Designing Incentive Systems for Schools

Derek Neal

University of Chicago and NBER [*]

June, 2008

forthcoming in

Performance Incentives: Their Growing Impact on American K-12 Education

edited by Matthew Springer, Brookings

[*] I thank the participants at the NCPI conference for useful comments. I thank Matthew Springer for detailed suggestions. I thank Margarita Kelrikh and Richard Olson for research assistance. I thank Gadi Barlevy, Edward Haertel, and Canice Prendergast for useful discussions. I thank the Seale Freedom Trust for research support. I also thank Lindy and Michael Keiser for their support through a gift to the University of Chicago's Committee on Education.

This chapter considers the design of incentive systems for educators. Much debate concerning the design of performance incentives in education centers on specific psychometric challenges. Advocates of the use of performance incentives in education often argue that student test scores provide objective measures of school output, but their opponents raise concerns about the breadth of assessments, the reliability of assessments, the alignment of assessments with curriculum, and the potential for schools to manipulate assessment results through various forms of coaching or even outright cheating. In sum, many doubt that school systems can or will construct student assessments that truly form a basis for measuring and rewarding educational performance.[1] These psychometric concerns are first order, and section one discusses these issues in detail. However, most of this chapter argues that policy makers and researchers must pay more attention to several challenges that would remain even in a world with perfect assessment technologies.

Assume for a moment that the only mission of schools is to foster the math skills associated with a particular curriculum. Further, assume that policy makers in this setting possess an ideal instrument for assessing math skill and are able to make assessments of every student at the beginning and end of each school year. Even these ideal assessments do not contain the information policy makers need to rank schools according to their performance.

If a factory produces 500 widgets today, we know that the value of this output is 500 times the price of a widget. If Johnny's math scale score rises

[1] See the Chapter by Rothstein (2008) in this volume for more on these issues.

from 140 to 145, we may conclude that Johnny's expected number of correct answers, in a setting that requires him to try all the items in a specific domain, has increased by 5. However, we do not know what this increase in expected correct answers is worth to Johnny or to society. In addition, we do not know whether a five-point increase would have been worth more to society if Johnny had begun the school year with a baseline score of 130 or 150 instead of 140. Finally, because Johnny may receive tutoring and support from his parents as well as his teachers, it is not straightforward to determine what portion of Johnny's score increase should be credited to his school rather than his family.

Education is not the only field where it is difficult to attach dollar values to the marginal contribution of a given worker or a group of workers that function as a production unit. Private firms that face these measurement issues often abandon the task of trying to produce cardinal measures of output for individual workers or teams. Rather, firms take on the more manageable task of forming performance rankings among workers or groups of workers and then deliver bonuses, raises, and promotions as a function of these rankings.[2]

However, section two explains that the task of constructing performance rankings in public education differs from the task of constructing performance rankings in most private firms because there is no clear way, a priori, to

[2] Lazear and Rosen (1981) began the economics literature that describes these forms of incentive pay as prizes associated within rank order performance tournaments.

collapse the multiple dimensions of school performance into a single index. Private sector firms may not be able to precisely measure the contribution of a worker or group of workers to overall profits, but firms know that this is the criterion by which they seek to rank performance. In public education, policy makers must begin the process of designing incentive systems by developing a clear definition of performance and then pay close attention to the mapping between this definition and the performance ranking procedures they adopt.

Section three discusses the benefits of building incentive pay systems around competition among schools rather than competition among individual teachers. Because the potential benefits of cooperation among teachers are large relative to the costs of cooperation, incentive systems should foster cooperation and not undermine it. Section four discusses the benefits of allowing individual schools or organizations that manage groups of schools to compete not only in academic performance contests but also in the labor market for teachers. Systems that assign reward pay at the school level but allow each school to allocate resources among teachers according to their own personnel policies foster competition in the market for teachers that may speed the rate of social learning about the best ways to hire, mentor, and motivate teachers.

Most of the analyses offered here rest on the implicit assumption that there exists a benevolent education authority that faithfully represents the interests of taxpayers, but the concluding section considers whether or not the public provision of education invites political corruption that contaminates the design of incentive systems. This observation raises the possibility that voucher systems serve as complements to rather than substitutes for incentive pay and accountability systems.

1 The Limits of Performance Statistics

Private firms have the ability to hand out bonuses, promotions, and other forms of reward pay based not only on objective information and procedures but also on the subjective evaluations of owners or the managers that work for them. This arrangement is possible because workers know that owners are losing their own money when firms fail to retain, motivate, and promote their best employees. However, there are no residual claimants in government agencies, and officials that run public organizations may suffer no harm if they hand out bonuses and reward pay to their friends and family instead of those who are most deserving. This feature of public agencies generates demands by public employees that performance incentive systems in government tie rewards and punishments to objective performance measures, and these performance statistics are often reported to the public.

In 1976, Donald Campbell put forth a pair of observations concerning government statistics that are often referenced as *Campbell's Law*:

> 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.[3]

Campbell makes two assertions. First, when governments attach important stakes to specific government statistics, actors within governments often face incentives to engage in activities that improve these statistics without actually improving the conditions that the statistics are meant to monitor. These activities corrupt the statistics in question because the statistical improvements induced by the activities do not coincide with real improvements in welfare. Second, the same activities that corrupt performance statistics may actually cause direct harm.

Campbell provides numerous examples of this phenomenon, and in a chapter in this volume, Rothstein (2008) provides more detail concerning Campbell's observations and related observations from several different fields. Because Rothstein's summary of existing evidence suggests that *Campbell's Law* may be an appropriate label, education policy makers should be wary of performance incentive or accountability systems that rely heavily on performance statistics. Workers change their behavior in response to the adoption of any particular performance measurement system, and these responses often compromise the value of the performance measures in question. This section describes exactly how these responses compromise performance measures, and the conclusion of this chapter discusses how the political process may corrupt decisions concerning what types of information are aggregated

[3] See Campbell (1976) p. 49.

into performance measures in the first place.

