentage gain in average test scores exceeded 5%. | Muralidharan and Sundararaman (2010) | After 2 years, students in incentive schools scored better than the control group by .28 SDs in math, and .17 SDs in language. These students also tended to do better on questions of all difficulty. Students at incentive schools also did better in non-incentive subjects. |
| Achievement Challenge Pilot Project (ACPP) | Little Rock, Arkansas, 2004–2007 | Individual teachers were awarded bonuses based on their students' improvement on the Iowa Test of Basic Skills. Awards were determined by the level of growth and number of students a teacher had. | Winters (2008) | There was statistically significant improvement in all three subjects (math, reading, language) tested. Students increased 3.5 Normal Curve Equivalent (NCE) points in math (.16 SDs), 3.3 NCE points in reading (.15 SDs), and 4.6 NCE points in language (.22 SDs). |

*Continued*

Table 6.1 Existing Evidence on Performance Pay Systems and Their Effects—continued

| Program | Place/Time | Description | Study | Results |
|---|---|---|---|---|
| POINT | Nashville, TN, 2006–2009 | Teachers volunteered to participate in a performance pay experiment. Bonuses of 5 K, 10 K, and 15 K were awarded for surpassing the 80%, 90%, and 95% threshold in the historic distribution of value-added. | Springer et al. (2010) | Program involved 5th- through 8th-grade math teachers. Some evidence of achievement gains in 5th-grade math in years two and three, but these gains did not persist over the next school year. No evidence of positive program impacts in other grades. Attrition rates from the study were high years two and three. Attrition is concentrated among inexperienced teachers. |
| NYC School-Wide Bonus Program | New York City, 2007–2011 | Random sample of "high-need" schools participated in a bonus pay scheme. The scheme involved team incentive pay at the school level linked to growth targets, but school compensation committees distributed the bonus money among teachers. The two bonus levels were $3000 per teacher and $1500 per teacher. The program was added on top of an accountability program that already put performance pressure on schools. | Goodman and Turner (2010) | Performance scores were weighted averages of improvements in test score performance and inspections of school environment. Target scores required lower-performing schools to make greater improvements. 2008–2009 was the only full year of implementation. 89% of eligible schools won the maximum bonus. There is no clear evidence that the program improved student achievement. |

*Continued*

**Table 6.1** Existing Evidence on Performance Pay Systems and Their Effects—continued

| Program | Place/Time | Description | Study | Results |
|---|---|---|---|---|
| Portugal's Scale Reform | Portugal, 2007–Present | Abandoned single pay scale in favor of two scale system. Promotion to higher pay scale involved a level jump of about 25% of monthly salary. Teachers in the same school who already worked on the higher pay scale performed the performance assessments for junior teachers. | Martins (2009) | Using schools in the Azores and Madeira as well as private schools as controls, there is no evidence of achievement gains induced by the program and consistent evidence that the program harmed achievement on national exams. |
| MAP | Florida | Districts choose their own method for measuring teacher contribution to achievement. | No Independent Study | |
| Procomp | Denver, 2006–Present | Teachers and principals negotiate achievement targets for individual students. Teachers can also earn bonuses for meeting state growth expectations for their students based on statistical targets. Finally, teachers may earn bonuses for serving in a "distinguished" school. School ratings are determined by test scores, parent surveys, and attendance. Maximum performance bonus is 5%. | No Independent Study | |
| Qcomp | Minnesota, 2007–Present | Much of performance pay linked to evaluations of lesson plans and their implementation. Schools or districts develop their own plans for measuring teacher contributions to measured students achievement. | No Independent Study | |

or have been implemented in the recent past. The table devotes particular attention to schemes that have been evaluated by teams of independent scholars, and it covers performance pay schemes from several different countries. Most of these studies address schemes implemented in the developed world, but a few address performance pay in developing countries. As I work through various implications of agency theory for the design of incentive schemes, I will refer back to related empirical results in Table 6.1.

## 2.3. Generalizability

Table 6.1 shows that many assessment-based performance pay schemes do generate noteworthy increases in student performance on the particular assessment used for incentive provision. Thus, Table 6.1 provides much evidence against the notion that educators simply do not respond to incentives. The exceptions are the PRP system in England, the recent pay scale reform in Portugal, and two recent experiments in New York City and Tennessee. I will comment below concerning the unique features of these schemes that may have muted incentives.

I begin my review of the empirical studies in Table 6.1 by asking how many studies provide evidence that a particular incentive scheme induced changes in teacher effort allocations that improved results on a particular assessment but did not improve students' actual skill levels. In the framework set out above, it seems natural to assume that, given most interpretations of $t_2, f_1 > 0, f_2 \geq 0, g_1 \geq 0, g_2 > 0,$ and this implies that any incentive scheme with $b > 0$ will induce teachers to supply more total effort $t_1 + t_2 > \bar{t}$, since the marginal costs of both efforts are zero given $t_1 + t_2 = \bar{t}$. However, the choice of $b > 0$ is clearly not welfare improving if the increased total effort by teachers improves measured performance, $p$, without generating improvements in actual student human capital, $h$. This combination of outcomes implies that teachers are expending more effort in response to the incentive scheme without improving student learning, which suggests that improvements in measured performance are coming through increases in $t_2$ that crowd out time devoted to $t_1$ and result in lower student human capital.

If the ex post evaluation of a given incentive scheme reveals that student learning did improve, this is not clear evidence that the introduction of the scheme improved welfare. Such a finding constitutes evidence that the scheme created real benefits for students, but these benefits may or may not be greater than the costs of the program. These costs include not only the resources required to implement the program but also any losses of student skill in areas that are not assessed and therefore given less attention after such schemes are implemented.

On the other hand, if studies that evaluate the effects of a given incentive plan reveal no real improvements in student skill, then there is good reason to suspect that the plan is not efficient. Implementing incentive schemes usually requires new resources, and schemes that do not generate real improvements in student skills that are the targets of assessments are not likely sources of improvements in skills that are not directly assessed.

Empirical research on these questions is fundamentally difficult because neither policy makers or researchers observe true skill, *h*. Nonetheless, a significant body of research on high-stakes testing systems attempts to make inferences about changes in *h* by exploring whether or not assessment gains induced by particular incentive systems generalize to other assessments. For example, assume that a school district decides to implement a performance pay system for fifth-grade math teachers, and the district links teacher pay to student results on assessment A. Further, assume that following the introduction of this program, student results on assessment A improve. The generalizability issues that interests many researchers in educational statistics are variations on the following counterfactual question;

> *"Suppose that in every period, the fifth-grade students in this district had also taken a second math assessment, B, and teachers were not rewarded or punished as a result of student outcomes on assessment B. Would one have observed gains on assessment B following the introduction of incentive pay that were comparable to the gains observed on assessment A?"*

In sum, do gains measured on the assessments used to determine incentive payments reflect increases in skill that create general improvements in math assessment results or only improvements specific to one assessment format or a particular set of questions?

If gains on a particular high-stakes assessment do not generalize, this is not clear evidence that the incentive system induced no changes in teacher behavior that created real increases in skill. Assessments differ in terms of their relative focus on various topics and assessment B may simply not cover the skills assessed on A that improved. Nonetheless, if one finds that gains on a particular high-stakes assessment do not generalize at all to other assessments that are designed to cover the same curriculum, it is possible that the gains on the high-stakes assessment represent no lasting contributions to student skill. In this case, the district likely induced socially wasteful allocations of teacher effort that improved high-stakes assessment results without improving student skills.

Koretz (2002) summarizes results from several studies of generalizability, and he discusses three different types of teacher behavior that could generate gains on high-stakes tests that do not generalize to other assessments of the same subject matter. To begin, teachers may narrow their instructional focus. If teachers respond to incentives by devoting more class time to topics listed in the curriculum and stressed on a related high-stakes assessments, then scores on these assessments may rise substantially while scores on broader assessments of the same subject may show only modest improvements. This scenario is a plausible explanation for the results found in some generalizability studies, but it seems far fetched as an explanation for why some studies document large improvements on high-stakes assessments while scores on contemporaneous low-stakes assessments of the same domain remain flat or even fall.

Here, it seems more likely that teachers are engaging in what Koretz calls coaching. Coaching involves activities that improve scores on a given assessment without improving student mastery of a subject. Stecher (2002) reviews observational studies of coaching behaviors and cites a striking example of such behavior in two Arizona school districts that introduced high-stakes assessment systems. Shephard and Dougherty (1991) report that teachers in these districts reduced the number of writing assignments they gave to students and increased the number of assignments that involved having students find mistakes in prepared passages. This change in teaching practice likely harmed the development of writing skill among students, but it makes sense as a strategy for raising test scores on standardized tests.[17]

Koretz also noted that some educator responses to high-stakes assessment systems go beyond coaching and constitute cheating. I discuss specific examples of cheating in the next section, but for now, I note that both coaching and cheating should generate measured achievement gains that do not generalize to other assessments.

Clean evidence on the generalizability of assessment gains is rare, and the existing literature does not speak with one voice. Some studies provide fairly persuasive evidence that the measured gains induced by a particular performance pay program represented little or no improvement in actual subject mastery. Others provide suggestive evidence that at least a portion of the measured gains induced by particular programs reflects real skill gains.

I begin by considering two programs in Table 6.1 that both involve performance pay that is determined by assessments results collected within state accountability programs. The ABC program in North Carolina allows all teachers in a given school to earn a bonus of up to $1,500 per teacher based on the test score performance of all the students in the school relative to targets determined by a statistical model that conditions on historical performance in the particular school in question and in the state as a whole. The KIRIS system in Kentucky began in 1992. This system also provided bonus pay for teachers based on team performance. All teachers in a school could earn bonuses if the overall performance of students in their school surpassed targets determined by KIRIS formulas.

Koretz and Barron (1998) examine the effects of KIRIS on achievement during the period 1992–1996. Vigdor (2009) examines the effects on ABC of student achievement in North Carolina. Both studies compare trends in NAEP scores with trends in scores on the state specific assessments used to create school accountability measures and

[17] Stecher (2002) reviewed several related practices that have been documented in other states. In math, a related practice involves working only on math problems that follow a format or rubric know to be present on a particular high-stakes assessment.

determine bonus payments. Koretz and Barron (1998) report results in standard deviation units. Vigdor (2009) reports trends in proficiency rates. These studies provide evidence that KIRIS and ABC produced noteworthy gains in reading and math scores on state assessments. Further, in some subjects and grades, the improvements on the KIRIS exams were extremely large.

Nonetheless, NAEP scores in Kentucky improved by only modest amounts and at rates no greater than one would have expected based on national trends, and reading proficiency rates in North Carolina follow a similar pattern. In fact, eighth grade reading proficiency levels on NAEP in North Carolina have been lower than for most of the past decade than they were in the late 1990s when the state introduced the ABC system. Still, since the introduction of ABC, proficiency rates in math on both the state assessment and the NAEP have risen steadily, and although Vigdor does not compare North Carolina NAEP trends in math with trends in other states, the math results from ABC are at least consistent with the hypothesis that ABC generated gains in math achievement that are not entirely specific to the ABC assessment system.

