DISCUSSION PAPER SERIES

IZA DP No. 4383

Pay for Percentile

Gadi Barlevy
Derek Neal

August 2009

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Pay for Percentile

Gadi Barlevy
Federal Reserve Bank of Chicago
and IZA

Derek Neal
University of Chicago,
NBER and IZA

Discussion Paper No. 4383
August 2009

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 4383
August 2009

ABSTRACT

Pay for Percentile*

We propose an incentive pay scheme for educators that links educator compensation to the ranks of their students within appropriately defined comparison sets, and we show that under certain conditions our scheme induces teachers to allocate socially optimal levels of effort to all students. Because this scheme employs only ordinal information, our scheme allows education authorities to employ completely new assessments at each testing date without ever having to equate various assessment forms. Thus, our scheme removes incentives for teachers to teach to a particular assessment form and eliminates any opportunities to influence reward pay by corrupting the equating process or the scales used to report assessment results. Having shown that cardinal measures of achievement growth over time are not a necessary ingredient of incentive systems for educators, we note that education authorities can employ our scheme as a means of providing incentives for educators while employing a separate system for measuring growth in student achievement that involves no stakes for educators. This approach creates no incentives for educators to take actions that contaminate the measurement of student progress.

JEL Classification: J33, I20

Keywords: compensation, education, tournaments

Corresponding author:

Gadi Barlevy
Economic Research Department
Federal Reserve Bank of Chicago
230 South LaSalle
Chicago, IL 60604
USA
E-mail: gbarlevy@frbchi.org

* We thank Fernando Alvarez, Ann Bartel, Roger Myerson, Kevin Murphy, and Phil Reny for helpful comments and discussions. Neal thanks Lindy and Michael Keiser for research support through a gift to the University of Chicago's Committee on Education. Neal also thanks the Searle Freedom Trust. Our views need not reflect those of the Federal Reserve Bank of Chicago or the Federal Reserve System.

1. Introduction

In modern economies, most wealth is held in the form of human capital, and publicly funded schools play a key role in creating this wealth. Thus, reform proposals that seek to enhance the efficiency of schools are an omnipresent feature of debates concerning public policy and societal welfare. In recent decades, education policy makers have increasingly designed these reform efforts around measures of school output such as test scores rather than measures of school inputs such as computer labs or student-teacher ratios. Although scholars and policy makers still debate the benefits of smaller classes, improved teacher preparation, or improved school facilities, few are willing to measure school quality using only measures of school inputs. During the 1990s many states adopted accountability systems that dictated sanctions and remediation for schools based on how their students performed on standardized assessments. In 2001, the No Child Left Behind Act (NCLB) mandated that all states adopt such systems or risk losing federal funds, and more recently, several states and large districts have introduced incentive pay systems that link the salaries of individual teachers to the performance of their students.

We pose the provision of teacher incentives as a mechanism design problem. In our setting, an education authority possesses two sets of test scores for a population of students. The first set of scores provides information about student achievement at the beginning (fall) of a school year. The second set provides information about the achievement of the same population of students at the end (spring) of the school year. Taken together, these test scores provide information concerning the effort that teachers invested in their students.

We begin by noting that if the authority knows the mapping between the test score scale and the expected value of student skill, the authority can implement an incentive scheme that pays teachers for the skills that their efforts helped create. Some will contend that social scientists have no idea how to construct such a mapping and therefore argue that such performance pay systems are infeasible.[1] However, even if policy makers are able to discover the mapping between a particular test score scale and the value of student skill, the authority will find it challenging to map scores to skills in practice because attempts to minimize opportunities for coaching behaviors make it difficult to maintain a mapping between test scores and the value of student skill that is consistent across assessments. In order to deter "teaching to the test" and related behaviors, an education authority must employ a series of assessments over time that differ in terms of specific item content and form. But, in order to map results from each assessment into a common scale that measures the value of skills, the authority must equate the various assessment forms, and equating requires common items that link the various forms. If designers limit the number of common items, they advance the goal of preventing teachers from coaching students for specific questions or question formats, but they hinder the goal of properly equating

[1]See Balou (2009) for more on difficulties of interpreting psychometric scales. Cawley et al (1999) addresses the task of using psychometric scales in value-added pay for performance schemes. Cunha and Heckman (2008) describe methods for anchoring psychometric scales to adult outcomes. Their methods cannot be applied to new incentive systems involving new assessments because data on the adult outcomes of test takers cannot be collected before a given generation of students ages into adulthood.