Modern economists typically use Holmstrom and Milgrom's (1991) multitasking model to organize their analyses of the phenomena that Campbell describes. Holmstrom and Milgrom built their model to explain why private firms often choose not to attach incentives to performance statistics even when they have access to statistics that are highly correlated with actual performance. Their insights concerning the settings in which private firms are reluctant to attach high stakes to performance measures help us understand why Campbell drew such pessimistic conclusions about the use of performance statistics in government.

In the multi-tasking model, agents may take many different actions at work, and by assumption, employers have two tools for directing the efforts of their workers. First, firms may pay the costs required to monitor their workers' actions directly. Second, firms may link worker pay to performance statistics. These statistics are noisy signals of worker outputs, and the key assumption is that the relationships between worker actions and measured output are not the same as the relationships between worker actions and actual output. Some actions have a greater impact on measured output than actual output while the reverse is true for other actions.

Advocates of recent trends in education reform hope that high-stakes assessments will prompt teachers to allocate more effort toward activities like teaching math and less toward activities that amount to leisure for teachers and extra recess for students, and any fair assessment of test-based accountability programs would likely conclude that accountability systems do create these types of changes in effort allocation.[4] However, the logic of the multi-tasking model suggests that other re-allocations should also be expected.

Everyone knows the saying that "you get what you pay for," but the multi-tasking model takes this line of reasoning a step further. Because effort is costly, if firms pay for performance as measured by a statistic, workers will not only allocate more effort to the actions that have the greatest impact on the statistic in question but also allocate less effort to actions that do not have large direct impacts on this statistic. Further, this reallocation will occur even if it involves allocating less effort to actions that have large positive impacts on actual output. In education, these reallocations may involve teachers spending less time on activities that foster creativity, problem solving skills, the ability to work well in teams, or other important skills that are not directly assessed. Thus, even if teachers put forth more total effort following the introduction of assessment based incentive pay, these types of reallocations may result in students being worse off.[5]

Campbell's empirical observations combined with the insights of the multi-tasking model are warning signs for those who wish to design incentive pay systems for public schools. However, the balance of this chapter argues that,

[4] See Hanushek and Raymond (2004) and Hanushek (2005).

[5] This scenario may be avoided if teachers find that the process of preparing children for specific assessments actually lowers the cost of building other skills. If the process of preparing children for a high-stakes assessment makes it easier to teach critical thinking skills, social skills, etc, it may be possible to design an accountability system that generates improved performance on a specific assessment without taking attention and effort away from the skills that are not directly assessed.

when one considers the design challenges inherent in constructing incentive pay systems for educators, the corruption of assessments is only the tip of a large iceberg. The following sections discuss important challenges that remain even in a world with ideal assessment technologies.

2 Required Ingredients

2.1 The Best of All Possible Worlds

Incentive pay systems for educators require two components. First, these systems require a method for ranking schools or teachers according to performance. Second, these systems require the assignment of specific rewards and penalties to the various performance ranks that schools or teachers may receive. This section focuses only on the task of constructing performance rankings, and it begins by analyzing the construction of performance rankings in a world with ideal assessment technologies.

In this ideal setting, the following are true:

- There are exactly K skills that schools are supposed to foster.
- Each school has N students.
- There exists an assessment for each of the K skills.
- Each of the K assessments has perfect reliability.
- Neither students nor teachers can corrupt the assessment results.

- Variation in achievement growth among students is determined entirely by school performance.

The no corruption assumption implies that the only way that schools can enhance their students' test scores is to engage in activities that create more skill among their students. The final assumption implies that policy makers can isolate the contribution of schools to student skill development.

This ideal setting brings to the forefront a fundamental issue that must be settled before the process of designing incentive systems for schools can begin. In order to design a system that rewards performance, one must first define performance. Note that, at a point in time, all the assessment information for the students in a given school can be placed in in NxK matrix. Each row contains all the assessment results for a particular student. Each column contains the results of a particular assessment for all the students in that school. If we index schools by $s = 1, 2, ...S$, we can define a set of S matrices $X = (X^1, X^2, ..X^S)$. Each matrix is NxK , and together these matrices contain all skill assessments for all students in all schools at a point in time. For simplicity, assume that these measures are taken on the first day of school. Next, define X' as the collection of measurements $X' = (X^{1'}, X^{2'}, ..X^{S'})$ taken among the same students on the last day of the same school year. Given that society began the school year at X, how does society evaluate the relative values of ending the year at any one of the billions of possible X' outcomes? Further, if we take the matrices of test scores from the beginning and end of the school year for any two schools, how do we use these four matrices to decide which school performed better?

In a truly perfect world, an incredibly skilled team of econometricians possessing the largest research grant in the history of social science research would have already devised a method for estimating the social value (in dollars) of moving the students in school s from any X^s to any $X^{s'}$, and given this method, it would be easy to design incentives for educators. Education authorities could simply allow competing school districts or school management companies to bid for the opportunity to operate schools in given locations and then pay each of these entities a lump sum at the end of the year equal to the social value of the change in human capital among all of its students minus the total amount bid for the right to operate schools.

This simple approach is not possible in education, and this idealized setting shows that the central reason is orthogonal to common observations concerning the difficulty of accurately assessing all the skills produced in schools. Even if policy makers possessed measures of all skills produced in schools, and these measures were reliable and expressed on interval scales, policy makers would still have no idea how to value various improvements on these scales in monetary terms.