The ABC and KIRIS programs are of particular interest here because they involved cash payments to educators and independent researchers have explored the generalizability of the gains induced by these systems. However, there is a larger literature on the generalizability of gains induced by high-stakes accountability systems generally. Jacob (2005) concludes that an accountability system introduced in the Chicago Public Schools in 1996 generated noteworthy gains in scores on high-stakes assessments, but he reports that scores on low stakes assessments did not improve among third and sixth grade students relative to what one would have expected based on pre-existing trends in Chicago test scores. Jacob finds that both high and low stakes scores rose sharply among eighth graders, and he concludes that the Chicago accountability program generated increases in general skills among older students but not among younger students.

Klein et al (2000) examine data from Texas during the 1990s. The Texas Assessment of Academic Skills (TASS) program began in the 1990s, and this accountability program received considerable attention because scores on TASS exams rose dramatically following the introduction of state wide accountability measures. However, Klein et al demonstrate that, between 1992 and 1996, changes in NAEP reading and math tests did not always square with corresponding changes in TASS scores. Fourth grade math scores on the NAEP rose sharply in Texas relative to scores in other states, but changes in NAEP fourth-grade reading scores and changes in NAEP eighth grade math scores in Texas followed the same pattern in Texas that one observes nationwide.

Hanushek and Raymond (2005) analyze differences among states in NAEP scores trends during the 1990s and conclude that accountability systems improve student learning if they contain real threats of sanctions for educators when students perform poorly. They reach this conclusion by comparing the time pattern of state-level changes in NAEP scores with the timing of the introduction of state level accountability systems of different types. They conclude that accountability systems that only create public report cards for schools generate at most small gains in achievement but systems that contain real sanctions for poor educator performance generate noteworthy gains in NAEP scores.

In 2001, the No Child Left Behind Act (NCLB) forced all states to adopt accountability systems that contained, at the least, threats of serious sanctions for poorly performing schools. Because NCLB is a nationwide program, it is nearly impossible to precisely assess its impact on general skill development, but it is clear that measured achievement gains on most state assessments have greatly exceeded gains on the NAEP since 2001.

In sum, the United States literature suggests that assessment based incentives schemes typically generate measured improvements on the assessments used to determine rewards and sanctions for educators, and in some cases but not all, these gains appear to reflect improvements in general student skills. Readers may be less than surprised to learn that results from generalizability studies outside the United States also provide mixed results.

Glewwe (2009) argues that agency problems between public school teachers and education authorities are often much more severe in developing countries than in the developed world. In many developing countries, teachers earn wages that are many times greater than per capita GDP, yet teachers are often absent from school and often absent from their classrooms even when they attend school. He summarizes evidence from a number of developing countries and makes a compelling case that public school teachers in many developing countries perform poorly while earning large wage rents.

Given these stylized facts, policy makers and researchers are interested in learning more about the potential benefits of performance pay schemes for educators in less-developed countries. Two recent studies employ data from field experiments in Kenya and India. These settings are interesting because, in both countries, teachers earn much more than the typical worker and also work within civil service systems that provide extraordinary job security and few performance pressures. The high wages offered to teachers in these countries permit both governments to fill teaching positions with well educated people, but the civil service systems in both countries create widespread concern about teacher effort. As in many other developing countries, absenteeism is a significant problem, and policy makers have concerns about the effort level of teachers who do show up for work. Given the status quo in both countries, some may conjecture that the introduction of incentive pay should create real benefits in both countries. However, the results from these experiments are quite mixed.

Glewwe, Ilias, and Kremer (2010) evaluate an incentive pay program run as an experiment in Kenya by International Child Support (ICS).[18] The program began by

[18] ICS is a Dutch organization that funds education and health interventions that seek to help children in developing countries.

selecting 100 schools that appeared to be performing less than optimally. From these schools, the program administrators chose 50 schools to participate in a program that awarded prizes to teachers based on student test score performance in their schools. The plan involved team incentives since all teachers who worked with students in grades 4 through 8 received common prizes based on an aggregate measure of the performance of all of their students.

The prizes ranged in value from 21 to 43 percent of the monthly earnings of a typical teacher. Students took two types of exams. The program linked teacher prizes to scores on government exams, but ICS also created another set of exams that involved no stakes for teachers. The program generated little evidence of test score improvements during the first year or the program. In the second year, the program created large score gains on government tests but no improvements in scores on the low stakes exams.

Glewwe et al conclude that teachers responded to the program by increasing the number of test preparation sessions held for students. They find no evidence of improvements in teacher attendance or classroom practice. Further, they report that even the improvements on government exams did not persist in year three after the incentive program ended.

The fact that the ICS experiment generated measured improvements in student achievement that did not generalize to a parallel low stakes assessment is not shocking given the results reviewed above. However, it is noteworthy that relative student performance on high-stakes exams returned to pre-program levels when the incentive experiment ended. Thus, the test preparation sessions and other activities that generated the measured improvements in high-stakes test performance during the program did not even generate lasting improvements in test taking skills or knowledge specific to the government exams.

While the Glewwe et al results provide suggestive evidence that the Kenyan program was socially wasteful, a recent incentive pay program in India appears to have generated some real gains for students. The Andhra Pradesh Randomized Evaluation Study (APRES) is a large study of experimental interventions in government primary schools located in the Indian state of Andhra Pradesh. School years begin in the middle of June in Andhra Pradesh, and the program began in late June of 2005 with baseline testing of students in treatment and control schools. Two of the treatments specified in this project involved bonus pay schemes based on student test score outcomes in future years.

Let $\Delta\bar{s}$ equal the percentage point increase in the average score of the students in a given classroom or school. Teachers received bonus pay equal to $500 * \max[0, \Delta\bar{s} - 5]$. Teachers who participated in the group incentive plan received bonuses based on school-level average improvements, so, if the average score in a team-incentive school increased by .07, all teachers received a bonus of 1,000 rupees. Teachers in the individual incentive program received bonuses according to the same formula based on the performance of their own students.

Muralidharan and Sundararaman (2010) estimate the impacts of these two incentive programs by comparing test score outcomes in treatment schools over the subsequent two years to outcomes in a group of control schools. The APRES design randomly assigned 100 schools to each of the two treatments and the control sample. Both incentive programs generated significant improvements in student tests scores. Taken as a whole, the incentive programs raised scores over a two-year period relative to the control group by .27 standard deviations in math and .17 standard deviations in language. The measured impacts in year two are somewhat larger in schools treated with the individual incentive program.

The APRES experiment did not collect test score data from any parallel set of low-stakes math and reading exams. Thus, it is not possible to perform a direct analysis of the generalizability of these gains. However, Muralidharan and Sundararaman provide much suggestive evidence that these gains do reflect at least some real contributions to students' subject mastery. Scores on social studies and science tests also rose significantly in incentive schools relative to control schools even though teachers received bonus pay based only on the math and language results. Further, there is evidence that teachers in incentive schools assigned more work and conducted classes beyond regular school hours. On the other hand, there is evidence that part of the extra class time was devoted to taking practice tests, which may have involved some coaching behaviors. Further, there is no evidence that teachers in incentive schools improved their attendance rates, which remained far below levels found in developed countries.

The contrast between the results from Kenya and India points to the need for more research on what features of the design and implementation of incentive programs improve outcomes. One obvious difference between the two programs is that the Kenyan program tied reward pay to results on national examinations that had been in place for a long time while the APRES experiment developed their own exams for the program. The greater apparent prevalence of coaching as opposed to improved teaching in Kenya may signal that familiarity with the national exam system greatly raised the relative returns to coaching. This conjecture is quite speculative at this point, but I argue below that coaching is less of a concern if education authorities implement assessment systems such that the specific item content and format of each assessment is not predictable.

## 2.4. Other Hidden Actions and the Contamination of Information

In the previous section, I discussed implicit and explicit evidence that teachers coach students for specific exam questions and question formats in response to high-stakes assessments. Although coaching is typically not an optimal allocation of teacher effort, some forms of coaching may generate some lasting human capital gains for students, and if coaching activities reflect reduced leisure on the part of teachers rather than reductions in effective teaching time, it is possible that these incentive schemes are improving educator performance relative to the performance one expects given public sector monitoring alone. Nonetheless, a different literature documents other ways that some teachers respond to assessment based incentive schemes that are almost certainly wasteful from a social perspective. In Koretz's (2002) taxonomy, these activities constitute cheating.

Jacob and Levitt (2003) provide clear and compelling evidence that some teachers or principals in Chicago responded to the introduction of high-stakes accountability in 1996 by simply changing their students' answer sheets before returning them. It is worth noting that these cheaters were not terribly sophisticated. Jacob and Levitt found that some classes got entire blocks of questions correct even though their performance on the remaining questions implies that it should have been almost impossible for the whole class to get any one set of even two or three questions correct. The scores for students linked with cheating often reflect large increases from the previous year, and these same students experience small improvements or declines in the following year. Jacob and Levitt conclude that cheating took place in between three and five percent of Chicago classrooms following the introduction of high-stakes testing.

Figlio and Winicki (2005) present evidence that schools in Virginia that faced the most serious threats of sanctions under a state accountability system responded by increasing the sugar content of student meals on the day the state administered high-stakes tests. They also cite several media reports of similar behavior in response to high-stakes assessment systems in other areas of the country. School officials appear to be responding to a literature that links test score performance to glucose level in test takers, and these actions represent a textbook example of how agents may respond to the presence of an incentive system by taking hidden actions that inflate their measured performance but contribute nothing to their actual performance.

Jacob and Levitt (2003) and Figlio and Winicki (2005) show that high-stakes assessment systems induce some educators to engage in behaviors that are socially wasteful. These socially wasteful behaviors as well as the coaching activities described above contaminate public information about school performance in two ways. First, since these types of manipulations inflate assessment results, these behaviors contaminate measures of how student achievement is evolving over time on average in a state, district, or school. Second, because some educators are likely more prone to engage in these manipulations than others, these manipulations also distort our understanding of the relative performance of different districts, schools, and teachers. This second point is often missed in current policy debates. The case studies that Campbell (1976) reviewed involve scenarios in which gaming behaviors contaminate information about the performance of some unit or agency over time. However, if the teachers and principals in various schools differ in their personal norms concerning their distaste for coaching or cheating behaviors, then heterogeneity in coaching or cheating contaminates the information that assessments provide concerning relative performance levels in a cross-section of teachers or schools at a point in time.

Suppose school A has higher measured performance than school B under a low stakes assessment regime, but the measured performance of school B exceeds that of school A after the introduction of a high-stakes assessment program. There are two possibilities. School B may have instituted real improvements and now is more effective than school A, or the staff of school B may simply be more willing than the staff of school A to engage in coaching or cheating behaviors that inflate measured performance. This last possibility may be thought of as Campbell's Law turned on its side because it points to the possibility that hidden responses to incentive schemes may contaminate not only time-series information concerning the true evolution of average performance but also cross-sectional information about the true relative performance of various units at a point in time.