and thus properly scaling the various assessment forms. In addition, because equating is a complex task and proper equating is difficult to verify, the equating process itself is an obvious target for corruption.[2]

Given these observations, we turn our attention to mechanisms that require authorities to make incentive payments based only on the ordinal information contained in assessment results, without any knowledge of how the fall and spring assessments are scaled. Because such systems involve no attempt to equate various assessment forms, they can include completely new assessment forms at each point in time and thus eliminate incentives to coach students regarding any particular form of an assessment.

We describe a system called "pay for percentile," that works as follows. For each student in a school system, first form a comparison set of students against which the student will be compared. Assumptions concerning the nature of instruction dictate exactly how to define this comparison set, but the general idea is to form a set that contains all other students in the system who begin the school year at the same level of baseline achievement in a comparable classroom setting. At the end of the year, give a cumulative assessment to all students. Then, assign each student a percentile score based on his end of year rank *among the students in his comparison set.* For each teacher, sum these within-peer percentile scores over all the students she teaches and denote this sum as a percentile performance index. Then, pay each teacher a common base salary plus a bonus that is proportional to her percentile performance index. We demonstrate that this system can elicit efficient effort from all teachers in all classrooms to all students.

The linear relationship between bonus pay and our index does not imply that percentile units are a natural or desirable scale for human capital. Rather, percentiles within comparison sets tell us what fraction of head-to-head contests teachers win when competing against other teachers who educate similar students. For example, a student with a within-comparison set percentile score of .5 performed as well or better than half of his peers. Thus, in our scheme, his teacher gets credit for beating half of the teachers who taught similar students. We propose a linear relationship between total bonus pay and the fraction of contests won because all of the contests share an important symmetry. Each pits a student against a peer who has the same expected spring achievement when both receive the same instruction and tutoring from their teachers.

The scheme we propose extends the work of Lazear and Rosen (1981). They demonstrate that tournaments can elicit efficient effort from workers when firms are only able to rank the performance of their workers. In their model, workers make one effort choice and compete in one contest. In our model, teachers make multiple effort choices, and these choices may simultaneously affect the outcomes of many contests, but we still find that a common prize for winning each tournament can induce efficient effort. Further, we show that pay for

[2]A significant literature on state level proficiency rates under NCLB suggests that political pressures have compromised the meaning of proficiency cutoff scores in numerous states. States can inflate their proficiency rates by making exams easier while holding scoring procedures constant or by introducing a completely new assessment and then producing a crosswalk between the old and new assessment scale that effectively lowers the proficiency threshold. See Cronin et al (2007)

percentile can elicit efficient effort in the presence of heterogeneous gains from instruction, instructional spillovers, and direct peer effects among students in the same classroom.

Within our framework, it is natural to think of teachers as workers who perform complex jobs that require multiple tasks because each teacher devotes effort to general classroom instruction as well as one-on-one tutoring for each student. Our results offer new insight for the design of incentives in this setting. If employers can form an accurate ordinal ranking of worker performance on each task that defines their job, these rankings are a set of performance indices that may provide a basis for efficient incentive pay.

Many engaged in current education policy debates implicitly argue that proper equating of different exam forms is fundamental to sound education policy because education authorities must be able to document the evolution of the distribution of student achievement over time. However, our results suggest that education authorities should treat the provision of incentives and the documenting of student progress as separate tasks. Equating studies are not necessary for incentive provision, and the equating process is more likely to be corrupted when high stakes are attached to the exams in question.

Even though our scheme does not map transparently into the multi-tasking framework developed by Holmstrom and Milgrom (1991), some may worry that it too provides incentives for teachers to engage in activities that inflate assessment results relative to student subject mastery. Here, we make the assumption that, if the authority gives a new assessment at each point in time, the only way teachers can directly affect the rank of their students is by teaching. Our aim is to address the design of optimal performance pay systems in settings where at least the ordinal information in assessments cannot be contaminated by actions that the education authority cannot observe. Further, it is important to note that we are proposing a mechanism designed to induce teachers to teach the material in a given curriculum. We do not address the often voiced concern that potentially important dimensions of student skill, e.g. creativity and curiosity, may not be included in curricular definitions.