Even psychometrically perfect assessments provide no rational basis for constructing pay for performance systems that look like piece rate or commission systems, and further, they do not provide the information required to simply rank schools or teachers according to performance. Because school output is multi-dimensional, i.e. there are NxK outcomes at each point in time in each school, it is not clear a priori how one collapses this information into a one-dimensional performance ranking for schools or teachers.

Many owners and managers in the private sector also operate in environments that do not permit them to assign a dollar value to the marginal contributions of each of their employees, but the task of constructing performance rankings is likely more complicated in education than in these private firms. If the partners in an accounting firm sit down to form a ranking of their associates, each partner knows the criterion they are supposed to use. They are supposed to rank associates based on their best guesses concerning how much each associate could add to the total value of the partnership. However, if the superintendent of a large school district or even a state decides to rank schools or teachers according to their performance, she must first construct a definition of performance.

2.2 Defining Performance First

Any sensible method of constructing performance rankings in education must be guided by three principles that are all variations on the same theme. One must develop a coherent definition of performance that serves as an anchor for the procedures used to construct performance rankings.

2.2.1 Spelling Out Priorities

First, the documents describing any accountability or incentive pay system should spell out the priorities of policy makers. These documents should clearly delineate the types of achievement that the system is intended to foster, and to the extent possible, these documents should explore how policy makers view the relative importance of achievement in various subjects or by various types of students. Thus, policy makers should begin by formulating clear answers to questions like the following:

- Is progress in reading more valuable than progress in math or civics, and if so, how much?
- Is it more socially valuable to bring a disadvantaged student closer to grade level than to bring a gifted student closer to her full potential, and if so, how much?
- What are the relative values of non-cognitive traits like persistence versus cognitive skills?

Schools are supposed to simultaneously foster many skills in hundreds of students at the same time. Without clear answers to these questions and many others, the task of objectively ranking the overall performance of any two schools is a hopeless endeavor.[6]

[6] Dixit (2002) correctly notes that many different advocacy groups act as performance monitors in public education, and these groups do not always have the same priorities. Seen in this light, the typical failure of existing incentive pay systems to take clear and coherent stands on how performance should be defined and measured is not completely surprising. However, my goal is not to explain why current government policies are what they are but rather to outline normative criteria that incentive policies should meet.

2.2.2 A Clear Mapping Between Priorities and Procedures

Second, the mapping between the policy priorities that define an incentive system for educators and the procedures used to create performance rankings for schools and teachers should be clear and precise. This step is quite challenging, but those who design and implement incentive systems risk failure if they do not devote enough attention to this essential task.

Consider the No Child Left Behind Act of 2001 (NCLB) as an example. The language of the act, beginning with its title, impresses upon the reader that addressing the educational needs of the most academically disadvantaged is a high priority. However, Neal and Schanzenbach (2008) argue that, in states that measure school performance by counting the number of students who score above a state-wide proficiency standard, the levels of the proficiency standards on various assessments determine which students are most pivotal for a school's performance rating. Students who are below the proficiency standard but are able to achieve the standard given modest interventions are the students whose achievement gains matter most in determining their school's status under NCLB. Thus, even though the rhetoric surrounding NCLB highlights the need to improve outcomes among our most disadvantaged students, NCLB implicitly places greater social value on the learning of students in the center of the achievement distribution than the progress of students who are currently far below their state's proficiency standard.[7]

[7] Neal and Schanzenbach (2008) draw their conclusions based on data from Chicago, and

In states that use value-added systems to measure school or teacher performance, choices of scales for the exams combined with choices concerning how to weight improvements that occur in different ranges of the test score distribution determine the rewards that schools receive for directing attention to different students and different subjects, but policy makers often fail to offer a rigorous justification for these choices.

A concrete example helps make this point clear. In the 2006-2007 school year, Florida implemented the Special Teachers Are Rewarded (STAR) incentive system. An important component of the STAR program involved assigning performance points to teachers based on their students' gains on standardized tests using the Value Table method. Table 1 is an example of a Value Table. The Florida Department of Education (FODE) offered this table as a model for how points should be assigned to teachers under STAR based on their students' reading outcomes.

There are six levels of reading achievement for students in Florida elementary schools, and the table specifies points associated with each of the 36 possible student transitions. The table indicates that if a student goes from Level 2 to Level 3 in one year, her teacher receives 205 points. However, if another student moves from Level 1b to Level 2, his teacher receives only 145 points. The FDOE intentionally gave more points for improvements that

Reback (2007) draws similar conclusions based on earlier data from Texas. Springer (2007) does not find similar patterns using data from Idaho, but he cannot replicate the Neal and Schanzenbach (2008) methodology because he does not have access to assessments take prior to the introduction of NCLB.

are less common, but it is hard to see why these particular gradients are the right ones.

Calculate the difference between columns 3 and 5 for each level of Year 1 performance. The additional reward for bringing a student past Level 3 and up to Level 5 in Year 2 varies greatly depending on the baseline achievement level. The marginal reward is much greater if the student began at Level 1b than either Levels 1a or 2. Why would this be the case? Shouldn't one expect that the value to society of bringing a child from Level 3 to Level 5 is roughly the same regardless of the child's identity? If Johnny began the year behind Sue, but both Johnny and Sue are at the same reading level by January, is there any reason that society should value Johnny's learning during the spring more or less than Sue's?

Because the STAR proposal did not contain a detailed discussion of the relative social importance of different types of progress among different types of students, it would be easy to generate an equally plausible set of point allocations for the entries in Table 1 that would imply notably different results in terms of which teachers are ranked among the top performers in their district. STAR and other systems that do not create clear ties between how performance is defined and how performance is measured inevitably yield performance rankings that lack credibility.