Some policy makers may have their own preferred strategies for minimizing cheating or coaching through the use of independent testing agencies or other devices. However, if the assessment system used to measure student performance or educator performance relative to public standards is a no stakes system that is completely separate from any system of incentives or sanctions for educators, there is no reason for educators to engage in coaching or cheating in the first place. Any assessment-based performance pay system must contain strategies for minimizing gaming behaviors, but the best strategy for making sure that public measurement systems actually produce reliable measurements is to make these systems separate from any systems that reward or punish educators.[19]

## 2.5. Choosing Targets and Prizes

In section 2.1 above, I presented a model where the education authority takes the performance metric as given and must choose an optimal linear piece-rate given this performance metric. Most assessment based incentive programs in education do not resemble piece rate schemes where educators earn bonus pay as a linear function of some scaled performance metric. Instead, most incentive programs for educators are contest schemes, and more often than not, these contests do not involve competition among educators but rather competition to surpass a performance target.

In section 2.1, I described a human capital production function in which teacher actions are the only source of growth in human capital or increase in measured performance, and I used this model to discuss the alignment of incentive schemes when teachers can take multiple hidden actions. Now, I want to set aside the issue of alignment and focus on the choice of performance targets and prizes given a well aligned

[19] Cullen and Reback (2006) showed that schools may also alter the results of assessment-based accountability systems by manipulating the distribution of students who are tested.

performance metric. The existing literature contains little discussion of these issues. The papers summarized in Table 6.1 contain no formal analyses of how to set performance targets for contests or prizes given certain performance targets in order to maximize some clearly defined social objective function.

I assume that teachers can engage in only one action, $t$, which one can think of as time spent employing optimal teaching practices. Further, because I am concerned with how the authority sets performance targets, I also model changes in student human capital and measured academic performance that do not result from teacher effort but rather from baseline learning that reflects activities directed by a student's parents or the student himself.

For now, I continue to assume that each teacher teaches one student and specify the model as follows:

$$h = \gamma(h_0) + ft + e$$
$$p = \varphi(h_0) + gt + v$$
$$U = X - \frac{c}{2}(t - \bar{t})^2$$

Here, as in section 2.1, $h$ is the human capital the student possesses at the end of the period, and $p$ is the measured performance of the student at the end of the period. But now, the educational production function includes $\gamma(h_0)$, which captures baseline learning that is not attributed to teacher effort, and $\varphi(h_0)$, which captures the effect of baseline learning on measured achievement. Both of these baseline learning factors are functions of the student's human capital stock at the beginning of the period, $h_0$. The parameters $f$ and $g$ capture the effects of $t$ on human capital growth and changes in measured performance respectively. The terms $e$ and $v$ are mean zero error terms that reflect shocks to the creation and measurement of human capital. Both are drawn identically and independently over all student-teacher pairs, and both distributions are unimodal and symmetric around zero. Let $\Phi(.)$ denote the cumulative distribution function of $v$. Realizations of $v$ determine the outcomes of contests in equilibrium.

As before, $U$ is teacher utility and $X$ denotes expected teacher income. The cost of effort function is quadratic around the effort norm $\bar{t}$. Given this setup, the condition $(t^* - \bar{t}) = \frac{f}{c}$ defines socially optimal teacher effort. To keep things simple, I have chosen a setting such that optimal teacher effort is the same for all teachers regardless of the levels of $h_0$ their students possess. However, performance standards in this setting will vary with $h_0$.[20]

[20] In versions where students learn at different rates given the same instruction, the efficient level of instruction will vary among students even when all teachers are homogeneous. If both teachers and students are heterogeneous, the social optimum also involves not only a specification of instruction levels for each student but also the assignment of students to teachers.

I begin by discussing the design of optimal contests against performance standards in a setting where the authority understands teacher preferences and the technology of instruction, knows the quantity $\varphi(h_0)$ for each student and observes $p$ but not $h$ at the end of the period. Let the authority define $\hat{p}(h_0, t^*)$ as the expected measured performance for a student who begins the period with human capital $h_0$ and receives efficient instruction from his teacher. Assume the authority knows $\hat{p}(h_0, t^*)$ for each student and let the authority announce the following contest scheme. Teachers receive their base salary independent of their effort choice. They also receive bonuses, $\pi$, if the measured performance of their students is greater than or equal to the relevant values of $\hat{p}(h_0, t^*)$. The problem facing each teacher is

$$\max_t \ \pi[1 - \Phi(\hat{p}(h_0, t^*) - \varphi(h_0) - gt)] - \frac{c}{2}(t - \bar{t})^2$$

and the teacher's first order condition is

$$\pi g \phi(\hat{p}(h_0, t^*) - \varphi(h_0) - gt) = c(t - \bar{t})$$

Now, suppose that the authority chooses $\pi = \frac{f}{g\phi(0)}$, then the solution to this first-order condition becomes $(t^* - \bar{t}) = \frac{f}{c}$, and it is straightforward to show that the second order condition for a local maximum is also satisfied at $t^*$. However, more work is required to demonstrate that $t^*$ is a global solution to the teacher's problem. If the density $\phi(v)$ falls too quickly as $|v|$ increases, the teacher may find that the total cost of choosing $t^*$ is greater than the expected return.[21]

Nonetheless, for reasonable parameterizations of this model, when the authority sets the performance standard for a given teacher at $\hat{p}(h_0, t^*)$, there is a prize associated with this standard that elicits efficient effort from the teacher, and the teacher will win the prize with probability one half. This contest scheme is a rather obvious place to begin, but there may be many other combinations of prizes and targets that also elicit efficient effort from teachers. Consider changing the performance standard by an amount $\Delta$ while choosing a new prize level $\frac{f}{g\phi(\Delta)}$, the first order condition for a teacher choosing optimal effort when facing a contest of the form $\left(\frac{f}{g\phi(\Delta)}, \hat{p}(h_0, t^*) + \Delta\right)$ is satisfied at $t^*$, and for many values of $\Delta$, $t^*$ may remain the teacher's optimal effort choice.

Let $\Omega$ denote the set of all values of $\Delta$ such that teachers choose $t^*$ when facing the contest $\left(\frac{f}{g\phi(\Delta)}, \hat{p}(h_0, t^*) + \Delta\right)$. Given $\Delta \in \Omega$, each contest is associated with an expected

[21] It is well established that one can design two-person contests such that both workers chose efficient effort as part of a pure strategy Nash equilibrium. See Lazear and Rosen (1981). However, these equilibria require that chance play a sufficient role in the outcome of these contests. A contest against the standard $\hat{p}(h_0, t^*)$ is analogous to a game against a machine that always chooses efficient effort, and by making sure that chance plays a sufficient role in determining the outcome of such contests, the authority can ensure that the teacher's best response is to also choose efficient effort.

payoff $\frac{f}{g\phi(\Delta)}[1-\Phi(\Delta)]$. Based on the history of performance pay experiments in public education, I assume that the base salary of teachers is fixed and that any prize money teachers earn from the introduction of performance pay systems is additional income. Further, I assume that the education authority's goal is to minimize the additional payroll cost of introducing a contest scheme that induces efficient effort. Thus, the optimal $\Delta$ minimizes the expected prize payoff $\frac{f}{g\phi(\Delta)}[1-\Phi(\Delta)]$, subject to the constraint that $\Delta \in \Omega$. A complete characterization of the solution to this problem is rather tedious, but several features of the optimal solution are worth noting.

To begin, the optimal prize involves a scaling factor, $\frac{f}{g}$, that parallels the scale factor in our optimal piece rate formula in section 2.1. The issue of scaling is front and center in any piece rate scheme, but the issue must also be confronted in contest schemes. The authority needs to understand how to translate the scale of the performance metric into values of student skill stocks in order to choose prizes correctly.[22]

Turning to the choice of performance standard, the optimal $\Delta$ cannot be negative. Since the authority is considering only contests, $\Delta \in \Omega$, that elicit efficient effort, $\Delta < 0$ implies that teachers win more often than in the $\Delta = 0$ contest. Further, the prize $\frac{f}{g\phi(\Delta)}$ is larger than the prize in the $\Delta = 0$ contest because $\phi(\cdot)$ is maximal at zero. These results imply that the expected cost of the $\Delta = 0$ contest is less than the expected cost for any $\Delta < 0$ contest that ellicits efficient effort.

Although the optimal contest involves $\Delta \geq 0$, the authority must be careful not to choose a $\Delta$ that is too large. If $\Delta$ is too demanding, teachers may find it optimal to choose some $t < t^*$ because the total cost of choosing $t^*$ exceeds the expected increase in prize winnings from choosing $t^*$. For example, let $\Phi(v)$ represent a normal distribution with variance $\sigma_v^2$. Then, it is straightforward to show that $t^*$ is not an optimal response to any contest $\left(\frac{f}{g\phi(\Delta)}, \hat{p}(h_0, t^*) + \Delta\right)$ if $\Delta > \frac{c\sigma_v^2}{fg} = \frac{\sigma_v^2}{g(t^* - \bar{t})}$.[23]

To provide some insight into this condition, note that if $f = 1$ and $g = 1$, then the unit of time used to measure $t$ is the unit such that teachers raise the expected value of a student's human capital by one dollar when they allocate one more unit of effective instruction to the student. Further, the units of $v$ are such that one can think of these shocks to measured

[22] Cunha and Heckman (2008) discuss methods that allow researchers to map test scores for youth into expected values of future adult outcomes like earnings. These methods cannot provide direct evidence on the meaning of scales associated with new assessments unless there are ways to equate the new assessment scales to the scales of tests taken when the current generation of adults was in school. More work is needed in this area to provide better guidance concerning the correct pricing of the psychometric performance measures used in performance pay schemes.

[23] To see that the second order condition is violated for these values of $\Delta$ when $t = t^*$ use the fact that $\phi'(\Delta) = -\frac{\Delta}{\sigma_v^2}\phi(\Delta)$. The optimal choice for $t$ may be greater or less than $t^*$, but cases involving inefficiently low levels of effort are the main concern here.

human capital as the equivalent of deletions or additions to the total amount of instruction they receive. Here, in contests where $\Delta > \frac{\sigma_\nu}{(t^* - t)}\sigma_\nu$, the teacher's second order condition is violated at $t^*$. Thus, if one is willing to assume that the effort innovation $(t^* - \bar{t})$ offsets a one standard deviation shock of bad luck, then the optimal $\Delta \in [0, \sigma_\nu]$.

Even though I have modeled a rather simple contest, a full characterization of the optimal $\Delta$ is beyond the scope of this chapter. The point of deriving bounds on the optimal $\Delta$ for this case is to demonstrate that it takes little effort to construct environments in which education authorities can easily make one of two mistakes by choosing performance standards in an ad hoc way. First, since $\pi(\Delta)$ is the only prize such that the teacher's first order condition is satisfied when she chooses effort $t^*$ in response to a contest against the performance standard $\hat{p}(h_0, t^*) + \Delta$, an authority that began by choosing $\Delta > \frac{c\sigma_\nu^2}{fg}$ would find that no prize exists such that both the first and second order conditions of the teacher's problem are satisfied at $t^*$. It is possible to set standards that are too high in the sense that, given such standards, there are no prizes that elicit efficient effort. The typical outcome in these cases is that the authority chooses a prize level that elicits less than efficient effort, but it is possible that the authority could set a prize so large that teachers supplied more than efficient effort.[24] Second, authorities can set standards that are clearly too low and waste resources relative to the $\Delta = 0$ benchmark. Any contest that results in significantly more than half of the contestants winning a prize is either not eliciting efficient effort or is wasting prize money.

