

Credentials Versus Performance: Review of the Teacher Performance Pay Research

Michael Podgursky

*Department of Economics
University of Missouri–Columbia*

Matthew G. Springer

*Department of Leadership Policy, and Organizations
Peabody College of Vanderbilt University*

In this article we examine the economic case for merit or performance-based pay for K–12 teachers. We review several areas of germane research. The direct evaluation literature on these incentive plans is slender; highly diverse in terms of methodology, targeted populations, and programs evaluated; and primarily focused on short-run motivational effects. It is nonetheless fairly consistent in finding positive program effects. The general personnel literature highlights potentially significant selection effects of employee compensation systems. This is particularly relevant for teaching, because a growing body of production function research points to large, persistent, but idiosyncratic differences in teacher productivity. Thus, along with motivation effects, there is potential for substantial positive long run selection effects from teacher performance pay systems. The evaluation literature is not sufficiently robust to prescribe how systems should be designed (e.g., optimal size of bonuses, mix of individual vs. group incentives). However, it is sufficiently positive to support much more extensive field trials, pilot programs, and policy experiments, combined with careful follow-up evaluation.

We thank Samantha Dalton and Kelly Fork for research assistance. The usual disclaimers apply.

Salary schedules for teachers are a nearly universal feature of public school districts. Data from national surveys show that nearly 100% of public school teachers are employed in school districts that use salary schedules for pay setting. In large school districts the pay of thousands of teachers in hundreds of schools—from kindergarten up to secondary teachers in math and science—is typically set by a single district schedule (Podgursky, 2006).

These salary schedules for teachers contrast with the situation in most other professions. In medicine, the pay of doctors and nurses varies by specialty. Even within the same hospital or HMO, pay will differ by specialty field. In higher education there are large differences in pay between faculty by teaching field. Faculty pay structures in most higher education institutions are flexible. Starting pay is usually market driven, and institutions will often match counteroffers for faculty whom they wish to retain. Merit or performance-based pay is commonplace. Ballou and Podgursky (1997) and Ballou (2001) reported generally similar findings for private K–12 education. Even when private schools report that they use a salary schedule for teacher pay, payments “off schedule” seem commonplace.

Indeed, by comparison with pay determination in teaching, even government civil services schedules are more flexible and market based. Starting pay in the federal General Schedule system reflects market factors. New PhDs in economics or finance typically will start at higher levels of pay than will PhDs in most other disciplines. Professionals advance through the General Schedule steps not only within a grade but also between grades based on merit as well as experience.

The current era of academic standards and accountability ushered in by the No Child Left Behind has been an important stimulus for experiments in performance-based pay. Many districts, and even entire states, are experimenting with the concept of pay for performance in an attempt to bolster administrator and teacher performance. Pay for performance comes in many different forms, from compensation based on supervisor evaluations to payments awarded on the basis of portfolios created by teachers and/or self-evaluations. By some journalistic estimates (perhaps exaggerated), at least one third of the nation’s school districts appear “poised” to participate in local, state, or federal-initiated performance incentive policy.¹ Poised or not, it’s clear that many states and districts are actively considering this option.

¹Local and state officials are currently exploring pay for performance initiatives in states such as Arizona, Arkansas, California, Florida, Minnesota, Iowa, Idaho, New Mexico, North Carolina, North Dakota, Texas, and Wisconsin. Moreover, the federal government has appropriated \$99 million for a Teacher Incentive Program. Concomitantly, 60% of states have

This is not the first time that the American education system has entertained such reforms. In the wake of the *A Nation at Risk* report (National Commission on Excellence in Education, 1983), a number of school districts began experimenting with teacher incentives as a means to improve student outcomes and promote a more flexible compensation system. Teacher career ladders and merit pay programs were among the most visible programs employed (Dee & Keys, 2004). Research on these programs suggests it was difficult to create a reliable process for identifying effective teachers, measuring the value-added to a student by an individual teacher, eliminating unprofessional preferential treatment from evaluation processes, and standardizing assessment systems across schools. Moreover, past programs included insignificant dollar amounts awarded to successful teachers, faced opposition to alternative compensation systems by teacher unions, and lacked rigorous evaluations to assess and possibly recalibrate programmatic components more effectively to bring the program to scale. As a result, teacher compensation reform efforts that included incentives encountered opposition and were generally short-lived (e.g., Hatry, Greiner, & Ashford, 1994; Murnane & Cohen, 1986).

In this article we examine the economic case for merit or performance-based pay in K–12 education. Although our focus is on teachers, by far the largest group of employed professionals in K–12 education, many of the arguments generalize to school administrators as well. We begin by reviewing several strands from the growing economics literature on performance pay that have particular relevance for K–12 education.

One important issue, often ignored in discussions of teacher performance pay, is motivation versus selection effects in an incentive system. A long-run effect of any performance system will be to draw in employees who have higher earnings capacity in the incentivized regime and shed those with lower earnings capacity. A second important theme is the role of credentials versus performance in pay determination. A growing body of research points to large but idiosyncratic teacher effects. This potentially makes teacher incentives an attractive policy option. We review the research to date on teacher effects generally, their relationship to employment evaluations, and the overall effect of performance-pay systems on outcomes. Although the evaluation literature is not sufficiently robust to prescribe how systems should be designed (e.g., optimal size of bonuses, mix of individual vs. group incentives), it is sufficiently positive to suggest that further experiments and pilot programs by districts and states are in order.

enacted legislation requiring localities to explore alternative compensation systems (Wallace, 2003).

Current Experiments in Reform

As previously above, there is growing interest in performance-based compensation in K–12 education in states and districts, fueled in part by pressures from state accountability systems and the No Child Left Behind Act. We are aware of no systematic compilation of teacher incentive schemes. The Education Commission of the State is tracking some reforms and issues periodic reports on the topic (Azordegan, Byrnett, Campbell, Greenman, & Coulter, 2005). It does seem to be the case that interest in states and districts is growing. Several examples illustrate this trend. Florida statutes for several years have required school districts to use at least 5% of their teacher wage bill for merit or performance-based pay. It is clear that some, perhaps many, Florida districts have been slow to comply with the law. This has led the Florida legislature to implement a statewide teacher merit pay plan, E-Comp, which is supposed to provide pay bonuses to the top 10% of Florida teachers based on value-added estimates.