2.2.3 Define Sensible Comparison Sets

Third, incentive systems should group schools according to the types of students and families they serve, and then rank schools either relative to other schools that serve similar students or to a performance standard designed for such schools. Any attempt to create a single performance ranking over all schools in an entire state or large district necessarily encounters serious conceptual problems. When school A is ranked above school B, the implication is that school A performed better than school B. However, if the two schools are educating students from extremely different backgrounds, one must ask, "better at what?"

In 2006, Hillsborough County, Florida decided to participate in the STAR merit pay system described above. Although STAR's Value Table approach sought to place all teachers on a level playing field, the 2006-07 results in Hillsborough suggest that the STAR procedures generated performance rankings that overstated the true performance of teachers in affluent schools relative to the performance of teachers in disadvantaged schools. County officials moved quickly to modify the plan, and the revised plan involves schools being placed in leagues according to their Title I status.[8]

The Hillsborough experience is not surprising when one realizes that the original plan sought to make performance comparisons among teachers who, in important respects, were not performing the same job.[9] The tasks of defin-

[8] See Stein (2008) for details. The new Hillsborough plan is part of the Merit Awards Program (MAP) that replaced STAR statewide.

[9] Another chapter in this volume, McCaffrey et al (2008) explores in more detail how

ing and measuring job performance in education are necessarily complicated because educators perform complex jobs, but these tasks become quixotic when policy makers insist on making performance comparisons among persons who are not working with comparable students.

The gains that students make, in a given year, on any particular assessment scale reflect the interaction of their initial skill level and the quality of the instruction they receive. Thus, data from two classrooms where students began the year at extremely different levels of achievement do not provide any information that allows one to directly compare the quality of instruction in the two classrooms. One can never rule out the possibility that students in one class room simply began in a region of the scale where it is easier to make significant gains.

This section has delineated several guidelines for constructing performance rankings in education. To begin, the process of constructing valid performance rankings requires that policy makers formulate a clear definition of performance. Further, ranking procedures should faithfully reflect the priorities contained in this definition of performance, and these procedures should be applied to sets of schools or teachers who work in comparable environments.

rankings of teacher performance vary depending on numerous choices that policy makers must make when building an empirical model to produce the rankings.

2.3 Auxiliary Benefits of Competition Within Leagues

Some will worry that a system requiring schools to compete only against other schools that draw from similar student populations may do little to improve performance in disadvantaged communities because it may be possible to outperform most schools in disadvantaged communities without actually performing at an exceptionally high level. However, this line of reasoning does not incorporate how teachers and principals might change where they choose to teach in response to such a system.

Imagine that there are ten different leagues in a state and that these leagues are defined by the pre-school preparation and family backgrounds of entering students. If an “easy” league exists where it is less costly to win reward pay or avoid sanctions, talented principals and teachers face a strong incentive to move to a school in this league. More importantly, teachers and principals who are best suited to teaching in the schools that belong to that particular league face the strongest incentive to move.

Further, in a system with league-specific tournaments, one can use differences in reward pay across leagues as an effective means of attracting the right teachers and principals to serve in disadvantaged communities. Those who respond to the extra reward pay are not only those who are willing to teach in disadvantaged communities but also those who are willing to bet that they know how to do it successfully.

Finally, by using schools with observationally similar students to define the performance standard for any given school, one minimizes an important performance measurement problem that has been assumed away in the analyses presented so far. If we observe that Johnny's math score rose by 10 points this year, it is hard to know what part of this gain should be attributed to the efforts of Johnny's teacher versus the inputs that Johnny receives outside school from parents, grandparents, or other adults.

To the extent that teachers and principals have information about the backgrounds of their children that are not reflected in the measures of preschool preparation or family background available to policy makers, it will not be possible to form perfect comparison groups for any school. However, to the greatest extent possible, whenever school A receives a better ranking than school B, this outcome should imply that school A performed better than school B and not simply that school A worked with more advantaged students.

The state of California actually produces a performance ranking for each school that is a ranking within a set of schools that are similar in terms of resources and the background of their students. Although the Similar Schools Rank (SSR) for a particular school gives a performance rank for that school within a set of schools that are deemed an appropriate comparison set, policy makers in California treat SSR data as simply "additional contextual information."[10] Neither the state accountability system nor the state implementation of NCLB attaches important rewards or sanctions to SSR outcomes.

[10] See PSAA Technical Report 00-1, p. 4.

The federal government, in its implementation of NCLB, and numerous states continue to make the mistake of asserting that rewards and punishments for educators must be determined by measures of how schools perform relative to either state-wide standards or every other school in the state. Defenders of this approach argue that it is the only way to implement high achievement standards for all children, but this argument confuses two distinct uses of statistics.

Statistics that provide accurate information concerning whether or not an organization is reaching stated goals are not necessarily the same statistics that organizations should employ in their incentive pay systems. If one wants to determine whether or not the children in a given state are reaching a minimum level of achievement that the state has set as an important target for all its citizens, then one obviously wants to measure the performance of each student against a common standard that reflects this target. However, if one wants to use assessment results as part of a set of personnel practices that rewards and punishes teachers and principals for their job performance, then one must make comparisons among persons who are working in comparable environments and thus doing comparable jobs.

It is important to note that neither value-added models nor growth models offer a way around this concern. The original Hillsborough approach sought to rank teachers using measures of achievement growth, and it still produced results that were not credible. If the baseline achievement distributions for two classrooms have little overlap, the data permit few direct comparisons between students who began the year at similar achievement levels but attended different classrooms. Although researchers or policy makers can always write down models that produce estimates of how the teachers in these two classrooms are performing relative to one another, the modeling choices of analysts, not data on actual relative performance, drive these estimates. Some will argue that Hillsborough simply chose the wrong growth model, but their real mistake was trying to make performance comparisons among teachers who were not working in comparable classrooms.