## 2.6. Common Design Flaws

Most performance pay programs adopt prizes and performance standards without conducting any formal analyses of expected responses by teachers, and the prevailing view seems to be that simply providing incentives through standards and prizes should improve effort allocation among teachers. However, the model outlined here raises concerns about ad hoc approaches to the design of performance contests. Contests that may seem reasonable to many can actually be wasteful.

### *2.6.1 Setting Standards Too High*

Political forces often create pressure for "high standards" in education, but these pressures can be counterproductive. Although it is clearly wasteful to set standards too low, standards well beyond $\hat{p}(h_0, t^*)$ may induce no additional effort from teachers.

[24] Further, although teachers in the model above never respond to incentives by choosing $t=\bar{t}$ because the quadratic cost function assumed here imposes a marginal effort cost of zero at $\bar{t}$, any fixed cost associated with adjusting effort away from the norm, $\bar{t}$, introduces the possibility that teachers would respond to excessively demanding standards by staying at $\bar{t}$.

The POINT program (2006–2009) allowed math teachers in grades 5 through 8 in Nashville, TN to volunteer for a performance pay program. The volunteers were randomly assigned to treatment and control groups. Those in the treatment group were eligible for three levels of bonus pay: $5,000, $10,000, and $15,000, dollars. The reward levels were linked to value-added performance targets associated with the 80th, 90th, and 95th percentiles in the historical distribution of student gains on the Tennessee Comprehensive Assessment Program (TCAP). Although these prizes are significant relative to base levels of teacher pay, Springer et al (2010) report that the students in treatment classrooms typically performed no better than students in control classrooms.[25] Further, Springer et al (2010) are not reporting that the gains induced by the program did not generalize. Rather, they report no clear pattern of gains on the high-stakes assessment, TCAP.

It is tempting to say that the POINT results are quite puzzling, given the size of the prizes involved in POINT and the balance of the existing literature on educator responses to incentive schemes. However, it is just as important to note that POINT may have set targets so high that teachers responded optimally by doing roughly what they had done before, $t \approx \bar{t}$. Springer et al (2010) provide an appendix which claims that the expected marginal gains from more teacher effort were likely significant for many teachers. However, their figures suggest that roughly one half of the teachers in the experiment faced less than a twenty percent chance of winning a bonus based on their past record of performance. Although it is quite difficult to determine what the marginal gains to effort were for any of the POINT teachers, it takes little creativity to choose cost and density functions such that the one half of teachers who faced less than a twenty percent chance of winning based on their past performance would have found it optimal to remain at or near $\bar{t}$.

In the model above, all teachers are equally talented and thus share the same cost of effort, and this is likely not true of teachers in the POINT project. However, the presence of teacher heterogeneity only increases the likelihood that *at least some* teachers responded to the system with no change in effort. The estimated treatment effects in the POINT project almost certainly reflect a weighted average of many different changes in teacher effort, but researchers who work with POINT data in the future should carefully investigate the possibility that a significant portion of POINT teachers *optimally* chose not to change their effort levels.

This observation is closely related to the literature on educational triage in accountability systems. Many systems, including the implementation of NCLB in many states, hold all students to a single proficiency standard. However, this "high standards" for all approach often induces teachers to divert resources away from some students who are currently in great need of special attention and who also have no realistic chance of

[25] There was some indication of improvement among fifth graders after year one. However, these impacts did not persist into sixth grade.

reaching proficiency in the near term. Gillborn and Youdell (2000) began this literature with work on the responses of English schools to the structure of national exam systems. Neal and Schanzenbach (2010) document this behavior among Chicago teachers following the introduction of NCLB.

### 2.6.2 Incorrect Handicapping

The POINT system used a simple value-added approach to transform test scores into performance metrics for teachers. In terms of the model above, this procedure is an attempt to condition on $\varphi(h_0)$ when setting standards for teacher performance. It is obvious that all performance pay schemes based on targets must solve this measurement problem. However, it is surprising how often policy makers have adopted rather heuristic approaches that produced less than desirable results.

The MAP system now in place in Florida replaced an earlier system called STAR. The STAR system attempted to assign points to teachers for different possible innovations in reading levels that their students might experience during a given year. These point allocations formed a Value Table with rows for each initial reading level, columns for each terminal reading level, and entries that specified performance points for each possible outcome. The Value Table methodology represented an attempt to make sure that all teachers competed "on a level playing field" as the law required.[26] However, Neal (2009a) reports that the initial results from Hillsborough County provided strong suggestive evidence that the point allocations overstated the relative performance of teachers who worked in affluent schools, and the STAR system was altered and then replaced shortly after its introduction.

The ABC system in North Carolina sets performance targets at the school level and also uses rather ad hoc statistical procedures to attempt to control for baseline differences in school characteristics. Vigdor (2009) reports that this system may also be biased against schools that serve economically and academically disadvantaged students, and Clotfelter et al (2004) report that the introduction of ABC created a dramatic relative decline in the retention rates of faculty in schools serving disadvantaged student populations. These changes in retention rates are quite large, and there is no evidence that these departures were concentrated among weak teachers.

Systems that employ statistical procedures to set performance targets must be implemented with care. Any performance pay scheme that employs a statistical procedure to set performance targets will create incentives for even good teachers to leave their current students if the procedures set performance standards for these students that are too demanding relative to the standards set for others.

[26] The method assigned positive points to student improvements and assigned more points to improvements that are less common. Teachers received point deductions for students who regressed.

### 2.6.3 Using the Wrong Sample

The response of many in the education research community to these observations is that the STAR scheme, the ABC formula, and other ad hoc adjustment schemes are transparently flawed. Advocates of value-added models (VAM) contend that performance metrics without these flaws are available. Consider the following regression model:

$$y_{ijt} = x_{ijt}\beta + D_{ijt}\theta + \varepsilon_{ijt}$$

Here, $y_{ijt}$ is the test score of student $i$ in classroom $j$ at time $t$, and $x_{ijt}$ is a set of student, peer, and resources variables that serve as controls for the baseline growth expected from student $i$. The matrix $D_{ijt}$ includes a set of dummy variables that indicate the assignment of students to classrooms at time $t$, and $\varepsilon_{ijt}$ is an error term that captures shocks to measured performance. Regression models of this form produce vectors $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2 \dots \hat{\theta}_J)$ that contain metrics of classroom performance for all classrooms, $j = 1, 2, \dots J$. Although I use $j$ to index classrooms, $j$ can also index schools in systems where teachers receive bonus pay for team performance. In either case, performance pay systems built around the VAM approach award prizes to the teachers who work in schools or classrooms associated with values of $\hat{\theta}_j$ that exceed some target level.

VAM advocates contend that this approach is the best way to produce performance metrics for educators that correctly control for differences among students in the expected growth in measured achievement attributable to differences in baseline growth among students, $\varphi(h_0)$. However, in order to set appropriate performance targets, policy makers also need to control for the expected measured gains from efficient instruction, $gt^*$, and many implementations of VAM fail to address this second issue.

The most widespread and statistically sophisticated assessment based incentive program in the United States is TAP. TAP involves several components, but the assessment based component involves running a regression like the one above and giving teachers a score of one through five based on their rank in the vector $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2 \dots \hat{\theta}_J)$. Teachers are then rewarded if they earn a score of three or more. If one ignores the rounding procedure, TAP is paying a bonus to teachers with measured performance above the median, and thus some may see this system as analogous to the $\Delta = 0$ contest above, i.e. the scheme that employs $\hat{p}(h_0, t^*)$ as the performance standard. Note that, in the $\Delta = 0$ contest, all teachers choose efficient effort and win the bonus with probability .5.

However, TAP is not analogous to the $\Delta = 0$ contest. The VAM models TAP uses to produce performance metrics for TAP teachers employ data from both TAP and non-TAP schools. Because non-TAP teachers typically work in schools without performance pay systems and because TAP addresses the widely held belief that teacher effort is not efficient in many traditional public schools, it makes sense to assume that many of the teachers in the VAM samples are not supplying efficient effort. Thus, TAP is using VAM in a manner that sets performance standards below the expected value of measured performance given efficient effort, and the analysis above shows that contest schemes built around standards set below $\hat{p}(h_0, t^*)$, i.e. $\Delta < 0$, waste resources.

Nonetheless, the simple model above does predict that teachers will respond to systems with "low" standards if the prizes are high enough, and in a recent study, Hudson (2010) reports that TAP does improve measured student performance. Hudson compares school-wide improvements in test score performance following the introduction of TAP to test score changes in a composite set of control schools that did not introduce TAP, and she finds that TAP does raise math scores and may improve reading scores as well.

### 2.6.4 Holding the Line

So far, I have focused on how difficult it may be for education authorities to specify an efficient system of performance standards and prize payments using standard psychometric performance metrics. However, even if an education authority were endowed with an efficient system at a point in time, the authority would find it difficult to maintain the integrity of its performance standards over time. We have already discussed how coaching on the part of teachers can inflate assessment results, but even in a world with no coaching or gaming, placing the results of different assessment forms on a common scale over time is technically quite difficult, which implies that it is difficult to verify the integrity of psychometric scales as well as any performance metrics derived from them.

Suppose that a political organization representing teachers put hidden pressure on testing agencies to make assessments less challenging over time while scoring them in the same manner. If this organization were successful, the scores associated with various performance targets would correspond to lower levels of teacher effort and actual student skill, and the fact that the performance standards had been compromised would be hard for the public or the education officials that represent them to detect. Those who think this concern is far-fetched should consult the literature on the integrity of proficiency standards under NCLB. A detailed review of this literature is beyond the scope of this survey, but there is considerable evidence that political pressures have compromised the meaning of proficiency scores over time under NCLB.[27]

In addition, the School-Wide Bonus Program in New York City may be an example of problems that arise when changes in exam difficulty compromise performance standards over time. Goodman and Turner (2010) describe an experiment in New York City that began during the 2007–2008 school year and continued through the 2008–09, school year. Schools could earn bonuses of either $1,500 per teacher or

[27] See Cronin et al (2007) and Neal (2010).

$3,000 per teacher if they met targets for school improvement scores. The improvement scores were weighted averages of measures of progress in student achievement and measures of school environment factors such as attendance and safety. The program required schools that began at lower performance levels to meet higher improvement targets in order to win a bonus. The program involved competition at the school level, but compensation committees at each winning school divided the bonus monies among teachers, and in many instances, the committees exercised their discretion and deviated from an equal sharing rule.

Because the program was announced in the middle of the 2007–2008 school year, 2008–2009 was the first year that the program was in place before the school year began. In 2008–2009, 135 of 152 treatment schools (89%) won the maximum prize of $3,000 per teacher, and there is scant evidence that the program had any positive impacts on student achievement. The program is quite complicated, and it was layered on top of the New York City accountability system. The failure of the program to impact student achievement may reflect confusion about exactly how the program worked as well as the fact that many schools in the treatment and control samples already faced significant performance pressures from the accountability system. However, there is another possibility. It is widely believed that the state assessments used to generate student achievement and school performance measures became easier over time starting in 2007 and that the scoring and scaling of these assessments did not reflect these changes in exam difficulty. It is almost impossible to know the extent to which teachers were aware of this trend, but ex post, the program operated almost like a change in base pay. Although treatment schools did not perform better than control schools, more than 91% of treatment schools won a bonus, and 89% of treatment schools won the maximum prize.