The Denver performance pay plan, ProComp, has received a great deal of national attention. In this plan, annual teacher pay increases are determined by a variety of performance measures, including student test scores. This new system is optional for incumbent teachers but mandatory for all teachers hired after 2005–06. Minnesota has introduced a Q Comp system that provides \$260 per student to fund incentive pay systems that meet requirements established by the state. These include multiple career paths and objective assessment systems. There has been a great deal of interest in performance pay in Texas. Houston public schools is rolling out a teacher merit plan in the 2005–06 school year that will be based in part on achievement gain scores. Statewide, a new program, the Governor's Educator Excellence Awards, provides funds for teacher performance pay to 100 high-poverty, high-achieving Texas public schools in the 2005–06 school year. This program is scheduled to expand to roughly 1,200 schools in the 2006–07 school year.

Many of these teacher-level incentive plans have been stimulated by a special Teacher Incentive Fund enacted by Congress, which provides \$100 million annually on a competitive basis to school districts, charter schools, and states to fund experiments and pilot performance-based pay projects.

Although states and districts are approaching performance pay in a variety of ways, one model program that is attracting attention in many states is the Teacher Advancement Program (TAP) developed by the Milken Family Foundation.² In this system, teachers advance up a promotion ladder

²The TAP Web site states "There are over 125 TAP schools nationwide, impacting more than 3,500 teachers and 56,000 students" (<http://www.tapschools.org/newsroom>).

based on numerous performance evaluations based on classroom observations as well as student achievement gains (i.e., from career, mentor, to master teacher). Most schools continue to use traditional salary schedules along with TAP career ladder. TAP recommends a pay differential of \$15,000 between master and career teachers (Azolbeydan et al., 2005).

All of these examples are programs designed to encourage individual-level teacher bonuses. However, there is a longer history of states and districts providing school-level bonuses—paid across the board to professional staff—to high-performing schools. Dallas and South Carolina had such programs in place in the mid-1990s. The size of the bonus awards, however, was relatively modest (Clotfelter & Ladd, 1996). A similar school-wide program was implemented in the Charlotte–Mecklenburg school district (Kelley, Heneman, & Milanowski, 2002).

Challenges of Incentive Design

A growing literature on incentive design, compensation, and performance measurement in economics, organizational theory, and psychology offers a useful lens for viewing the K–12 debate. For example, economics principal-agent theory initiated in seminal works by Berhold (1971), Ross (1973), and Jensen and Meckling (1976) concerns the relationship between a principal and an agent, whereby the principal contracts the agent to act toward an agreed-upon set of outcomes. Principal-agent theory presupposes an informational asymmetry exists between the principal and agent whereby the agent has an informational advantage over the principal. This information asymmetry is most salient given that education is characterized as a complex and multidimensional enterprise (Baker, 1992; Holmstrom & Milgrom, 1991). That is, it is possible that a contract (i.e., pay for performance program) designed by the principal does not encompass all relevant aspects of an organization mission, and as a result, “the use of explicit contracts could cause agents to focus too much on those aspects of the job included in the contract to the detriment of those that are excluded” (Prendergast, 1999, p. 21).

In the wake of the *A Nation at Risk* report in 1983, a number of school districts began limited experiments with merit pay. Most of these were short-lived. Case studies of districts that implemented merit pay were undertaken by Murnane and Cohen (1986) and Hatry et al. (1994). In

taf?page=release.20060530.Algiers). The \$100 million Teacher Incentive Fund implemented by the U.S. Department of Education is likely to accelerate the expansion of TAP schools in coming years.

fact, one of the more influential critiques of merit pay was Murnane and Cohen, who, drawing on strands of the personnel literature just described, argued that teaching is not a field that lends itself to performance-based compensation methods. Their objections included the following:

- *Difficulty in monitoring performance.* Teacher performance is difficult to monitor. Unlike, say, the sales of a salesman or the billable hours of a professional such as a doctor or lawyer, the output of teacher is not marketed, and thus we cannot readily measure the value of the services provided by an individual teacher.
- *Team production.* To a considerable extent teachers work as members of a team. Introducing individual merit pay would reduce incentives for teachers to cooperate and overall performance of the school will suffer.

Murnane and Cohen (1986) also believed that teacher merit pay fails another important test. They wrote

Merit pay is efficient when the nature of the activity in which workers are engaged is such that supervisors can provide relatively convincing answers to these two questions posed by workers:

1. Why does worker X get merit pay and I don't?
2. What can I do to get merit pay? (p. 7)

Of course, some of these criticisms are specific to individual merit pay. For example, a performance bonus given to an entire team of teachers would not undermine team morale. On the other hand, it is well understood that as the size of the team grows, so does the "free rider" problem (Prendergast, 1999). In the ensuing 2 decades, the measurement of teacher and school performance has become considerably more reliable. States are rapidly developing massive longitudinal student databases that permit more precise estimation of value-added at the building; grade; and, in a growing number of states, at the teacher level. As these measurement systems grow increasingly sophisticated, it is hard to imagine that they will not begin to play a role in personnel policy.

A more recent, and elegant, conceptual discussion of the performance pay problem in education is found in Lazear (2003). Lazear considered the economic arguments for and against two alternative regimes—payment for input and payment for output. In the absence of externalities or information problems, payment for output always trumps payment for input in terms of raising overall productivity. The most obvious reason is that incentives are appropriately lined up for incumbents. Of course, there has

been a great deal of attention paid to potential “multitasking” problems with using tests or other quantitative measures of teacher performance (e.g., Dixit, 2002; Hannaway, 1992; Holmstrom & Milgrom, 1991).

However, Lazear (2003) also pointed out a more subtle but important factor in the gains from a performance pay system that arise from labor market selection. A performance pay system will tend to attract and retain individuals who are particularly good at the activity being incentivized and repel those who are not. He noted that this effect on the workforce can be very important in explaining productivity gains. For example, in one of his own case studies outside of teaching, he found that the sorting effect was substantial and roughly equal in magnitude to motivation effect. In other words, although the incentive system raised the productivity of the typical worker employed, it also tends to raise the overall quality of the workforce. After introduction of the plan, workers who quit were replaced by much more productive workers.