3 School versus Teacher Performance

Thus far, this chapter has not drawn distinctions between incentive systems that operate at the school level versus the teacher level and has often discussed incentive pay and accountability systems as if they operate at the school level. Nonetheless, the process of designing incentive systems in education involves making choices concerning the extent to which policy makers attach incentives to measures of overall school performance versus individual teacher performance.

Three different scenarios form interesting baselines. First, consider state or district-wide systems that link measured performance for individual teachers to their pay and job security. Second, imagine district or state-wide policies that tie incentive pay for teachers to measures of how their school or department performs. Finally, imagine a system that links all government performance incentives to school level outcomes but allows those who run schools to adopt their own policies concerning how incentive payments at the school level are allocated among different teachers within schools. This section and section four highlight several reasons that the latter two approaches are likely preferable to the first.

3.1 Cooperation and Information Sharing

It seems reasonable to assume that the teachers in a school possess a great deal of information concerning how the performance of their peers could be improved. However, incentive systems that rely solely on rewards and punishments for individual teachers do not provide any motivation for teachers to share this valuable information with their peers. Thus, even if an assessment-based system can accurately tell you that teacher A is not performing as well as her peers, the system will not foster efficient improvement in teacher A's performance if teacher A is the only person affected by her performance.

For at least two reasons, an efficient system will provide incentives for teacher A's principal and peers to help her improve. First, they likely have the best information concerning how she might improve. Second, the costs of sharing this information are often low relative to the benefits. When one teacher shares lessons learned from experience and experimentation with another teacher, the time costs required to convey information may often be quite low relative to the benefits, and it takes little imagination to come up with numerous examples. Information concerning pedagogy, organization, or even the personalities and needs of particular students in the school may often be shared at low cost but to great benefit.[11]

Incentive systems based on measures of individual teacher performance not only provide no incentive for teachers to engage in this type of information sharing but may also provide clear incentives to withhold such information. Any system that makes performance comparisons among teachers working in the same school actually creates incentives for teachers to sabotage the performance of their peers. Although some may view this possibility as far-fetched, economists point to this possibility as one reason that incentive systems used in the private sector are often designed to avoid direct competition among workers who are supposed to cooperate with each other in joint production activities.[12]

Some may argue that one can avoid these undesirable effects by having individual teachers compete against a fixed performance standard rather than each other. However, competition against fixed performance standards creates other problems. To begin, competition against a performance standard is competition against some historical notion of what was possible in the

[11] Itoh (1991) shows that when the cooperation or helping costs among workers are low enough relative to benefits, it is optimal for firms to adopt incentives policies that operate *only* at the team level. In another chapter in this volume, Muralidharan and Sundararman (2008) find no difference in achievement gains associated with teacher incentives versus school incentives using experimental data from India. However, the schools involved in their experiment contained only a handful of teachers, and the organization of these schools differs greatly from that of modern schools in developed countries. Gains from cooperation may be greatest in larger schools where a number of teachers are teaching similar material to different students. Lavy (2002) documents noteworthy responses to a school level incentive plan in Israel.

[12] See Lazear (1989).

past in a particular type of classroom. This form of competition cannot require educators to perform at efficient levels unless standards are constantly revised to reflect what is possible in different schooling environments given new methods of pedagogy, instructional technologies, and other resources.[13] Further, this need for revision and updating creates opportunities for political forces to build low performance expectations into the system. Competitions that allow the possibility that everyone can be a winner invite mischief that lowers standards.

In contrast, when incentive systems involve direct competition among schools for reward pay, individual teachers have clear incentives to help their peers improve because they receive no reward pay unless their school performs better than other schools. Further, if principals have the freedom to hand out different shares of their school's total reward pay based on their own aggregation of test score outcomes and their subjective evaluations of each teacher, principals can build reputations for rewarding not only individual performance but also cooperation among teachers. Principals have strong incentives to pursue this course of action if their pay and job security depend on their schools' overall performance rankings, and principals who follow this course strengthen incentives for teachers to help each other improve.

None of the above arguments against attaching incentive pay to measures of individual teacher performance deny that variation in individual teacher

[13] The tournament model of Green and Stokey (1983) clarifies the potential drawbacks of the performance standard approach.

performance is an important factor in determining variation in student outcomes. Everyone who has ever been a student knows that some teachers are much better than others, and recent work by Rivkin, Hanushek, and Kain (2005) provides clear evidence that this is the case. Identifying, training, and retaining talented teachers is key to running an effective school, and these tasks are too difficult to accomplish within systems that do not encourage all agents in a given school to use their information in ways that improve not only their individual performance but also the performance of others.

3.2 An Easier Measurement Problem

Incentive pay systems based on school performance are also easier to implement than systems built around measures of individual teacher performance because it is so difficult to measure differences in performance among teachers. The existing empirical literature provides clear evidence that teachers differ in efficiency but less clear evidence that statisticians can build reliable measures of teacher performance that form a credible basis for incentive pay. Several issues complicate the task of creating performance measures for individual teachers.

3.2.1 Noise

Estimates of individual teacher effects for a given year are quite noisy when one attempts to include reasonable controls for student and classroom characteristics. Although a number of researchers have argued that a particular type of Value-Added model can produce more reliable estimates of individual teacher effects by using multiple years of data, I do not see how a method that delivers precise estimates of teacher performance over periods of three to five years is useful as a basis for making personnel decisions and handing out reward pay.[14]

Most professionals in the private sector work in environments that involve some form of reward pay on at least an annual basis that comes in the form of bonuses, raises, or profit-sharing. Although decisions about promotions are made at less frequent intervals, one must remember that promotion systems not only provide incentives for current effort but also affect the efficiency of the entire organization by allocating the most talented and productive people to positions in which success depends most heavily on talent and productivity. Performance measures for individual teachers derived from many years of data may be useful inputs for a tenure evaluation process, but they are not useful as a means of providing incentives for existing teachers, especially tenured ones, to provide efficient effort levels on a continuous basis.