### 2.6.5 Subjective Targets

Private-sector firms also face difficult performance measurement issues, and these firms rarely rely solely on statistical methods to solve these problems. Instead, schemes that link rewards to subjective performance evaluations are common in the private sector. Several of the entries in Table 6.1 describe systems that link performance pay for educators to subjective evaluation schemes, and despite their prominence in private industry, the results of these schemes in public education are not impressive.

The Performance Related Pay (PRP) system in England involves two forms of bonus pay for teachers who have already reached the maximum pay level in the standard pay scale. The first is a permanent increase in base salary. The second involves future opportunities to move up to even higher levels of base pay dictated by an extended salary schedule. Atkinson et al. (2009) examine the performance of eligible versus ineligible teachers following the introduction of this system in 1999, and they find no significant effects of eligibility on student performance. When they attempt to correct these estimates for experience differences between eligible teachers and non-eligible teachers, their results imply that the program increased teacher performance in science but may have harmed teacher performance in math and English.

In sum, there is little clear evidence that the PRP system improved instruction in English schools, and given the process that determined awards, some may not be surprised by this result. The initial cohort of eligible teachers were those who were already at the top level of the standard pay scale. These teachers applied for a permanent increase in base pay and movement to a new promotion and salary schedule by submitting cases which contained evidence that their "students are making progress as good or better than similar students nationally." Wragg, et al. (2001) report that 88% of eligible teachers applied and 97% of those who applied received the award. Unless the returns to teacher experience in England are quite exceptional, the officials who reviewed these cases adopted a lenient interpretation of "as good or better."

Martins (2009) describes a similar performance pay program implemented in Portugal in 2006–2007. This program linked promotion to a higher salary schedule and one-time cash prizes to individual teacher performance evaluations. These evaluations were supposed to consider the performance of students on internal and external exams, feedback from parents on teacher performance, attendance records, and participation in research and professional development activities. However, these evaluations were not conducted by independent third party inspectors. Martins writes that "criteria for progression (promotion and prizes) were to be assessed at each school, by those teachers (already) in the higher pay scale."

Using private school students and students on Portuguese Islands,[28] Martins finds that student exam scores on internal tests remained flat or fell slightly following the reform, and scores on national exams fell substantially. Martins does not have an experimental control sample, but the results he reports are so negative that it is difficult to believe that the Portuguese system produced any real achievement gains for students, and students may have been harmed.

In private firms, the person who evaluates a worker's performance is either an owner of the firm or an agent of the owner. In public education, subjective performance evaluation is more problematic because many principals and administrators work under employment and salary rules that create only weak links between the quality of their personnel decisions and their own compensation. Thus, some may not be surprised that performance pay systems that involve one group of public employees making subjective determinations about the bonus payments given to another group of public employees did not generate noteworthy gains in student achievement.[29]

[28] Azores and Madeira implemented weaker versions of the performance pay reforms.

[29] See Prendergast (1999) for a discussion of problems that may arise in subjective evaluation systems within large private organizations if agency problems exist between managers and owners.

Still, the English and Portuguese systems are not unique. Many of the entries in Table 6.1 involve systems in which educators are involved in creating the performance standards that they or their coworkers are required to meet in order to earn a bonus. The ProComp system in Denver, the Qcomp system in Minnesota, and the MAP system in Florida all involve district or school level discretion in defining the performance standards that determine performance pay. These programs have not been formally evaluated, but one must worry that these systems may morph into vehicles for raising the base pay of most or all teachers whether or not these teachers improve their performance.

### 2.6.6 The Value of Relative Performance Systems

Education officials can avoid some of the problems highlighted thus far in section 2.6 if they commit to performance pay schemes that are true relative performance systems. In relative performance schemes, there is a fixed amount of prize money set aside, and all of the prize money is distributed to some worker or workers ex post based on relative performance comparisons among the workers. The reliance on relative performance measures means that some teachers will win and others will lose by construction. Thus, there is no way to manipulate these systems so that every worker receives a bonus even if no worker improved their performance. It is quite difficult to convert relative performance schemes into changes in base pay through corruption activities, whether the activities involve corruption of psychometric standards or manipulation of subjective performance evaluations.[30]

Further, relative performance schemes can provide information that the education authority needs to maintain the value of incentive schemes over time. Even if an authority knew the level of measured performance associated with efficient effort at a point in time, developments in pedagogy, changes in assessments, or contamination of performance metrics may cause this level to rise or fall over time. In some environments, the authority can use movements in average measured performance to infer how levels of measured performance associated with efficient effort are moving over time. Competition among teachers in relative performance systems may provide valuable information about the levels of measured performance that are associated with efficient classroom effort. Thus, VAM methods on samples of teachers who all face the same incentive system may create adequate control for both student differences in expected baseline achievement growth and the efficient levels of instruction that the system induces teachers to allocate to students.[31]

[30] If all teachers could collude on low effort, then the prizes would be handed out each period based on measurement error and each teacher would enjoy an increase in expected base pay without changing their effort. However, it seems unlikely that teachers in an entire school district or state could maintain such collusion.

[31] See Holmstrom (1982). Barlevy and Neal (2011) describe specifically how this insight applies to the design of incentive systems for educators.

Table 6.1 describes three systems that involve both competition among educators for a fixed set of prizes and the use of VAM methods to rank schools or teachers. Ladd (1999), Lavy (2002), and Lavy (2009) all contain evaluations of experimental performance pay schemes. Ladd (1999) describes a system implemented in Dallas in 1991. The Dallas system was a tournament among schools. Schools received performance scores based on estimates of average value added in the school as well as measures of attendance and dropout rates. The VAM estimates of school performance employed scores from several different assessment systems, and the procedure produced measures of relative school improvement in performance. Each year, about 20 percent of the schools won performance bonuses. All staff in winning schools received a bonus. Principals and teachers received one thousand dollars.

Lavy (2002) describes a tournament among secondary schools in Israel that took place in 1995–1997. Here, secondary schools received performance scores based on estimates of their contributions to improvement in three areas: credit units per student, the fraction of students receiving a matriculation certificate, and the school dropout rate. The top one third of schools received awards that varied with the overall performance ranking of the schools. The largest prize resulted in bonuses for teachers that equaled roughly five percent of the starting salary for a new teacher. The smallest prize generated bonuses that were one fourth as large.

Lavy (2009) describes a tournament among individual Israeli secondary school teachers in 2000–2001. Individual teachers received performance scores based on the average score of their students on the matriculation exam and their students' pass rate. Teachers who taught the same subject competed against each other. Further, because the regression models used to produce relative performance measures included school fixed effects, teachers were competing against other teachers in their school and were rewarded for being exceptional relative to their peers. The program ranked teachers according to pass rate performance and average score performance and used a point system to form an aggregate ranking. The pass rate score contributed more to the overall teacher ranking. Winners received performance pay bonuses based on their total performance index, and the top performers received large bonuses.

None of these programs involved random assignment of schools or teachers to treatment. Thus, the authors employ several empirical strategies that attempt to pin down the causal impacts of these programs. Although some may quibble with the details of any one of these three papers, the results taken as a whole paint a fairly consistent picture. All three papers find that these programs generated significant increases in measured achievement among students, but all three also report significant heterogeneity in estimated treatment effects for different sub-populations. Ladd (1999) reports that the Dallas program generated large gains for white and Hispanic students but not for Black students. Lavy (2002) and Lavy (2009) find that both Israeli programs generated larger improvements among students with lower baseline performance as well as students from less educationally advantaged families, but Lavy also notes that both programs included design features that generated stronger incentives for teachers to direct relatively more attention to weak students.

None of these three papers have access to the type of low-stakes assessment data required to make definitive statements about the generalizability of the measured gains induced by these programs. However, the Dallas system may have been more difficult than many to game because it involved test data from several different assessment systems as well as measures of attendance and dropout rates. Further, Lavy (2009) presents evidence that the Israeli program induced substantial changes in teacher effort and pedagogy.

All of these systems represent components of experimental programs. I know of no ongoing large scale performance pay systems in education that are true relative performance pay schemes. This outcome may reflect the fact that teachers and their unions recognize that relative performance schemes cannot be manipulated into systems that simply change base pay for all teachers.

### 2.6.7 Aggregation

Although the programs described in Ladd (1999), Lavy (2002), and Lavy (2009) appear to have worked fairly well, the tournament structure of these programs raises important implementation questions. In a world where each teacher has only one student, tournaments would be relatively easy to implement. One could define leagues based on baseline student characteristics, and the within-league rank of each student would determine whether or not his teacher won a prize.

However, because teachers and schools work with many students at one time, the construction of performance rankings based on assessment data is not so straightforward. Imagine a setting with assessments that produced perfectly reliable measures of student skill. Further, suppose one teacher had two students who both began the year with a math score of 150 and then ended the year with scores of 155 and 160. Finally, suppose another teacher had two students who began the year with scores of 100 and 200 respectively and ended the year with scores of 110 and 205 respectively. Based on such data, how could one rank the performance of the two teachers without understanding the values to society of bringing students from 100 to 110, 150 to 155, 150 to 160, or 200 to 205?

The VAM methods used in all three experiments assert that our two hypothetical teachers performed equally well simply because the average score improvement in both hypothetical classrooms was 7.5. The experiments in Dallas and Israel took the average of VAM residuals to create performance ranks for classrooms and sometimes schools, and one must ask when averages that are expressed in units of a particular psychometric scale provide valid rankings of total performance for schools or teachers. These averages provide valid rankings if the VAM model is correctly specified and if scores on a given psychometric scale are a fixed affine transformation of the social value of the underlying skill levels associated with various scores. Put differently if $p_{ijt} = y_{ijt} = ah_{ijt} + c$, where $y_{ijt}$ is the test score for student $i$ in class $j$ in period $t$, $h_{ijt}$ is the social value of this student's skills at the end of period $t$, and $a > 0$ and $c$ are constants, then VAM rankings of classroom of school performance will be accurate.

Yet, if an education authority could create a psychometric scale with these magical properties, then pay for performance schemes based on piece rates must be considered as serious policy options.[32] The absence of piece rate schemes in practice may reflect many factors, but I conjecture that a key factor is that the use of piece rates would focus attention on the fact that education authorities do not know whether or not a teacher who moves a child from 150 to 155 on a given developmental scale is creating greater, lesser, or equal social value than a teacher who moves a child from 200 to 205. But, if this is one reason that we do not observe piece-rates schemes based on VAM estimates of teacher performance metrics, there is no reason to accept VAM rankings as ex post performance rankings that determine the allocations of prizes in a tournament. Many VAM estimators are quite complex, and the literature contains lengthy debates about the relative value of different VAM approaches, but the results from all VAM models are sensitive to the psychometric scaling of assessment results, and this fact should give advocates of these models pause.[33]

Further, in some contexts, the literal interpretation of VAM performance rankings indicts the whole enterprise. Imagine two fifth grade math teachers in a large district. Both are supposed to take their students as far as they can through a common curriculum, but one teacher works with children in a disadvantaged school who began elementary school not knowing how to count and the other teaches in a selective magnet school designed for gifted children. Now, assume that the test score results from both teachers' classes are part of a state or district wide sample used as inputs into a VAM model that produces a vector $\hat{\theta}$ which contains a performance measure for all fifth grade math teachers in the district. The elements of $\hat{\theta}$ associated with our two hypothetical teachers are supposed to tell us which teacher performed better during the year or at least which teacher one should expect to have performed better. However, these two teachers did not do the same job because they worked with students who were at completely different places in their academic development, and thus it seems almost nonsensical to ask which teacher did better. Functional form assumptions and the assumption that the units of a given psychometric scale serve as a welfare index allow VAM to rank the performances of these two teachers, but the fact that some

[32] Many tournament schemes, like those in the Israeli and Dallas experiments, cannot elicit first best effort from all participants unless all teachers are equally talented, but piece-rate systems are efficient even in the presence of worker heterogeneity.