2. Basic Model

Here, we describe our basic model and derive optimal teacher effort for our setting. Assume there are J classrooms, indexed by $j \in \{1, 2...J\}$. Each classroom has one teacher, so j also indexes teachers. We assume all teachers are equally effective in fostering the creation of human capital among their students, and all teachers face the same costs of providing effective instruction. This approach allows us to focus on the task of eliciting effort from teachers, but it does not allow us to address other issues that arise in settings with heterogeneous teachers, such as how teachers should be screened for hiring and retention and who should be assigned to teach which students.

Each classroom has N students, indexed by $i \in 1, 2...N$. Let a_{ij} denote the initial human capital of the i-th student in the j-th class. Students within each class are ordered from least to most able, i.e.

$$a_{1j} \leq a_{2j} \leq \cdots \leq a_{Nj}$$

We assume all J classes are identical, i.e. $a_{ij} = a_i$ for all $j \in \{1, 2, ...J\}$. However, this does not mean that we are restricting our attention to an environment where all classes share a common baseline achievement distribution. The task of determining efficient effort in an entire system can be accomplished by determining efficient effort for each classroom. Thus, the planner may solve the allocation problem for the system by solving the problem we analyze for each baseline achievement distribution that exists in one or more classes.[3]

Teachers can undertake two types of efforts to help students acquire additional human capital. They can tutor individual students or teach the class as a whole. The tutoring instruction is student-specific, and any effort spent on teaching student i will not directly affect any student $i' \neq i$. Classroom teaching benefits all students in the class. Examples include tasks like lecturing or planning assignments.

Let e_{ij} denote the effort teacher j spends on individual instruction of student i, and t_j denote the effort she spends on classroom teaching. The human capital of a student at the end of the period, denoted a'_{ij}, depends on his initial skill level a_i, the efforts of his teacher e_{ij} and t_j, and a shock ε_{ij} that does not depend on teacher effort, e.g. random disruptions to the student's life at home. For now, we assume the production of human capital is separable between the student's initial human capital and all other factors and is linear in teacher efforts.

$$a'_{ij} = g(a_i) + t_j + \alpha e_{ij} + \varepsilon_{ij} \tag{2.1}$$

where $g(\cdot)$ is an increasing function and $\alpha > 0$ measures the relative productivity of classroom teaching versus individual instruction. Here, the productivities of both tutoring effort and classroom instruction are not a function of a student's baseline achievement or the baseline achievement of his classmates. This specification provides a useful starting point because it is analytically tractable. Further, given this production environment, the incentive scheme that we propose below is quite easy to implement. In later sections, we consider more general production technologies.

The shocks ε_{ij} are pairwise independent for any pair (i, j). Let $F(x) \equiv \Pr(\varepsilon_{ij} \leq x)$. We assume there is an associated density $f(x) = \dfrac{dF(x)}{dx}$ that is unimodal and symmetric around 0.

Let X_j denote teacher j's expected income. Then her utility is

$$U_j = X_j - C(e_{1j}, ..., e_{Nj}, t_j) \tag{2.2}$$

where $C(\cdot)$ denotes the teacher's cost of effort. We assume $C(\cdot)$ is increasing in all of its arguments and is strictly convex. We further assume it is symmetric with respect to

[3]Here, we assume the planner takes the composition of each class as given. One could imagine a more general problem where the planner chooses the composition of classrooms and the effort vector for each classroom. However, given the optimal composition of classrooms, the planner still needs to choose the optimal levels of effort in each class. We focus on this second step because we are analyzing the provision of incentives for educators taking as given the sorting of students among schools and classrooms.

individual effort, i.e. let $\mathbf{e_j}$ be any vector of tutoring efforts $(e_{1j}, ..., e_{Nj})$ for teacher j, and let $\mathbf{e'_j}$ be any permutation of $\mathbf{e_j}$, then

$$C(\mathbf{e_j}, t_j) = C(\mathbf{e'_j}, t_j)$$

We also impose the usual boundary conditions on marginal costs. The lower and upper limits of the marginal costs with respect to each dimension of effort are 0 and ∞ respectively. These conditions ensure the optimal plan will be interior. Although we do not make it explicit, $C(\cdot)$ also depends on N. Optimal effort decisions will vary with class size, but the tradeoffs between scales economies and congestion externalities at the center of this issue have been explored by others.[4] Our goal is to analyze the optimal provision of incentives given a fixed class size, N, and here, we suppress reference to N in the cost function.