3.2.2 Ignoring Classroom Assignments

Rothstein (2007) highlights a second challenge for those who wish to use statistical methods to rank teachers based on their contribution to student

[14] See McCaffrey et al (2003) for a comprehensive review of Value-Added methods. See McCaffrey et al (2008a), in this volume, for a detailed case study that explores how variation in methods used to measure teacher effects as well as policies that link reward pay to different performance ranks can, in practice, generate noteworthy variation in distributions of reward pay among teachers. See McCaffrey et al (2008b) for a detailed treatment of the stability of estimated teacher productivity effects.

achievement. Rothstein shows using North Carolina data that the identity of a student's teacher in a future grade helps predict performance in the current school year. This pattern is consistent with the hypothesis that the allocation of students to teachers within a school is driven, at least in part, by how individual students are progressing through school. Rothstein presents evidence that this sorting of students to teachers is not solely driven by fixed student characteristics but also by how the student develops academically over time, and he argues that estimated teacher effects based on methods that seek to control for this type of student tracking over time are not highly correlated with estimates from more standard models.

Standard methods that researchers use to measure the relative performance of individual teachers rely on the assumption that the assignment of children to teachers is a random process given standard student background variables and can thus be ignored. However, the assignment of teachers to students within schools reflects a set of choices made by principals based on information that researchers cannot see. Some teachers excel at working with children who are naturally hard workers while other teachers have a comparative advantage in working with kids who are struggling in school or at home. Thus, when researchers assume that the assignments of teachers to students is ignorable, they are, in effect, assuming that principals systematically fail to do their jobs.

Ignorable assignment is still a challenge at the school level. The schools that parents chose for their children likely reveal information about unmeasured family characteristics that influence academic outcomes for their children. However, there are scenarios that make ignorable assignment at the school level much more palatable.

Recall that California already has a set of procedures that are designed to identify a comparison set of similar schools for any given school in California. In large states, it may be possible to form comparison sets that are not only homogeneous with respect to student characteristics but also geographically separated. Imagine a set of 50 elementary schools that serve as the comparison set for elementary school A. Assume that all 50 schools are similar to A with respect to the pre-school preparation and demographic characteristics of students and also assume that no student in any of these 50 schools lives within a one hour commute of school A. The fact that students in school A did not attend one of the 50 schools in this comparison set provides no information about school A or the comparison schools. The comparison schools were not realistic options for the students in school A. Further, the fact that students in the comparison schools did not attend A is not informative about A or the comparison schools because school A was not an option for these students.

If one uses such a comparison set to create a performance measure for school A, unmeasured factors at the community level may still create problems. However, there is no set of decisions facing parents, teachers, or principals that one expects to directly generate correlations between these unobserved factors and the assignment of students to either school A or the schools in its comparison set.

4 The Value of Labor Market Competition

Assume that a state or large district allows independent companies and non-profit organizations to bid for opportunities to manage public schools. Further, imagine an incentive system that provides reward funds at the school level based on an index of school performance and also provides for the termination of an organization's management contract if the same index falls below a specified level. This index might be based entirely on assessment results or a combination of assessment results and the results of school inspections, parent surveys, and other information. Regardless, the key assumption is that the index is a reliable indicator of how a particular school performs relative to similar schools that serve students from the same backgrounds.

In addition, assume organizations that manage schools are responsible for distributing reward money to teachers and for designing their own policies and procedures for evaluating teachers, screening new hires, terminating existing teachers, granting tenure, and determining career ladders for teachers within their schools. Thus, school management organizations compete with each other not only in determining the educational practices used within their schools but also in developing and implementing the personnel policies and procedures that identify and retain the best teachers. Because the resources of these organizations are tied to their performance, they face clear incentives to select personnel policies that retain and reward the teachers who make the greatest contributions to overall school quality. Further, as different organizations experiment with different management models, successful innovations will spread to other organizations and other schools.

This type of labor market competition among schools is almost never seen in the developed world. Although many European countries have education systems with voucher components that foster competition among schools for students, collective bargaining on a national or regional level sets most personnel policies for both private and public schools in these systems.[15]

The personnel economics literature describes many ways that private firms implement desirable performance incentive systems even in environments like education where it is impossible to precisely measure the marginal contributions of individual workers to the profits of firms. However, these papers usually describe incentive schemes that are only possible when firms know a great deal about both the preferences of their workers and the details of their production technologies.[16] Economists justify this approach to characterizing what firms actually do by noting that competition among firms for talented workers moves the actual personnel policies of firms toward the efficient ones.[17] Inefficient policies waste resources by either paying too much

[15] Denmark, Netherlands and Sweden are examples. See Neal (2008) forthcoming for details.

[16] In Lazaer and Rosen's (1981) seminal paper on tournaments, firms know the exact willingness of workers to supply different levels of effort and the precise relationship between effort and true output, even though neither the worker's contribution to output nor the worker's effort are observed. Similar assumptions are common in many models of bonus pay and promotions. See Prendergast (1999) for other examples.

[17] Here, efficient does not necessarily mean the first-best outcome in a world with perfect

for the effort that workers provide or by encouraging workers to provide effort that does not generate returns in excess of the incentive payments made to workers. Because firms that do not discover efficient ways to provide incentives for their workers waste resources and cannot compete long term with firms that do, competition in the product market enhances efficiency in the labor market.