[33] See Briggs and Betebenner (2009), Briggs and Weeks (2009), Reardon and Raudenbush (2009) for more on this issue.

applications of VAM provide clear answers to nonsensical questions should be a source of concern for VAM advocates and not a selling point for VAM methods.

## 2.7. Steps Forward

In the previous sections, I described how hidden actions like coaching contaminate the information in high-stakes assessments, and I also discussed how hidden manipulations or subjective determinations of performance targets may transform performance pay schemes into increases in expected base pay for teachers without commensurate changes in teacher effort. Finally, I discussed the benefits of performance pay schemes based on measures of relative performance but noted the problems that may arise when policy makers create performance metrics that depend on the implicit assumption that particular psychometric scales serve as proxies for welfare indices.

In recent work with Gadi Barlevy, Barlevy and Neal (2011), we describe a performance pay scheme for educators with the following properties: (i) educators compete against each other for a fixed set of prize money (ii) reward pay is based on rankings of individual student outcomes. No measure of classroom or school output is involved, and no composite ranking of educator performance is created (iii) the mapping between student assessment results and the performance pay given to specific teachers is invariant to the scale used to report assessment results, and (iv) because the system is scale invariant, it can be implemented using a series of assessments that contain no repeated items and no common format, which removes opportunities for teachers to coach students concerning particular formats or items used in previous assessments.

The system we propose is called "pay for percentile" and it works as follows. Consider the population of students taking fifth grade math in a state or a large school district. At the beginning of the year, place each of these students in a comparison set that contains other students with similar records of academic achievement, common family backgrounds, and similar peers. Then, at the end of the school year, give each student a percentile score that describes the fraction of students in his comparison set that performed less well than he did. Average these percentile scores over all the fifth grade math students in a given classroom or school and call this average a percentile performance index. This index is a winning percentage. It tells us how often students in a given unit perform better than students in other units who began the year at the same achievement levels. Finally, pay educators bonuses that are proportional to their percentile performance indices.[34]

[34] Classroom size and the efficient prize in a standard two-person contest determine this constant of proportionality. The Barlevy and Neal (2011) framework extends the two-contestant, single-output tournament model of Lazear and Rosen (1981) to a setting with many contestants and many distinct but jointly produced outputs. In the context of education, the human capital acquired by each student is a distinct output, but the set of outputs produced in the classroom are produced jointly by choosing a vector of time allocations to different tasks, e.g., lesson planning, lecturing, small group instruction, and individual tutoring.

Note that this system relies only on the ordinal information in assessment results, and because only ranks within comparison sets matter, this system does not require and never produces a measure or ranking of overall educator performance. All students compete in seeded contests against students in other schools, and performance pay for educators is determined by the overall winning percentage of their students in these contests. Even though some teacher actions, e.g. lesson planning, group tutoring, classroom lectures, simultaneously affect the expected contest outcomes for many of their students, we show that such a scheme can elicit efficient effort from all teachers on all tasks that create human capital in their students.

Because pay for percentile employs only information concerning relative ranks, it provides no information that allows education authorities to understand how student performance is evolving over time or how the performance of a school is evolving over time. However, as I argue above, separating incentive provision and performance measurement eliminates incentives for educators to take actions that contaminate performance measurements. Education authorities can always measure progress in student achievement using parallel assessment systems that involve no stakes for educators and also contain the overlap in item content and format that make proper equating possible. By making this system a no-stakes system, education authorities remove incentives for educators to engage in the coaching and manipulation activities that currently contaminate the information produced by many accountability systems.

### 2.7.1 Team Competition

Lavy (2009) reports some positive effects of an incentive scheme that forces teachers to compete against other teachers in the same school, and Muralidharan and Sundararaman (2010) report that the incentive scheme that linked piece rates bonuses to individual teacher performance in India generated larger measured achievement gains than the scheme that paid team piece-rates. While some may be tempted to conclude that individual incentives are important as a means for overcoming free rider problems, there are benefits from implementing pay for percentile as a team competition rather than competition among individual teachers. Although the experimental results appear positive, systems like the one Lavy (2009) describes could create serious problems if implemented as permanent policies.

The presence of school fixed effects in the Israeli VAM models used to create teacher performance measures implies that the performance of each teacher is being measured relative to the average performance of teachers in her school. This convention creates a clear incentive for teachers to sabotage the work of their peers. Sabotage may not have been a problem in a short-lived experiment where teachers may or may not have fully understood the construction of performance metrics. However, the Jacob and Levitt (2003) results suggest that one should not assume that teachers are unwilling to engage in such behaviors when permanent incentive schemes create clear incentives for such malfeasance.

Systems that involve individual incentive pay but no direct competition among teachers working in the same school are less problematic, but education authorities may still prefer to have teachers compete in teams. The persons who may possess the best information about how a particular fifth grade math teacher in a given school can improve are the other fifth grade math teachers in the same school. Incentive systems should encourage these teachers to share this information rather than withhold it. Thus, it makes sense to allow all the teachers who teach a given subject in a particular grade to compete as a team against teachers in other schools that serve similar communities and students. These teams are often so small that free riding should not be a huge concern and peer monitoring should be quite effective. The majority of incentive schemes described in Table 6.1 are team-incentive schemes, and all of the team incentive plans did generate improvements in measured achievement.

There are also statistical reasons to prefer inter-school rather than intra-school competition. Barlevy and Neal (2011) discuss how existing methods in educational statistics can be adapted to estimate percentile performance indices, and a key assumption in these methods, and other methods used to create educational performance metrics, is that the conditioning sets that define league competition are so rich that one can treat the assignment of teachers to students as random given these conditioning variables. It may be easier to satisfy this requirement when performance pay contests involve only inter-school competition. Rothstein (2010) presents evidence from North Carolina data that, within schools, unobserved dimensions of student aptitude affect the allocation of student among classrooms, and it makes sense that this would be the case. In order to maximize the human capital created in their schools, principals must use all the information at their disposal to make optimal matches between students and teachers. Furthermore, any system that asks teachers within the same school to compete against each other may create resistance from some teachers to accept the students who should optimally be assigned to them. However, at the school level or grade level within a school, every student must be assigned to some teacher, and inter-school competition for team bonuses creates incentives for teachers and principals to make sure that students are assigned optimally among teachers.

While it is true that there may still be concerns about selection among schools by parents, it may be possible when implementing performance pay schemes at the level of a state or country to form leagues for schools to compete in such that schools are well matched on the measured characteristics of students, communities, and parents, and no two schools in the same league serve geographic areas that intersect. Given this arrangement, no parents would have chosen their child's particular school over any of the other schools in their school's league, and concerns about selection into schools on unobserved family traits may be less severe.

### 2.7.2 Limitations of Assessment-Based Incentives

The design of pay for percentile removes opportunities for teachers to coach students for upcoming assessments based on the specific items and format found in previous assessment. Further, this scheme avoids many thorny issues that arise when education authorities attempt to build performance pay systems that are dependent on the scaling of psychometric performance measures. However, any assessment-based performance pay scheme for educators will create alignment problems, and pay for percentile is no exception. Educators still benefit from cheating, e.g. giving students answers during the exam. Further, assessment-based schemes do not reward teachers for building non-cognitive skills that are not assessed.

Concerns about cheating can potentially be addressed by mandating that all assessments be monitored by third party testing agencies, but concerns about teachers diverting effort away from the development of important social and emotional skills must be addressed by building systems that reward teachers for contributing to their students' non-cognitive development. Many of the systems described in Table 6.1 are systems involving multiple components, and while I have focused on the assessment-based components of each program, the presence of other components is an important design issue. Many reasonable social welfare functions imply that the optimal set of personnel policies for educators should create incentives for teachers to foster both the cognitive and non-cognitive development of their students. In the next section, I will discuss a strategy for eliciting information from parents concerning the performance of educators with regard to the social and emotional development of children.

### 2.7.3 Heterogeneity

All incentive pay schemes in education that are built around statistical performance metrics appear to be designed as mechanisms for eliciting effort from a homogeneous group of teachers. The objective incentive schemes described in Table 6.1 involve statistical targets that are the same for all teachers holding constant the characteristics of their students. Further, the tournament schemes employed in Israel and Dallas involve no handicapping. Given student characteristics, all teachers compete on equal footing. Pay for percentile is similar.

However, if teachers differ in the talent levels, one common set of performance standards cannot elicit efficient effort from all teachers. Further, simple tournament schemes typically do not elicit efficient effort from heterogeneous contests without some handicapping system.

Thus, if the education authority can observe teacher characteristics that serve as exogenous proxies for effective talent, then the authority can improve efficiency by seeding contests based not only on student characteristics but also on these teacher characteristics as well as measures of resources within the classroom that may affect teacher effectiveness.[35] If this seeding process creates competition among teams of teachers such that teams who compete against each other have symmetric beliefs about their true talent levels, then there will exist prizes such that these seeded contests elicit efficient effort from all teachers. However, if some teams of teachers know that they are either better or worse than the typical team of teachers that shares their characteristics, then more elaborate mechanisms are required.[36]

Some may advocate piece-rate schemes as a strategy for inducing efficient effort from heterogeneous teachers. While I have already noted that this approach requires that education authorities translate an entire psychometric scale into monetary units, another implementation concern may be even more important. Tournament schemes can be implemented using a fixed amount of money that the authority introduces as an addition to total teacher compensation. Thus, tournaments allow existing teachers to know that they will not receive wage cuts following the introduction of incentive pay, and they allow education authorities to know ex ante exactly how much the incentive scheme will cost.

These features are attractive politically, but no piece-rate scheme can provide both of these features at once. In piece-rate schemes that involve relative pay for performance, teachers who perform well below average must receive salary reductions, and it is possible that those who perform at the lowest levels would owe performance fines in excess of their base salaries. This observation may offer insight into the fact that none of the systems described in Table 6.1 involve piece rates linked to relative performance measures.

The two piece rate schemes in India and Arkansas link performance pay to absolute measures of teacher output. These schemes guarantee non-negative bonuses for all teachers. However, these programs create the possibility that total prize winnings will exceed the budget an authority has set aside ex ante. Further, although both programs were experiments that lasted only a few years, any absolute piece rate scheme implemented as a permanent policy would invite the corruption and cheating activities expected in all scale dependent incentive systems, and these activities could generate significant growth in total bonus pay over time even if the distribution of teacher performance remained fixed over time.