Lazear speculated that this selection effect may be a powerful factor in teacher labor markets. For example, studies of teacher turnover consistently find that high-ability teachers are more likely to leave teaching than teachers of lower ability (Murnane & Olsen, 1990; Podgursky, Monroe, & Watson, 2004). A recent provocative study by Hoxby and Leigh (2004) found evidence that the migration of high-ability women out of teaching from 1960 to the present was primarily the result of the “push” of teacher pay compression—which took away relatively higher earnings opportunities for teachers—as opposed to the pull of greater nonteaching opportunities. As they pointed out, the remunerative opportunities for teachers of high and low ability grew outside of teaching. However pay compression within education accelerated the exit of the higher ability teachers. To the extent that these high-ability teachers were more effective in the classroom, a performance-based pay regime would have kept relatively more of them in the classroom.

The selection effect also undermines the Murnane and Cohen case against individual merit pay. Merit pay might work even if individual teachers do not know what to do to improve their teaching performance (and supervisors have no advice to give). It can do so simply by attracting and retaining more productive teachers. We next consider further evidence regarding potential sorting gains.

Empirical Research

Economic theory can enhance our understanding of labor market dynamics and behavior of workers on the job; however, ultimately we

must take these theories to the data. What can we expect from pay-for-performance programs? How should performance be measured? What are the appropriate level and structure of the incentives? What is the proper mix of individual and group incentives? Not surprisingly, the research base is thin; nonetheless it does provide some interesting insights on this policy debate. Here we review three strands of research we believe are relevant to this debate: the teacher effects literature, studies linking teacher effects to performance assessments, and direct evaluations of individual and group performance pay schemes.

Teacher effects studies. Over the last decade, researchers have begun to exploit massive state longitudinal student achievement data files to undertake “value-added” studies of teacher effectiveness. Beginning with William Sanders’s work in Tennessee (Ballou, Sanders, & Wright, 2004; Wright, Horn, & Sanders, 1997), such studies have now expanded to Texas (Rivkin, Hanushek, & Kain, 2005), and large school districts such as New York City (Kane, Rockoff, & Staiger, 2005; Boyd, Grossman, Lankford, & Loeb, 2006) and Chicago (Aaronson, Barrow, & Sanders, 2003). A consistent finding in the value-added studies is that there are large and somewhat persistent differences in achievement gain scores between classrooms and teachers. This has led to the oft-stated belief that teachers can have a large effect on student achievement growth.

Although researchers have found substantial variation in teacher effects within school districts and even within schools, they have also consistently found that these effects are highly idiosyncratic, that is, the estimated teacher effects are largely unrelated to measured teacher characteristics such as the type of teaching certificate held by the teacher, teacher education (e.g., MA), licensing exam scores, and experience (beyond the first 2 years). Indeed, nearly every researcher conducting rigorous teacher effect studies has taken note of this fact. (e.g., Aaronson et al., 2003; Goldhaber & Brewer, 1997; Kane et al., 2005; Rivkin et al., 2004). For example, in a large-scale study of certification status and effectiveness of new teachers in New York City public schools, Kane et al. (2005) wrote

In other words, there is not much difference between certified, un-certified, and alternately certified teachers overall, but effectiveness varies substantially among each group of teachers. To put it simply, teacher vary considerably in the extent to which they promote student learning, but whether a teacher is certified or not is largely irrelevant to predicting their effectiveness. (p. 40)

We have reproduced from their study a chart of estimated teacher effects that demonstrates clearly this point (see Figure 1). Here they report

variation in estimated teacher effects for new teachers by type of teaching certificate held by the teacher.³ Clearly the distributions overlap almost entirely, illustrating the point of this quote. However, for our purposes it is worth noting the very wide variation in teaching effectiveness within each of the certification groups. Any policy that can retain and sustain the performance of the upper-tail teachers and enhance or discard teachers in the lower tail has the potential for substantial effects on student achievement.

A recent study of Chicago public teachers by Aaronson et al. (2003) illustrates this point as well. Like other such studies, this work is based on a very large longitudinal file of linked student achievement scores. What makes this study unique is that the authors also have very extensive administrative data on teacher characteristics that are unavailable in other studies, including education, experience, types of teaching licenses, and selectivity of the teacher's undergraduate college. Aaronson and colleagues found that more than 90% of teacher effects are not explained by any measured teacher characteristics.

The implications of highly dispersed, idiosyncratic teacher effects for the design of teacher performance pay are important. First, the research to date provides little support for a "credential-based" system of teacher compensation. For example, there is little evidence that teacher graduate degrees, the most common educational credential, have any effect on student achievement (Hanushek, 2003). The evidence is similar for teacher certification. Some have proposed bonuses for teachers certified by the National Board for Professional Teaching Standards as an alternative to merit pay.⁴ Indeed, states are investing substantial sums in bonuses for Board-certified teachers. Yet even here the evidence concerning performance is mixed (Goldhaber & Anthony, in press; Sanders, 2007).

An important consequence of this wide dispersion of teacher effects is the potential importance of selection on teacher performance. A policy that ties remuneration to performance over time will tend to pull many more teachers in the upper tail of this into the teaching workforce, whereas low

³Another team examining New York City public schools reached a similar finding (Boyd et al., 2006).

⁴For example, the National Commission on Teaching and America's Future, in their influential 1996 report, stated, "It has always been difficult to recognize and reward good teachers in ways that are credible and objective. The merit pay plans of the 1980's (like those of the 1950's and 1920's) have already disappeared because local evaluators did not have useful standards, or the time or expertise, to make reliable judgments about teacher competence In contrast, the careful process of National Board certification—based on evaluation by experts according to well-developed standards and a collaborative process—provides an alternative that teachers find credible, helpful, and an extraordinary learning experience" (p. 74).