For this reason, systems that promote competition among schools while allowing schools to compete for teachers by experimenting with different personnel policies offer greater promise than systems that impose a single set of incentive pay policies on all schools. Imagine that a state or district superintendent must design a single incentive pay system for an entire state or disitrict. Even if she possessed an ideal system for creating teacher performance rankings based on peer evaluations, principal evaluations, student assessment results and other relevant information, she would need a second crystal ball to help her determine the rewards and penalties that should be attached to particular performance ranks. In competitive labor markets, efficient innovators thrive and prosper while those who pursue inefficient personnel policies either abandon them or go out of business, but few competitive forces discipline the personnel policies adopted by nations, states or even large school districts.

This observation also raises concerns about the ability of large government

information but rather the best firms can do subject to the information constraints they face.

agencies to determine the reward structures and ranking procedures that govern competition among schools. The benefits of competition among schools will be determined in part by the extent to which policy makers not only choose valid ranking procedures but also attach the correct reward structure to various ranks. Policy makers require enormous amounts of information to perform these tasks well.

5 Conclusion

The great myth about incentive pay or accountability systems is that they bring "business practices" or "competitive pressures" to public education, but such claims are not true. In contrast to private firms, public school systems are not directly accountable to their customers, i.e. the families they serve. In the traditional public school model, teachers and principals, as public employees, are accountable to the appointed agents of elected officials. In accountability or incentive pay systems, teachers and principals are accountable to formulas and procedures created by these same agents. These systems may foster competition to earn the rewards governments offer, but if governments design these competitions poorly, there is no guarantee that governments will correct their mistakes in a timely manner.

Decades ago school boards began to adopt policies that guaranteed salary increases for all teachers that obtained Master's degrees in Education, and our university libraries are now filled with research papers that find no relationship between the acquisition of these degrees and the performance of teachers.[18] Yet, there is no indication that districts intend to break the link between Master's degrees and pay levels any time soon. If state education agencies or school districts adopt incentive pay systems that are just as ill-advised as the decision to grant automatic raises to teachers who obtain a Master's degree, what forces will correct such errors?

Section one describes how hidden actions of agents can corrupt performance statistics. The multi-tasking model demonstrates that once government agencies attach important incentives to a particular statistic, government employees will take actions to improve the value of this statistic even if these actions contribute nothing or do harm to those that their organization is intended to serve. However, the political process may corrupt government performance statistics in a more direct manner if interest groups exert influence over the adoption of specific performance measures and reward schemes for use in incentive pay systems.

The analyses presented thus far have implicitly assumed the existence of a benevolent education authority and described the policies this authority might adopt given its access to information. However, it is easy to imagine ways that ranking procedures and reward structures might be corrupted by the political process. Is it inconceivable that an alliance of teachers' unions and post-secondary schools of education could demand that state officials consider the number of Master's degrees among faculty members or the total

[18] See Walsh and Tracy (2005).

number of hours in professional development classes as a key factor in determining a school's overall performance ranking? It is easy to imagine the adoption of state or district-wide incentive pay systems that specify a mapping between certain performance statistics and total pay according to rules that do not vary at all with grade, subject taught, or school environment, even though it is almost impossible to justify this approach on efficiency grounds.

These observations suggest that voucher systems and state-wide performance measurement systems should not be seen as policy substitutes but rather policies that could work well together. Consider a system that provides parents with comprehensive performance rankings for schools but that also allows parents to use this information as only one of many factors when deciding where their children should attend school. In this scenario, the choices of parents determine the overall budgets of each school, and those who run schools engage in competition for resources by choosing the education and personnel policies that deliver educational services that parents value.

This approach gives parents the opportunity to act on information they possess that cannot be found in any database and also the opportunity to aggregate the information at their disposal based on their values and priorities. By granting parents control over the public funds that are allocated for their children's education, society gains an army of education performance monitors. If parents do not have this control, they have less incentive to acquire information about the quality of their child's school and no means to credibly communicate the information they possess.[19]

Those who are convinced that parents cannot possibly be trusted to choose schools for their children may wish to amend this system by making total school resources dependent not only on student enrollment but also on some government assessment of school quality. But even with such an amendment, a system of real competition among schools may serve as an important catalyst for improving the practices that determine the hiring, retention, and payment of teachers.

It is also worth noting that high-powered incentives may not even be the optimal approach to personnel policy in education. The Holmstrom and Milgrom (1991) multi-tasking model points directly to this possibility. In addition, Besley and Ghatak (2005) note that many non-profit organizations in education, health, or related services choose personnel policies that include relatively little incentive pay. They argue that, in these types of organizations, it is often efficient to devote considerable resources to the screening of potential hires and to then only hire candidates with high levels of personal commitment to the mission of the organization. When it is possible to identify such individuals, incentive pay is no longer necessary.

Current trends in education reform operate on the assumption that teach-

[19] Acemoglu, Kremer, and Mian (2007) argue that real competition among educators may cause harm. They reach this conclusion because parents in their model are not able to monitor schools directly and thus rely on public statistics like test scores. In this setting, Holmstrom and Milgrom's (1991) multi-tasking model suggests that intense competition among educators may waste resources and harm students.

ers should face high-powered performance incentives, but it is possible that this assumption is wrong. It is possible that schools do not need incentive pay systems but rather much better means for identifying and developing talented persons who enjoy helping children learn. Whether or not this is the case, real competition among schools and organizations that manage schools may be the best mechanism available for societal learning about desirable methods for identifying, training and motivating teachers.

Acemoglu, Daron; Kremer, Michael, and Atif Mian. “Incentives in Markets, Firms, and Governments.” *Journal of Law, Economics, and Organization,* forthcoming.