[35] Examples include class size, the presence of a teacher's aide, teacher experience, computer resources, etc.

[36] Barlevy and Neal (2011) discuss how heterogeneity in teacher talent affects the properties of pay for percentile and other tournament schemes. Several authors have proposed more complex tournament schemes that address heterogeneity directly but are also more difficult to implement. O'Keeffe et al (1984) and Bhattacharya and Guasch (1988) present contest schemes that involve heterogenous contestants selecting the measurement rules and payoff rules that they will compete under.

### 2.7.4 Back to Screening

I began section 2 by looking at models of screening in which teachers supplied effort inelastically but enjoyed different levels of talent, but most of section 2 implicitly addresses settings where teachers are homogeneous with respect to their talent levels, or at least homogeneous given a set of observed characteristics, and the goal is to design performance schemes that elicit efficient effort. The agenda for future research in this area should be the design of systems that dictate seeded relative performance contests at each stage of a teacher's career while permitting the entire history of winning percentages in these contests to affect not only performance bonuses but also base pay, pension benefits, retention decisions, and the seeding of future contests among remaining teachers. It is not clear how well education authorities can do if they seek to design systems that both screen and provide incentives. The dynamic aspects of such systems create new complications because teachers know that performance today may not only affect compensation today but also whom they compete against in the future. Further, team incentive schemes are useful for encouraging effective co-operation within schools, but measures of individual teacher performance may be most useful for retention policies. In sum, the existing economics of education literature contains considerable research on the construction of methods for evaluating the impacts of performance pay systems or other incentive systems in education, but the literature on the design of these systems remains quite small and limited in scope, and there is much work to be done.

## 3. MARKETS

I note above that, even if pay for percentile or some other assessment based incentive scheme can be used to induce all teachers in publicly funded schools to teach their students in ways that promote mastery of the topics specified in a common curriculum, most parents and public officials want teachers to be more than conduits of academic information. Parents want their children to feel safe at school, and they want their children to develop emotionally and socially as well as cognitively. Thus, even if education officials develop an assessment based incentive scheme that induces teachers to teach well, they must also address the concern that schools will spend too much time on academics at the expense of the social and emotional development of children.

This observation implies that assessment based incentive schemes can never be more than one component of the incentive systems that publicly funded schools face. However, it is not obvious how education officials should develop incentive schemes that direct the efforts of educators regarding the non-cognitive development of children. It is not at all clear that education officials will ever be able to design assessments of non-cognitive skills that are both extensive enough and reliable enough to use as a basis for incentive pay.

In the absence of systems that directly assess non-cognitive skills, education authorities need to consider indirect mechanisms. Although many education policy debates frame assessment based accountability and expansions of parental choice as opposing alternative mechanisms for eliciting better performance from publicly funded schools, I have written in Neal (2009a) that these policies are best seen as complements. Once policy makers recognize that assessment-based accountability proposals, almost by definition, ignore non-cognitive skill development, it is natural to consider these questions: Who possesses good information about the non-cognitive development of children, and who faces strong incentives to truthfully report information they possess about the non-cognitive development of children? "Parents" is a good answer to both questions, and the value of voucher systems, charter school expansions, and other policies that expand school choice is that they provide a means of enlisting millions of parents as performance monitors. Further, education officials can induce these performance monitors to reveal what they are observing using relative simple market mechanisms.

Three recent papers, Barrow and Rouse (2009), Figlio (2009), and Neal (2009b), review the literature on the effects of private schooling and the effects of access to private schools through voucher programs in particular. Three important conclusions stand out as themes concerning the impacts of vouchers in developed countries. First, the measured cognitive benefits of access to private schools through voucher programs are often modest. Second, the effects of voucher access on parental and student satisfaction are often large. Third, access to private schools often creates substantial gains in total education attainment.

Given the existence of at least three recent survey papers on this topic, I will not provide another literature review here. However, I do note that the literature as a whole implies that vouchers often allow parents to find schools for their children that are better matches on dimensions other than academic quality, and better matches apparently lead to more attainment. If parents do possess the ability to evaluate important non-academic aspects of school performance, then it makes sense to consider mechanisms that provide incentives and opportunities for parents to use their evaluations in ways that shape the behavior of educators who receive public funds.[37]

[37] Further, there is evidence that private schools offer an even broader set of benefits for students in developing countries. Andrabi et al (2010) examine outcomes for private school children in Pakistan. They do not have a voucher experiment that generates random variation in private school access, but they do build an instrumental variables strategy by exploiting interactions between the location of families, the location of public schools, and the historic pattern of settlement in rural villages. They find enormous positive effects of private schooling on achievement even though public schools are funded at much higher levels. Angrist et al (2006) report results from a voucher experiment in Colombia. The vouchers covered roughly half of the cost of private schooling and were assigned by lottery. The study used comparisons between lottery winners and losers to estimate the impacts of being offered access to private schooling. The implied achievement gains associated with private school access were large, and the authors conclude that the implied increase in expected adult earnings among recipients likely exceed the cost of the program.

Neal (2009a) outlines a framework for designing systems that distribute public funds among schools that combines features of assessment-based accountability systems and voucher systems. In this framework, all schools, both private and public, compete on multiple dimensions for public funding. Student assessment results, the results of school inspections, feedback from parents, and parental choices affect whether or not a given school is eligible to receive funding and the level of funding it receives in a given year. Much more work is required before researchers can offer specific guidance concerning the optimal mapping from these varied signals of school performance into the funding levels enjoyed by schools, but assessment based performance pay and vouchers may work well together in systems that require schools to compete for public resources on all relevant dimensions of school performance.

By creating competition among schools for students and public resources, such a system also creates competition among schools for teachers. I noted above that subjective performance pay schemes have produced questionable results in public education, and this presumably reflects the fact that educational administrators are not always penalized when they give raises or promotion to undeserving teachers. However, in a managed competition framework, all the teachers in a school as well as the administrators in the school know that the future capacity of the school to provide higher salaries for its employees is directly influenced by the quality of its personnel policies. The best solutions to the screening and incentive provision problems described above may arise as byproducts of a system that forces schools to compete for the public support they receive. A competitive market for teachers allows schools to build reputations as employers that reward teachers for excellent performance on all dimensions and also allows teachers to benefit from building their own personal reputations.[38]

Nonetheless, Neal (2009b) points out that, while many countries now have systems that operate like voucher systems and force schools to compete for students, no developed country with a large voucher system allows schools to compete for teachers by following different personnel policies. Systems that force schools to compete for public funding but also force all schools that receive public funding to hire, train, reward, and fire teachers according to a fixed set of personnel policies are incoherent from a design perspective. In any industry, increased competition among firms offers the possibility that the firms which remain in the market going forward will be those who have successfully adopted new and more efficient means of production. Teachers are the key input in educational production. Thus, policies that govern the hiring, training, retention, and motivating of teachers should have large impacts on the efficiency of schools. It makes no sense to promote competition among schools for students while restricting how schools may compete for teachers.

[38] See Hoxby (2002) for more on how competition for teachers could affect who teaches and how.

## 4. CONCLUSION

Current research in the economics of education devotes considerable attention to the methods that researchers use to evaluate the impacts of various innovations in public education policy. It is appropriate that researchers devote great energy to the tasks of discovering what works best and developing methods that actually help us discern what works best. However, economists should begin contributing more to debates among scholars and policy makers concerning how performance pay programs are designed before they are ever implemented and evaluated.

Most of the programs reviewed here provide some evidence that teachers responded to performance pay schemes by changing their effort allocations in some way, and in many cases, there is at least strong suggestive evidence that total teacher effort rose following the introduction of performance pay. Two of the exceptions to this rule are the bonus schemes in England and Portugal that relied on subjective assessments made by either education officials or peer teachers. Ex post, these programs appear to have been vehicles for increasing the baseline pay scale of experienced teachers without requiring improved teacher performance. Whether or not this outcome was anticipated by the political champions of these programs, the lesson taught by these programs, and a larger literature on performance pay in other organizations, is that subjective bonus schemes should not be expected to work well unless they are part of a larger incentive system that provides incentives for those who make subjective performance evaluations to make these evaluations accurately.

The POINT program also stands out as a program that generated few measurable impacts, but the lesson that POINT teaches is different. The performance standards in POINT are completely objective. However, it is not clear that these standards were set at levels that make efficient incentive provision possible. While there may be other plausible explanations for the POINT results, the simple model developed in section 2.5 highlights the possibility that POINT simply set the performance standards too high. Further, whether or not this is true in the case of POINT, the theoretical results from section 2.5 provide an important warning for those who design incentive schemes around psychometric performance targets. It is simply not true that education authorities can choose performance standards in an ad hoc manner and then experiment with different prize levels until they discover a prize level that will elicit efficient effort given their initial choice of standard. Given some performance targets, there is no prize level that would induce efficient effort.

Concerns about the choice of performance standards as well as the manipulation of performance standards can be mitigated to some extent if education authorities require that all incentive schemes involve pay for relative performance. When authorities force educators to compete for a fixed amount of reward money, well designed contests can reveal the expected level of measured performance that is associated with efficient effort levels among teachers. When officials allow competition to determine standards endogenously, they make it difficult for educators or their representatives to compromise performance standards or prevent standards from rising over time as new technologies and teaching methods make better performance possible.

Nonetheless, even in relative performance schemes, manipulation of performance metrics remains a concern. Although relative performance schemes weaken incentives for collusion among teachers, they may induce wasteful forms of competition. Educators in relative performance contests may take actions that are privately beneficial because they raise measured relative performance but socially wasteful because they crowd out teaching activities that create more lasting skills among students.[39] The literature suggests that educators often respond to assessment-based incentives by coaching students for specific assessment items or item formats. In fact, studies that examine scores on both high and low stakes assessments for the same population of students offer no evidence that any incentive scheme induced changes in measured performance on high-stakes assessments that even come close to fully generalizing to low stakes assessments of the same material.

Thus, it seems obvious that a key task for those who design future performance pay schemes for teachers is the creation of a series of assessments that consistently cover a well specified curriculum but vary substantially in terms of specific item content and format. Put more pointedly, the designers of assessment-based incentive schemes must take seriously the challenge of designing a series of assessments such that the best response of educators is not to coach but to teach in ways that build true mastery of the intended domain.

Many existing performance pay schemes cannot employ results from such a series of assessments because these systems are built around a particular psychometric scale, and it would typically not be possible to place results from assessments of varying formats on a common psychometric scale. However, ordinal contests like the pay for percentile scheme described in section 2.7 can employ the results from such assessments, and a commitment to ordinal contests and tests without repeated items and formats could go a long way toward eliminating the coaching and test preparation responses that appear to plague many current and previous systems.

This observation is related to the most obvious lesson generated by the material presented in section 2. Education authorities cannot reasonably expect to obtain reliable information about secular trends in performance from assessment series that are part of

[39] Further, Barlevy and Neal (2011) point out that although general score inflation does not benefit teachers who compete in a relative performance scheme, teachers as a group can still benefit from manipulating the dispersion of scales. If teachers can collectively pressure testing agencies to compress the distribution of performance metrics, after piece-rates have been set, the contaminated system will provide weaker incentives but pay out the same total prize money to teachers.

incentive systems. Systems that provide reliable information about secular trends in performance must involve assessments that can be properly equated over time, but the overlap in content and format that makes proper equating possible creates opportunities for the coaching behaviors that inflate scores and compromise the meaning of assessment scales. If education officials desire credible measures of secular progress, they must obtain these measures from a series of assessments that contain no stakes for educators.