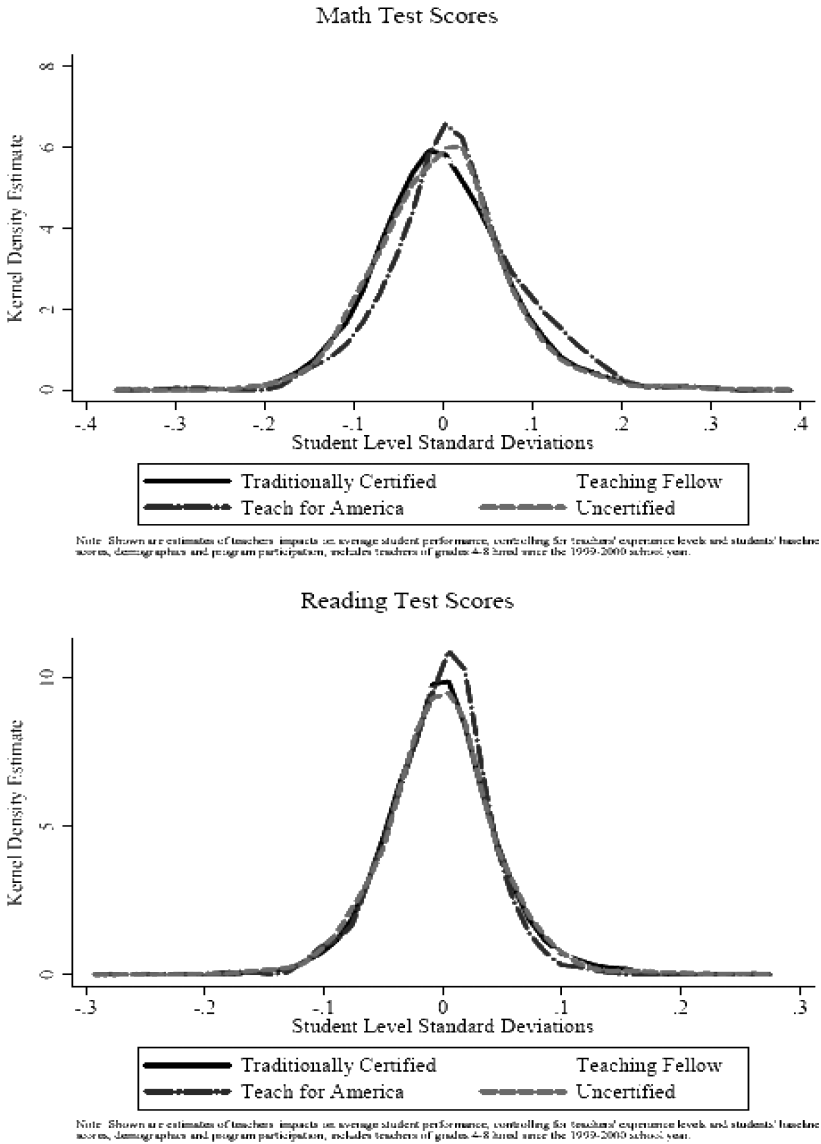


Figure 1 Variation in teacher effectiveness by type of teacher certificate: New York City public schools, 1998–99 to 2004–05. Source: Kane, Rockoff, & Staiger (2006, Figure 6).

productivity teachers will tend leave for nonteaching positions. To illustrate this point, suppose that a pay system is monotonically related to the teacher effectiveness measured on the horizontal axis in Figure 1. Further assume that all teachers have identical nonteaching earnings illustrated by a vertical bar somewhere horizontal axis, indicating the teaching productivity equivalent of the alternative wage. Ignoring nonpecuniary preferences for teaching versus other jobs, teachers with productivity to the right of the vertical bar would move to (or stay in) teaching and those to the left would leave. Teacher turnover would thus become part of a virtuous cycle of quality improvement, rather than a problem to be minimized.

Teacher effects, merit awards, and principal evaluations. As previously noted, the problem of multitasking a job makes subjective assessments by supervisors and peers highly important in many contexts (Prendergast, 1999). What does education research tell us the relationship between these subjective assessments and teacher performance as measured by test score gains? A number of value-added studies find that principal evaluations are a reliable guide to identifying high- and low-performing teachers as measured by student achievement gains. Two older studies using student longitudinal data by Armor et al. (1976) and Murnane (1975) found large effects of principal evaluations on student achievement gains. More recently, Sanders and Horn (1994) reported, "There is a very strong correlation between teacher effects as determined by the data and subjective evaluations by supervisors" (p. 2000). A particularly rigorous recent study is focused entirely on the predictive validity of supervisor evaluations. Jacobs and Lefgren (2005) examined the relationship between scores on a detailed principal evaluation of teacher performance and teacher effects in a medium-sized unidentified school district in the western United States. They estimated teacher effectiveness measures for 202 teachers in Grades 2 through 6 in math and reading. They demonstrated a significant positive relationship between value-added measures of teacher productivity and principals' evaluations of teacher performance.⁵ The principals' evaluations had a significant positive relationship teacher gains scores for math and reading. They noted that principals were much better at identifying teachers in the upper and lower tails of the effectiveness distribution. Another particularly interesting part of this study was an "out of sample" prediction of 2003 student achievement scores based on teacher value-added estimates from 1998 to 2002 and principal ratings. Not surprisingly, students had higher average scores in math and science if they had teachers

⁵ Among other things, principals were asked to assess the ability of teachers to raise student achievement on a scale from 1 (*inadequate*) to 10 (*exceptional*).

with higher measured effectiveness in prior years. The same is the case for teachers with higher principal ratings. Jacobs and Lefgren found that the principal evaluation remains a statistically significant predictor of current student achievement even when teacher value-added (in the previous year) is added to the model. This finding suggests that principal evaluations provide an important independent source of information on teacher productivity.

In sum, the research literature to date and particularly the recent work by Jacobs and Lefgren show that subjective evaluations of teacher performance are valid measures of teacher effectiveness as measured by student achievement gains. These assessments are, of course, readily available and are readily incorporated into formal teacher evaluations.

All of these results pertain to “low-stakes” principal evaluations. They suggest that principals have a relatively good notion of who the good and bad teachers are. A second question is whether this knowledge will find its way into a merit pay system. The fact that a principal identifies a teacher as “inadequate” on an anonymous survey does not mean that he will do so in a high-stakes context. Two other studies shed some light on whether “old-style” merit schemes based in part on supervisor evaluations (along with other factors) are positively associated with teacher value-added measures. Cooper and Cohn (1997) found that classroom gain score measures were higher for teachers who received merit pay awards in South Carolina.⁶

A more recent study by Dee and Keys (2004) examined the relationship between career ladder bonuses and student achievement gains in Tennessee Project STAR data. This study is unique because teachers’ students were randomly assigned to teachers in the experiment. Although the focus of the STAR experiment (and subsequent research studies) has been on the effect of class size, Dee and Keys took advantage of the fact that students were also randomly assigned to Tennessee career ladder teachers. Teachers advanced on the career ladder rungs primarily on the basis of statewide evaluations conducted by local administrators—typically principals. Dee and Keys found that teachers with career ladder status (who have thus passed one or more evaluations) were more effective than teachers who had not. Unfortunately, it is not possible to sort out the selection effect from the validity of the evaluation per se. It may have been the case that teachers who self-select to join the program or advance up the ranks are better teachers.