Besley, Timothy and Maitreesh Ghatak . “Public versus Private Ownership of Public Goods.” *Quarterly Journal of Economics* 116 (November 2001): 1343-1372.

Campbell, Donald T. “Assessing the Impact of Planned Social Change.” Paper #8, Occasional Paper Series. Dartmouth College, The Public Affairs Center, December 1976.

Dixit, Avinash. “Incentives and Organizations in the Public Sector.” *The Journal of Human Resources* 37:4 (Autumn 2002) pp. 696-727.

Clotfelter, Charles; Ladd, Helen; Vigdor, Jacob, and Aliaga Diaz. “Do School Accountability Systems Make It More Difficult for Low Performing Schools to Attract and Retain High Quality Teachers.” *Journal of Policy Analysis and Management* 23 (Spring 2004): 251.

Glewwe, Paul; Ilias, Nauman, and Michael Kremer. “Teacher Incentives In Developing Countries: Recent Experimental Evidence from Kenya.” Working Paper 2008-09 (Nashville, Tenn.: National Center for Performance Incentives, February 2008).

Green, Jerry R. and Nancy L. Stokey. “A Comparison of Tournaments and Contracts.” *Journal of Political Economy* 91 (June 1983): 349-364.

Hanushek, Eric A. and Margaret E. Raymond. “Does School Accountability Lead to Improved Student Performance?” Working Paper 10591 (National Bureau of Economic Research, June 2004).

Hanushek, Eric A. “Impacts and Implications of State Accountability Systems” in *Within Our Reach*, edited by John E. Chubb, Rowman & Littlefield. New York, 2005.

Holmström, Bengt, and Paul Milgrom. “Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design.” *Journal of Law, Economics, & Organization* 7 (January 1991): 24-52.

Itoh, Hideshi. “Incentives to Help in Multi-Agent Situations.” *Econometrica* 59 (May 1991): 611-636.

Lavy, Victor. “Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement.” The Journal of Political Economy 110 (December 2002): 1286-1317.

Lazear, Edward P. “Pay Equality and Industrial Politics.” *Journal of Political Economy* 3 (June 1989): 561-580.

Lazear, Edward P. and Sherwin Rosen. **“**Rank-Order Tournaments as Optimum Labor Contracts.” *The Journal of Political Economy* 89 (October 1981): 841-864.

McCaffrey, Daniel F., J.R. Lockwood, Daniel M. Koretz, and Laura S. Hamilton. *Evaluating Value-Added Models for Teacher Accountability.* Santa Monica, CA: RAND Corporation, 2003.

McCaffrey, Daniel F., Bing Han and J.R. Lockwood. "From Data to Bonuses: A Case Study of the Issues Related to Awarding Teachers Pay on the Basis of their Students' Progress." Working Paper 2008-14 (Nashville, Tenn.: National Center for Performance Incentives, February 2008).

McCaffrey, Daniel F., Tim R. Sass, and J.R. Lockwood. "The Intertemporal Stability of Teacher Effect Estimates." unpublished manuscript, June 2008.

Muralidharan, Karthik and Venkatesh Sundararaman. "Teacher Incentives In Developing Countries: Experimental Evidence from India." Working Paper 2008-13 (Nashville, Tenn.: National Center for Performance Incentives, February 2008).

Neal, Derek. "The Role of Private Schools in Education Markets," forthcoming in the *Handbook of Research on School Choice*, edited by Mark Berends, Matthew G. Springer, Dale Ballou, and Herbert J. Walberg, Lawrence Erlbaum Associates/Taylor & Francis Group.

Neal, Derek and Diane Whitmore Scanzenbach. "Left Behind By Design: Proficiency Counts and Test-Based Accountability." University of Chicago, February 2008.

Prendergast, Canice. "The provision of incentives in firms." *Journal of Economic Literature* 37 (March 1999): 7-63.

Reback, Randall. "Teaching to the rating: School accountability and the distribution of student achievement." *Journal of Public Economics* 92 (June 2008): 1394-1415.

Rivkin, Steven G.; Hanushek, Eric A., and John F. Kain. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (March 2005): 417-458.

Rothstein, Jesse. "Do Value-Added Models Add Value? Tracking, Fixed Effects, and Causal Inference." Princeton University, November 20, 2007.

Rothstein, Richard. "Holding Accountability to Account: How Scholarship and Experience in Other Fields Inform Exploration of Performance Incentives in Education." Working Paper 2008-04 (Nashville, Tenn.: National Center for Performance Incentives, February 2008).

Springer, Matthew G. "The Influence of an NCLB Accountability Plan on the Distribution of Student Test Score Gains," *Economics of Education Review,* forthcoming, 2007.

State of California, Department of Education. Office of Policy and Evaluation. "Construction of California's 1999 School Characteristics Index and Similar Schools Ranks," PSAA Technical Report 00-1, April 2000.

Stein, Letitia. "Hillsborough's merit pay experiment benefits affluent schools," *St. Petersburg Times,* February 24, 2008.

Walsh, Kate, and Christopher O. Tracy. "Increasing the Odds: How Good Policies Can Yield Better Teachers." National Council on Teacher Equality, December 2004.

TABLE 1

Elementary Reading Value Table

Elementary Reading							
	Year 2 Level - 2006						
Year 1 Level 2005	1a	1b	2	3	4	5	Average Score
1a	0	100	455	550	675	725	100.0
1b	-50	50	145	265	340	500	100.0
2	-100	-50	125	205	245	350	100.1
3	-175	-100	-90	170	210	250	100.2
4	-200	-150	-140	-75	195	215	100.0
5	-250	-200	-160	-125	25	210	100.2
All Levels							100.1

See http://www.fldoe.org/PerformancePay/pdfs/ValueTable2005-06.pdf