Finally, because taxpayers and their representatives want schools to build non-cognitive as well as cognitive skills, assessment based incentive schemes can never be more than one component of a broad system of incentives for educators. From this starting point, it is clear that assessment based incentive schemes and voucher systems should not be seen as policy substitutes but rather policies that may work well together as part of a broader system that requires schools to compete on several dimensions for access to government funds. By fostering competition among schools that rewards schools for fostering both the cognitive and non-cognitive skills of children, education authorities may create competition among schools for effective teachers that spurs innovation in the creation of new methods for screening, developing, and rewarding teachers.

## ACKNOWLEDGMENTS

I thank participants at the 2009 Handbook of Economics of Education Conference in Munich for helpful feedback. I thank Gadi Barlevy, Robert Gibbons, Canice Prendergast, and Phil Reny for useful conversations and guidance. I thank Luca Maini, Eric Nielsen, Armin Rick, and Richard Olson for excellent research assistance. I thank the Searle Freedom Trust for research support. I also thank Lindy and Michael Keiser for research support through a gift to the University of Chicago's Committee on Education.

## REFERENCES

Adams, G.J., 1996. Using a Cox Regression Model to Examine Voluntary Teacher Turnover. J. Exp. Educ. 64 (3), 267–285.

Andrabi, T., Bau, N., Das, J., Khwaja, A.I., 2010. Test Scores and Civic Values in Public and Private Schools. mimeo.

Angrist, J., Bettinger, E., Kremer, M., 2006. Long-Term Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia. Am. Econ. Rev. 96 (3), 847–862.

Atkinson, A., Burgess, S., Croxson, B., Gregg, P., Propper, C., Slater, H., Wilson, D., 2009. Evaluating the Impact of Performance-related Pay for Teachers in England. Labour Econ. 16 (3), 251–261.

Baker, G., 2002. Distortion and Risk in Optimal Incentive Contracts. J. Human Resour. 37 (4), 728–751.

Ballou, D., Podgursky, M., 1997. Teacher Pay and Teacher Quality. W.E. Upjohn Institute for Employment Research, Kalamazoo, Michigan.

Barlevy, G., Neal, D., 2011. Pay for Percentile. University of Chicago, forthcoming American Economic Review.

Barrow, L., Rouse, C.E., 2009. School Vouchers and Student Achievement: Recent Evidence, Remaining Questions. Ann. Rev. Econ. 1, 17–42.

Bhattacharya, S., Luis Guasch, J., 1988. Heterogeneity, Tournaments, and Hierarchies. J. Polit. Econ. 96 (4), 867–881.

Briggs, D., Betebenner, D., 2009. Is Growth in Student Achievement Scale Dependent? presented at the annual meetings of the National Council on Measurement in Education, April, 2009.

Briggs, D.C., Weeks, J.P., 2009. The Sensitivity of Value-Added Modeling to the Creation of a Vertical Score Scale. Educ. Finance Policy 4 (4), 384–414.

Campbell, D.T., 1976. Assessing the Impact of Planned Social Change. Occasional Working Paper 8. Dartmouth College, Public Affairs Center, Hanover, N.H.

Clotfelter, C.C., Ladd, H.F., Vigdor, J.L., Allan Diaz, R., 2004. Do School Accountability Systems Make it More Difficult for Low-Performing Schools to Attract and Retain High-Quality Teachers? J. Policy Anal. Manage 23 (2), 251–271.

Cronin, J., Dahlin, M., Adkins, D., Gage Kingsbury, G., 2007. The Proficiency Illusion. Thomas, B. Fordham Institute, Washington, DC.

Cunha, F., Heckman, J., 2008. Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation. J. Hum. Resour. 43, 739–782.

Cullen, J., Reback, R., 2006. Tinkering Towards Accolades: School Gaming under a Performance Accountability System. In: Advances in Applied Microeconomics. Vol. 14. Elsiever, pp. 1–34.

Dee, S.T., Keys, B.J., 2004. Does Merit Pay Reward Good Teachers? Evidence from a Randomized Experiment. J. Policy Anal. Manage. 23 (3), 471–488.

Dixit, A., 2002. Incentives and Organizations in the Public Sector: An Interpretative Review. J. Hum. Resour. 37 (4), 696–727.

Figlio, D.N., 2009. Voucher Outcomes. In: Mark Berends, Matthew G. Springer, Dale Ballou, Herbert J. Walberg, (Eds.), Handbook of Research on School Choice. Lawrence Erlbaum Associates/Taylor & Francis Group.

Figlio, D.N., Winicki, J., 2005. Food for Thought: The Effect of School Accountability Plans on School Nutrition. J. Public Econ. 89, 381–394.

Gibbons, R., 2010. Inside Organizations: Pricing, Policies, and Path Dependence. Annu. Rev. Econom. 2, 337–365.

Gillborn, D., Youdell, D., 2000. Rationing Education: Policy, Practice, Reform, and Equity. Open University Press, Buckingham.

Glewwe, P., 2009. Teacher Incentives in the Developing World. In: Springer, M.G. (Ed.), Performance Incentives: Their Growing Impact on American K-12 Education. Brookings.

Glewwe, P., Ilias, N., Kremer, M., 2010. Teacher Incentives. Am. Econ. J. Appl. Econ. 2 (3), 205–227.

Goodman, S., Lesley. T., Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program, mimeo, Columbia University.

Hanushek, E.A., Raymond, M.E., 2005. Does School Accountability Lead to Improved Student Performance? J. Policy Anal. Manage. 24 (2), 297–327.

Hanushek, E.A., Rivkin, S.G., 2006. Teacher Quality. In: Hanushek, E.A., Welch, F. (Eds.), Handbook of the Economics of Education. Vol. 2. Elsevier, B. V.

Holmstrom, B., 1982. Moral Hazard in Teams. Bell J. Econ. 13 (2), 324–340.

Holmstrom, B., Milgrom, P., 1991. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership and Job Design. J. Law Econ. Organiz. 7, 24–52.

Hoxby, C.M., 2002. Would School Choice Change the Teaching Profession. J. Hum. Resour. 37 (4), 846–891.

Hudson, S. 2010. The Effects of Performance-Based Teacher Pay on Student Achievement. SIEPR Discussion Paper No. 09-023.

Jacob, B., 2005. Accountability Incentives and Behavior: The Impact of High Stakes Testing in the Chicago Public Schools. J. Public Econ. 89 (5), 761–796.

Jacob, B., Levitt, S., 2003. Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. Q. J. Econ. 118 (3), 843–877.

Jovanovic, B., 1979. Job Matching and the Theory of Turnover. J. Polit. Econ. 87 (5), 972–990.

Klein, S., Hamilton, L., McCaffrey, D., Stecher, B., 2000. What Do Test Scores in Texas Tell Us? Rand Issue Paper 202.

Koretz, D.M., Barron, S.I., 1998. The Validity of Gains on the Kentucky Instructional Results Information System (KIRIS). RAND, Santa Monica.

Koretz, D.M., 2002. Limitations in the Use of Achievement Tests as Measures of Educators' Productivity. J. Hum. Resour. 37 (4), 752–777.

Ladd, H.F., 1999. The Dallas School Accountability and Incentive Program: an Evaluation of its Impact on Student Outcomes. Econ. Educ. Rev. 18 (1), 1–16.

Lavy, V., 2002. Evaluating the Effect of Teacher's Group Performance Incentives on Pupil Achievement. J. Polit. Econ. 110 (6), 1286–1317.

Lavy, V., 2009. Performance Pay and Teacher's Effort, Productivity, and Grading Ethics. Am Econ. Rev. 99 (5), 1979–2021.

Lazear, E., Rosen, S., 1981. Rank Order Tournaments as Optimum Labor Contracts. J. Polit. Econ. 89 (5), 841–864.

Martins, S.P., 2009. Individual Teacher Incentives, Student Achievement and Grade Inflation. IZA Discussion Paper No. 4051.

Muralidharan, K., Sundararaman, V., 2010. Teacher Performance Pay: Experimental Evidence from India. UCSD, mimeo.

Neal, D., 2009a. Designing Incentive Systems for Educators. In: Springer, M. (Ed.), Performance Incentives: Their Growing Impact on American K-12 Education. Brookings.

Neal, D., 2009b. The Role of Private Schools in Education Markets. In: Berends, M., Springer, M.G., Ballou, D., Walberg, H.J. (Eds.), Handbook of Research on School Choice. Lawrence Erlbaum Associates/Taylor & Francis Group.

Neal, D., 2010. Aiming for Efficiency Rather than Proficiency. J. Econ. Perspect. 24 (3), 119–131.

Neal, D., Schanzenbach, D.W., 2010. Left Behind by Design: Proficiency Counts and Test-Based Accountability. Rev. Econ. Stat. 92 (2), 263–283.

O'Keeffe, M., Viscusi, K.W., Zeckhauser, R.J., 2019. Economic Contest: Comparative Reward Schemes. J. Labor Econ. 2 (1), 27–56.

Prendergast, C., 1999. The Provision of Incentives in Firms. J. Econ. Lit. 37 (1), 7–63.

Reardon, S.F., Raudenbush, S.W., 2009. Assumptions of Value-Added Models for Estimating School Effects. Educ. Finance Policy 4 (4), 492–519.

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

Rockoff, J.E., 2004. The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. Am. Econ. Rev. 94 (2), 247–252.

Rockoff, J.E., Staiger, D.O., 2010. Searching for Effective Teachers with Imperfect Information. J. Econ. Perspect. 24 (3), 97–118.

Rothstein, J., 2010. Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. Q. J. Econ. 125 (1), 175–214.

Shepard, L.A., Dougherty, K.C., 1991. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. Effects of High Stakes Testing on Instruction.

Springer, M.G., Ballou, D., Hamilton, L., Le, V., Lockwood, J.R., McCafrey, D., Pepper, M., Stecher, B., 2010. Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching. National Center on Performance Incentives at Vanderbilt University, Nashville, TN.

Stecher, B.M., 2002. Consequences of Large-Scale, High-Stakes Testing on School and Classroom Practice. In: Hamilton, L.S., Stecher, M.B., Klein, S.P. (Eds.), Making Sense of Test-Based Accountability in Education. National Science Foundation.

Vigdor, J.L., 2009. Teacher Salary Bonuses in North Carolina. In: Springer, M. (Ed.), Performance Incentives: Their Growing Impact on American K-12 Education. Brookings.

Winters, M., Greene, J.P., Ritter, G., Marsh, R., 2008. The Effect of Performance-Pay in Little Rock, Arkansas on Student Achievement. National Center on Performance Incentives, Peabody College of Vanderbilt University, Working Paper 2008-02.

Wragg, E., Haynes, G., Wragg, C., Chamberlin, R., 2001. Performance Related Pay: The Views and Experiences of 1000 Primary and Secondary Headteachers. Teachers' Incentives Pay Project Occasional Paper 1. School of Education, University of Exeter.