⁶In the individual plan, teachers who applied for the award were evaluated on four criteria; one was a performance evaluation, and another was evidence of superior student achievement gains. Thus, this is not a strong test of the thesis.

Although no single study is definitive in this area, at this point a small literature set has developed showing that principal evaluations as well as promotions or awards in merits systems based in whole or in part on principal evaluations are associated with higher classroom effectiveness on the part of teachers. This finding does not entirely address the problem multitasking issue. The multitasking critique says that there are many valuable attributes to teacher performance that are not adequately measured by state assessments (Hannaway, 1992). To the extent that principal's subjective assessments capture these, it is useful to know that these principal's evaluations are also correlated with teacher productivity as measured by student gain scores.

Assessments of performance pay systems. Although there have been numerous experiments in individual and group incentive pay for teachers over the years, the evaluation literature is very slender. Table 1 lists studies that we have found in the literature that employ a conventional treatment and control evaluation design. It is interesting to note that, in contrast to the very mixed findings of studies of teacher characteristics such as certification or teacher education (Hanushek, 2003), the slender incentive literature generally finds positive achievement effects.

It should be noted that this literature, even more than the inputs literature reviewed by Hanushek, has highly diverse "treatments." In some cases it is clear that the incentive programs were poorly designed. However, in every case examined, even those labeled "mixed," teachers clearly responded to the incentives. The problem is that many of these schemes were not well designed, and incentives were not targeted to the appropriate educational outcomes.

Table 1 summarizes some characteristics of these incentive schemes and findings from the studies. We have not attempted a more sophisticated "meta-analysis" or analytical synthesis for several reasons. First, as previously noted, these are very different incentive schemes, and "effect size" would be problematic. Second, the outcome variables analyzed also vary considerably, sufficiently so that we do not feel it is useful to attempt convert them to standard deviations or other common measures. The last column represents our assessment of the outcome of the study. Although it is our subjective assessment, we do not believe that the authors, or most other careful readers, would dispute these interpretations. Because the number of studies is small, and the range of assessments considered is highly diverse, we discuss most of them briefly.

Ladd (1999) and Clotfelter and Ladd (1996) examined the effect of a schoolwide incentive scheme implemented in the Dallas school district in the mid-1990s. This scheme provide a modest pay boost to all teachers in

Table 1
Quantitative Studies of the Causal Effect of Teacher Incentive Programs on Measures of Student Achievement

<i>Study</i>	<i>Sample</i>	<i>Study Time Span</i>	<i>Type of Teacher Incentive</i>	<i>Size of Incentive (per Teacher)</i>	<i>Outcome Variable</i>	<i>Results</i>
Ladd (1999); Clotfelter & Ladd (1996)	Dallas Grade 7 schools relative to other Texas urban districts ^a	1991–1995	Schoolwide (tournament)	\$1,000	Math and reading test scores, dropout rates	Positive
Eberts et al. (2002)	2 MI alternative high schools (1 treatment, 1 control)	1994–95 to 1998–99	Individual	Up to 20% of base pay	Course completion rates, pass rates, daily attendance, GPA	Mixed
Lavy (2002)	Israel, high schools	1993–95 to 1996–97	Schoolwide (tournament)	\$200–\$715	Test scores, pass rates, dropout rates, course-taking	Positive
Lavy (2004)	Israel, high schools	1999–2001	Individual (tournament)	\$1,750–\$7,500+ ^a	Pass rates and test scores	Positive
Glewwe et al. (2004)	Primary schools, rural Kenya		Schoolwide	Up to 43% of monthly salary	Grade 4, 8 test scores	Mixed
Atkinson et al. (2004)	UK high schools	1997–2002	Individual	> 9% in salary base	English, science, math assessments	Positive
Figlio & Kenney (2006)	U.S. NELS88 matched, to FK survey or 1993–94 SASS, 12th grade; public and private	1993	Individual	Varied within sample	12th grade, composite reading, math science and history score	Positive

Note. MI = grade point average; SASS =

^aThese are winnings per class. However, a teacher could enter multiple classes.

high-performing schools. Because the program was intended to raise the performance of all schools in the district, the district is the appropriate treatment unit. Clearly this makes evaluation challenging. They nonetheless find that achievement in Dallas rose relative to other Texas school districts. The remaining studies focused on school- or teacher-level effects.

Two of the most methodologically rigorous evaluations are Lavy (2002, 2004). In both of these studies the incentive design is explicitly a tournament designed to raise pass rates on high school exit exams in low-socioeconomic status high schools in Israel. Although they were not randomized or true experiments, both programs were implemented in a way that permitted rigorous nonexperimental evaluation. In addition, the incentive schemes were carefully designed so as to minimize gaming or other opportunistic behavior. The first considers a tournament in which a selected group of low-performing schools competed on the basis of school-wide performance and bonuses were distributed equally to all teachers in the winning schools. This was found to have a positive effect on the participating schools as compared to nonparticipating control schools. The second article (Lavy, 2004) examined an individual teacher bonus program, also run as a tournament. Essentially teacher participants were ranked on the basis of value-added on a variety of exit exams and bonuses were given to top performers. These were substantial bonuses, as large as \$7,500 per class on an average base pay of \$25,000. Participant teachers' performance rose relative to a control group of nonparticipants.

In the latter paper Lavy also investigated whether the program had the type of negative spillover consequences often discussed in the contracting literature. First, as proposed the multitasking issue, test scores in other nontournament subjects did not fall. In addition, and consistent with the teacher value-added literature previously discussed, teacher characteristics such as experience or certification could not predict the winners. One nice feature of this study is that Lavy compared the cost effectiveness of the individual bonus scheme as compared to group bonuses or policies providing additional educational resources (aside from pay) for low-achieving schools. He found that the cost per unit gain in the incentive program dominated those found in the group incentive or added resource programs.