Let R denote the social value of a unit of a'. Assume that each teacher has an outside option equal to U_0, an omniscient social planner chooses teacher effort levels in each class $j = 1, 2, ...J$ to maximize the following:

$$\max_{\mathbf{e_j}, t_j} E\Big[R \sum_{i=1}^{N} [g(a_i) + t_j + \alpha e_{ij} + \varepsilon_{ij}] - C(\mathbf{e_j}, t_j) - U_0\Big]$$

Because we have normalized units of time so that $\frac{\partial a'_{ij}}{\partial t_j} = 1$, R may also be interpreted as the gross social return per student when one unit of teacher time is effectively devoted to classroom instruction. Since $C(\cdot)$ is strictly convex, first-order conditions are necessary and sufficient for an optimum. Since all teachers share the same cost of effort, the optimal allocation will dictate the same effort levels in all classrooms, i.e. $e_{ij} = e_i$ and $t_j = t$ for all j. Hence, the optimal effort levels dictated by the social planner, $e_1, ..., e_N$ and t, will solve the following system of equations:

$$\begin{aligned} \frac{\partial C(\mathbf{e_j}, t_j)}{\partial e_{ij}} &= R\alpha \qquad \text{for } i = 1, ..., N \\ \frac{\partial C(\mathbf{e_j}, t_j)}{\partial t_j} &= RN \end{aligned}$$

Given our symmetry and separability assumptions, the cost and returns associated with devoting additional instruction time to a student are not a function of the student's baseline achievement or the distribution of baseline achievement in the class. Thus, in this case, the social optimum dictates the same levels of instruction in all classrooms and the same tutoring effort for all students. Let e^* denote the socially optimal level of tutoring effort that is common to all students, t^* denote the efficient level of classroom instruction common to all classes, and ($\mathbf{e}^*$,t^*) denote the socially optimal effort vector common to all classrooms.

[4] See Lazear (2001) for example.

In section 6, we generalize the model to allow heterogeneity in returns from instruction, instructional spillovers, and peer effects, and in this more general setting, optimal tutoring effort and classroom instruction effort will generally vary with the baseline achievement of individual students and their classmates.

3. Performance Pay With Invertible Scales

Now consider the effort elicitation problem faced by an education authority that supervises our J teachers. For now, assume that this authority knows everything about the technology of human capital production but cannot observe teacher effort e_{ij} or t_j. Instead, the authority observes test scores that provide a perfect ranking of students according to their achievement at a point in time, $s = m(a)$ and $s' = m(a')$, where $m(a)$ is a strictly monotonic function.

Suppose the authority knows $m(\cdot)$, i.e. it knows how to invert the psychometric scale s and recover a. In this setting, there are many schemes that the authority can use to induce teachers to provide socially efficient effort levels. For example, the authority could induce teachers to value improvements in student skill correctly simply by paying bonuses per student equal to Ra'_{ij}. However, from the authority's perspective, this scheme would be wasteful because it compensates teachers for both the skill created by their efforts and for the stock of skills that students would have enjoyed without instruction, $g(a)$.[5]

If the authority knows both $m(\cdot)$ and $g(\cdot)$, the authority can form an unbiased estimator, V_{ij}, of teacher j's contribution to student i's human capital,

$$\begin{aligned} V_{ij} &= a'_{ij} - g(a_{ij}) \\ &= m^{-1}(s'_{ij}) - g(m^{-1}(s_{ij})) \end{aligned}$$

and elicit efficient effort by paying teachers RV_{ij} per student. Further, even if the authority does not know $g(\cdot)$, it can still provide incentives for teachers based on their contributions to student skill. For each student i, let the authority form a comparison group comprised of all students with the same initial test score as student i at the beginning of the period. Given the environment we describe above, this set contains the i-th student in each classroom. Next, define $\overline{a}'_i$ as the average achievement for this group at the end of the period, i.e.

$$\overline{a}'_i = \frac{1}{J}\sum_{j=1}^{J} a'_{ij}$$

and consider a bonus schedule that pays each teacher j bonuses linked to the relative performance of her students; specifically $R(a'_{ij} - \overline{a}'_i)$ for each student $i = 1, 2, ...N$. If our