It is interesting that even in those cases where we report "mixed" findings, the incentive scheme always had a positive effect on the behavior that was incentivized ("You get what you pay for"). Eberts, Hollenbeck, and Stone (2002) presented a case study of the effect of an incentive scheme in a single alternative high school in Michigan. Because dropout rates in courses were a problem, the school introduced a bonus system that paid teachers to raise course completion rates by their students. The researchers

compared the “treatment” school to another alternative high school that was considered comparable. In fact, the bonus program significantly raised course completion, but, not surprisingly, nontargeted variables such as student pass rates or grade point average dropped, as academically marginal students were induced to stay in school. Clearly a better performance pay plan would have incorporated a larger set of performance indicators. However, the results of the study show that teachers responded positively to a short-term incentive plan, even if poorly designed.

A second “mixed” study involved primary schools in rural Kenya. This study is unique in the group in that it used random assignment methods. Fifty schools were chosen at random for participation from among 100 relatively low-performing rural schools. The bonuses were tied to pass student pass rates on state exams in a variety of subject areas. The bonuses were substantial, ranging from 21 to 43% of monthly pay. However the program was of short duration (originally announced as 1 year’s duration, later extended to 2). The researchers found that pass rates on the state exams (the target of the incentive) increased, but the gains did not persist to subsequent years, which they took as evidence of gaming on the part of teachers. Although the targeted teachers provided more after school “test prep” sessions, the researchers found no evidence of differences in pedagogy or in teacher absenteeism (a major problem in these schools).

Both the Michigan and Kenya cases thus showed that teachers responded to the incentives in the program but that the incentive regime was poorly designed. A review of the principal–agent multitasking literature outside of education by Courty and Marschke (2003) highlights the dynamic learning context of these incentive systems. Their model highlights the importance of experimentation and trial and error required on the part of the principal in getting the performance measure right. These two teacher pay studies provide good illustrations of the need for such experimentation.

Atkinson et al. (2004) evaluated the effect of a teacher bonus pay scheme in the United Kingdom. The program was short-lived—only 1 year in duration. In addition, *ex post* it turned out that the bonus was provided to nearly all teachers who applied. The authors make a fairly convincing case that, *ex ante*, this not the perception. To win the performance teachers were required to provide evidence to the education ministry that the achievement gains of their students exceeded national averages in five areas. They compare gain score data for eligible teachers with benchmark data on gain scores prior to the implementation of the program. Unfortunately the authors had some difficulty in developing a representative sample with pre- and postprogram gains score data. With this caveat, however,

they nonetheless found a large and statistically significant effect of this 1-year program.

Figlio and Kenny (in press) analyzed a sample that is more broadly representative of U.S. schools. They creatively merged data from the National Educational Longitudinal Survey of 1988 with their own survey on merit pay, as well as data from the 1993–94 Schools and Staffing Surveys. They made use of natural variation in the sample in the use of incentive-based pay by both public and private schools. A unique feature of this study is that, whereas the other studies analyzed only the presence or absence of a given program, the natural variation in their sample in incentive programs allowed them to construct a measure of the strength of the teacher incentive “dosage” at the school that includes not only the existence of a merit pay scheme but also its pecuniary consequences. The effects of even modest doses of incentive pay are statistically significant in both public and private schools. The size of the effects, in comparison to effects of other inputs, are also noteworthy. They estimated the effect of a high level of implementation of incentives (relative to none at all) has an impact on achievement comparable to a 1 standard deviation increase in days absent by the student and an increase in maternal education of 3 years. It is much greater than the effect of any school input such as class size or teacher salaries.

Clearly using natural variation has its cost, as the variation may not arise exogeneously. However, one benefit of a study using natural variation is that many of the schools using the incentive plans may have had them in place for a sufficient length of time to pick up both motivation and selection effects.

In sum, this literature is extremely diverse in terms of incentive design, population, type of incentive (group vs. individual), strength of study design, and duration of the incentive. Yet in every case, the evidence suggests that teachers responded to the incentives. That and the fact that other spillover behaviors may not have improved suggest that the incentive schemes were not well targeted. Indeed, the lesson noted by Courty and Marschke (2003) is that, given gaming and multitasking, trial and error are probably required to get the right set of performance incentives.

Incentive pay in private and charter schools. If contracting problems in K–12 education such as performance monitoring, multitasking, and team production are inherent in the production process and sufficiently severe so as to preclude group or individual merit pay, then we would expect to see similar pay structures in charter schools and private schools as compared to traditional public schools (see Table 2). Several studies have

Table 2

Teacher Salary Schedules and Teacher Incentive Pay in Traditional Public, Charter, and Private Schools

	<i>Traditional Public (%)</i>	<i>Charter (%)</i>	<i>Private (%)</i>	<i>Nonreligious Regular School (%)</i>
Is there a salary schedule for teachers in this school?	96.3 (0.29)	62.2 (0.72)	65.9 (1.24)	45.1 (5.60)
Does this school currently use pay incentives such as cash bonuses, salary increases, or different steps on the salary schedule to reward:				
NBPTS certification?	8.3 (0.37)	11.0 (0.43)	9.6 (0.88)	14.8 (5.5)
Excellence in Teaching?	5.5 (0.35)	35.7 (0.65)	21.5 (0.93)	42.9 (5.5)
Completion of in-service professional development?	26.4 (0.70)	20.5 (0.56)	18.7 (0.88)	26.0 (5.67)
Recruit or retain teachers in fields of shortage?	10.4 (0.464)	14.9 (0.54)	7.9 (0.61)	15.0 (3.40)

Note. Standard errors are in parentheses. Source: 1999–00 Schools and Staffing Surveys (reported in Podgursky, 2006).

examined this question and found significant differences. Charter and private schools generally operate in a more competitive, less regulated, and nonunion environment. Hoxby (2002) hypothesized that increased competition leads to greater use of merit and performance-based pay and finds evidence in support of that thesis. Ballou (2001) directly explored the question of whether private schools make greater use of merit pay by analyzing the structure of earnings conditional on experience and education. His findings point to greater use of merit pay in private school wage setting. A more recent study by Podgursky (2006) examines data from the 1999–2000 Schools and Staffing Surveys. He found significantly higher use of performance pay bonuses by private and charter schools. Finally, Ballou and Podgursky (1993) examined survey data on teacher attitudes from the 1987–88 Schools and Staffing Surveys. They found that private school teachers are much more supportive of merit pay than public school teachers, which is consistent with the sorting hypothesized by Lazear (2003).

Issues for Further Research

All of the bonus studies surveyed in the previous section were basically short-run assessments. However, we should once again distinguish between motivation and selection effects. Because of the short time frame

of implementation for these programs, they are primarily estimating motivation effects. In the long run one might expect positive selection effects as well. Given the large dispersion of teacher effectiveness in Figure 1, the potential for strong positive sorting effects is substantial. Unfortunately, there has been little empirical work on this topic.⁷

At the same time, in the long term there may be more opportunities for gaming the system. Incentive schemes that tie teacher pay to achievement gains by students, either for the individual teacher or the “team,” provide incentives for cheating or other opportunistic behavior. Studies of high-stakes accountability systems have documented teachers focusing excessively on a single test and educators altering test scores and/or assisting students with test questions (Goodnough, 1999; Jacob & Levitt, 2005; Koretz et al., 1999). Related analyses have found evidence of schools’ strategic classification of students as special education and limited English proficiency (Cullen & Reback, 2002; Deere & Strayer, 2001; Figlio & Getzler, 2002; Jacob, 2002), use of discipline procedures to ensure that low-performing students will be absent on test day (Figlio, 2005), manipulation of grade retention policies (Haney, 2000; Jacob, 2002), misreporting of administrative data (Peabody & Markley, 2003), acceptance of test exemptions/waivers demanded by parents (Neufeld, 2000), and/or planning of nutrition enriched lunch menus prior to test day (Figlio & Winicki, 2005). As such, multiple mechanisms to minimize negative spillover effects need to be created for a high-stakes pay for performance program to be successful.

To the extent that these new performance pay regimes, particularly those focusing on individual teacher performance, and rely on estimates of teacher value-added, there are concerns about the statistical reliability and robustness of these estimates. First, we do not know how successfully statistically determined estimations of teacher performance effects can guide educational practice and provide incentives for teachers to change practice. Some researchers warn that we have to be careful in interpreting teacher effects as purely an attribute of the teacher without consideration of the school context (Ballou et al., 2004; Koedel & Betts, 2005; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004).

⁷Hanushek, Kain, O’Brien, and Rivkin (2005) examined the mobility of high- and low-productivity teachers as measured by gain scores in a set of urban Texas school districts. They found no evidence that the teachers lured to suburban districts by higher salaries are more effective teachers. Of course, these suburban districts are almost certainly paying these teachers off of salary schedules and not direct measures of their productivity.

Conclusion

In this article we examine the economic case for merit or performance-based pay in K–12 education. Our focus is on teachers, by far the largest group of employed professionals. However, many of the arguments generalize to school administrators as well. We began by reviewing several ideas from the growing economics literature on performance pay that have particular relevance for K–12 education. We considered some well-known problems in the use of incentive pay in any context and their relevance to K–12 education. One important theme, which is not widely considered in the education studies, is motivation versus selection effect in an incentive system. A second important theme is the role of credentials versus performance in pay determination. A growing body of research points to large but idiosyncratic teacher effects. This potentially makes teacher incentives an attractive policy option. We reviewed the research to date on teacher effects generally, their relationship to employment evaluations, and the overall effect of performance pay systems on outcomes. Although the literature is not sufficiently robust to prescribe how systems should be designed (e.g., optimal size of bonuses, mix of individual vs. group incentives), it is sufficiently positive to suggest that further experiments and pilot programs by districts and states are in order. As these are introduced, however, it is important to bring them out in a way that makes effective evaluation. The development of massive student longitudinal achievement database opens prospects for rigorous value-added assessment of these state experiments.

We would also suggest that, in addition to states and school districts, education philanthropies can make a unique contribution as well. In our survey of the research, we noted that the strongest finding to date comes from two experimental merit pay systems implemented in high schools in Israel (Lavy, 2002, 2004). Both of these were rank-order tournaments and involved substantial rewards for teachers. States and districts have been reluctant to implement tournaments, and unions are strongly averse to such schemes. By their very nature these are zero sum games, and many assert such a system would discourage teacher collaboration and cooperation, to the detriment of overall school performance. Foundations, on the other hand, routinely give out prizes, often with substantial sums. However, these awards are not implemented in a way that would permit evaluation of their incentive effects. This may be an area where the philanthropic community can do good and simultaneously add to our research knowledge base.

References

- Aaronson, D., Barrow, L., & Sanders, W. (2003). *Teachers and student achievement in Chicago public high schools*. Chicago: Federal Reserve Bank of Chicago.
- Armor, D., Conry-Oseguera, P., Fox, M., King, N., McDonnell, L., Pascal, A., et al. (1976). *Analysis of the school preferred reading program in selected Los Angeles minority schools*. Santa Monica, CA: RAND Corporation.
- Atkinson, A., Burgess, S., Croxon, B., Gregg, P., Propper, C., Slater, H., et al. (2004, December). *Evaluating the impact of performance-related pay for teachers in England*. UK: University of Bristol, Centre for Market and Public Organization.
- Azordegan, J., Byrnett, P., Campbell, K., Greenman, J. & Coulter, T. (2005, December). *Diversifying teacher compensation*. Denver, CO: Education Commission of the States.
- Baker, G. (1992). Incentive contracts and performance measurement. *Journal of Political Economy*, 100, 598–614.
- Ballou, D. (2001). Pay for performance in public and private schools. *Economics of Education Review*, 20(1), 51–61.
- Ballou, D., & Podgursky, M. (1993). Teachers' attitudes toward merit pay: Examining conventional wisdom. *Industrial and Labor Relations Review*, 47(1), 50–61.
- Ballou, D., & Podgursky, M. (1997). *Teacher pay and teacher quality*. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.
- Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 29(1), 37–66.
- Berhold, M. (1971). A theory of linear profit-sharing incentives. *Quarterly Journal of Economics*, 85, 460–482.
- Boyd, D., Grossman, P., Lankford, H., & Loeb, S. (2006). How changes in entry requirements alter the teacher workforce and affect student achievement. *Education Finance and Policy*, 1, 176–216.
- Clotfelter, C., & Ladd, H. (1996). *Recognizing and rewarding success in public schools*. In H. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education*. Washington, DC: The Brookings Institution.
- Cooper, S. T., & Cohn, E. (1997). Estimation of a frontier production function for the South Carolina educational process. *Economics of Education Review*, 16, 313–327.
- Courty, P., & Marschke, G. (2003, Summer). Dynamics of performance-measurement systems. *Oxford Review of Economic Policy*, 19, 268–284.
- Dee, T., & Keys, B. J. (2004). Does merit pay reward good teachers? Evidence from a randomized experiment. *Journal of Policy Analysis and Management*, 23, 471–488.
- Deere, D., & Strayer, W. (2001). *Putting schools to the test: School accountability, incentives, and behavior*. Unpublished manuscript, Texas A&M University.
- Dixit, A. (2002). Incentives and organizations in the public sector. *Journal of Human Resources*, 37, 696–727.
- Eberts, R., Hollenbeck, K., & Stone, J. (2002). Teacher performance incentives and student outcomes. *Journal of Human Resources*, 37, 913–927.
- Figlio, D., & Getzler, L. (2002). *Accountability, ability and disability: Gaming the system?* (NBER Working Paper No. 9307). Cambridge, MA: National Bureau of Economic Research.
- Figlio, D. & Kenny, L. (in press). Individual teacher incentives and student performance. *Journal of Public Economics*.
- Figlio, D., & Winicki, J. (2005). Food for thought? The effects of school accountability plans on school nutrition. *Journal of Public Economics*, 89, 381–394.

- Goldhaber, D., & Anthony, E. (in press). Can teacher quality be effectively assessed? National Board Certification as a signal of effective teaching. *Review of Economics and Statistics*.
- Goldhaber, D., & Brewer, D. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on education productivity. *The Journal of Human Resources*, 32, 505–523.
- Goodnough, A. (1999, December 8). Answers allegedly supplied in effort to raise test scores. *New York Times*.
- Haney, W. (2000). The myth of the Texas miracle in education. *Education Analysis Policy Archives*, 8(41).
- Hannaway, J. (1992, Spring). Higher order skills, job design, and incentives: An analysis and proposal. *American Educational Research Journal*, 29, 3–21.
- Hanushek, E. A. (2003, February). The failure of input-based resource policies. *Economic Journal*, 113(485), F64–F68.
- Hatry, H., Greiner, J., & Ashford, B. (1994). *Issues and case studies in teacher incentive plans*. Washington, DC: Urban Institute Press.
- Holmstrom, B., & Milgrom, P. (1991). Multitask principal agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organizations*, 7, 24–52.
- Hoxby, C. M. (2002). Would school choice change the teaching profession? *Journal of Human Resources*, 37, 846–891.
- Hoxby, C. M., & Leigh, A. (2004). Pulled away or pushed out? Explaining the decline of teacher aptitude in the United States. *American Economic Review*, 93, 236–240.
- Jacob, B., & Lefgren, L. (2005). *Principals as agents: Subjective performance measurement in education* (NBER Working Paper No. 11463). Cambridge, MA: National Bureau of Economic Research. Available from <http://www.nber.org/papers/w11463>
- Jacob, B., & Levitt, S. (2005). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, 118(3).
- Jensen, M., & Meckling, W. (1976). Theory of the firm: Managerial behavior, agency costs, and ownership structure. *Journal of Financial Economics*, 3, 305–360.
- Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2005). *Identifying effective teachers in New York City*. Paper presented at NBER Summer Institute.
- Kelley, C., Heneman, H., & Milanowski, A. (2002). Teacher motivation and school-based performance rewards. *Education Administration Quarterly*, 38, 372–401.
- Koedel, C., & Betts, J. (2005). *Re-examining the role of teacher quality in the education production function* (Working paper). University of California–San Diego.
- Koretz, D., et al. (1999). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS)*. Santa Monica, CA: RAND Corporation.
- Ladd, H. (1999). The Dallas school accountability and incentive program: An evaluation of its impacts on student outcomes. *Economics of Education Review*, 18, 1–16.
- Lavy, V. (2002). Evaluating the effect of teachers' group performance incentives on pupil achievement. *Journal of Political Economy*, 110, 1286–1317.
- Lavy, V. (2004, June). *Performance pay and teachers' effort, productivity and grading ethics* (Working Paper No. 10622). Cambridge, MA: National Bureau of Economic Research.
- Lazear, E. (2003). Teacher incentives. *Swedish Economic Policy Review*, 10, 197–213.
- McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value added models for teacher accountability, MG-158-EDU*. Santa Monica, CA: RAND.
- McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Education and Behavioral Statistics*, 29, 67–101.
- Murnane, R. J. (1975). *The impact of school resources on the learning of inner city children*. Cambridge, MA: Ballinger Press.

- Murnane, R., & Cohen, D. (1986). Merit pay and the evaluation problem: Why most merit pay plans fail and few survive. *Harvard Education Review*, 56, 1–17.
- Murnane, R. J., & Olsen, R. J. (1990). The effects of salaries and opportunity costs on length of stay in teaching: Evidence from North Carolina. *Journal of Human Resources*, 25, 106–124.
- National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Department of Education.
- National Commission on Teaching and America's Workforce. (1996). *What matters most: Teaching for America's future*. Washington, DC: Author.
- Peabody, Z., & Markley, M. (2003, June 14). State May lower HISD rating; Almost 3,000 dropouts miscounted, report says. *Houston Chronicle*, p. A1.
- Podgursky, M. (2006). Teams versus bureaucracies: Personnel policy, wage-setting, and teacher quality in traditional public, charter, and private schools. *Education and Policy Analysis Archives*. Available from <http://www.uark.edu/ua/der/EWPA/approved/Teams.v.B.html>
- Podgursky, M., Monroe, R., & Watson, D. (2004). The academic quality of public school teachers: An analysis of entry and exit behavior. *Economics of Education Review*, 23, 507–518.
- Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature*, 37, 7–63.
- Rivkin, S., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73, 417–458.
- Ross, S. (1973). The economic theory of agency: The principal's problem. *American Economic Review*, 63, 134–139.
- Sanders, W. J., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System: Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, 8, 299–311.
- Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement. *Journal of Personnel Evaluation in Education*, 11, 57–67.

Copyright of PJE. Peabody Journal of Education is the property of Lawrence Erlbaum Associates and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.