[5]Here, we take the assignment of students to classrooms as fixed, and we are assuming that the education authority cannot simply hold an auction and sell teachers the opportunity to earn Ra'_{ij} per student. However, absent such an auction mechanism, we expect any scheme that pays teachers for skills students possess independent of instruction should create wasteful activities by teachers seeking assignments to high achieving students.

comparison groups are large, teachers will ignore the effect of their choices on $\overline{a}'_i$, and it is straightforward to show that this bonus scheme elicits efficient effort, ($\mathbf{e}^*$,t^*).

Because $\operatorname{plim} \overline{a}'_i = g(a_i) + t^* + \alpha e^*$, the relative achievement of student i, $(a'_{ij} - \overline{a}'_i)$, is not a function of $g(\cdot)$ or a_i in equilibrium. Here, as in the pay for percentile scheme we propose below, teachers receive rewards or penalties for how their students perform relative to comparable students. Both schemes can be implemented without knowledge of the particular baseline scores associated with each baseline achievement level or the score gains achieved by any student.

If the authority knows R and $m(\cdot)$, it can implement this bonus scheme using a standard regression model that includes fixed effects for baseline achievement levels and classroom assignment.[6] Teachers associated with a negative classroom effect will receive a negative bonus or what might be better described as a performance based fine, and total bonus pay over all teachers will be zero. Because expected bonus pay per teacher is zero in this scheme, teachers must receive a base salary that covers their costs. Let X_0 denote the base salary per student. The authority could minimize the cost of eliciting efficient effort by choosing X_0 to satisfy $NX_0 = C(\mathbf{e}^*, t^*) + U_0$.

In this scheme, the focus on variation within comparison sets allows the authority to overcome the fact that it does not know how natural rates of human capital growth, $g(a_i)$, differ among students of different baseline achievement levels, a_i. In the following sections, we demonstrate that by focusing on rank comparisons within comparison sets, the authority can similarly overcome its lack of knowledge concerning how changes in test scores map into changes in human capital at different points on a given psychometric scale.

4. Tournaments

The scheme described above relies on the education authority's ability to translate test scores into the values of students' skills. In order to motivate why the authority might have limited knowledge of how scores map into human capital, suppose the education authority cannot create and administer the assessments but must hire a testing agency to provide $\mathbf{s}$ and $\mathbf{s}'$, the vector of fall and spring test scores for all students. In order to implement the relative performance scheme we describe above, the authority must announce a mapping between the distribution of student test scores, s', and the distribution of reward pay given to the teachers of these students. But, once the authority announces how it will invert $s' = m(a')$, it must guard against at least two ways that teachers may attempt to game this incentive system.

To begin, teachers may coach rather than teach. We have not modeled this activity explicitly, but the existing literature provides much evidence that teachers can inflate student assessment results by giving students the answers to specific questions or having students

[6]For example, if the authority regresses a'_{ij} on only a set of $N+J$ indicator variables that identify baseline achievement groups and classroom assignment, the estimated coefficient on the indicator for teacher j will equal $\frac{1}{N}\sum_{i=1}^{N}(a'_{ij} - \overline{a}'_i)$, and the authority can multiply these coefficients by RN to determine the total bonus payment for each teacher j.

practice taking tests that contain questions in a specific format.[7] Teachers have opportunity and incentive to engage in these behaviors whenever the specific items and format of one assessment can be used to predict the items and format present on future assessments.

In order to deter this activity, the education authority could instruct its testing agency to administer exams each fall and spring that cover the same topics but contain different questions in different formats. However, this instruction would be difficult to follow since existing psychometric methods for equating results from different assessments require that the assessments in question contain common items. Further, even if the testing agency possessed a technology that made it possible to correctly equate results from a series of assessments containing no overlap in item content or format, teachers face a strong incentive to lobby the testing agency to alter the content of the spring assessment or its scaling in a manner that weakens effort incentives. For example, if the pay for relative performance scheme described in the previous section is in place, teachers may secretly pressure the agency to correctly equate various assessments but then report spring scores equal to the correct spring scores divided by some constant greater than one. If teachers believe their lobbying efforts have been successful, each individual teacher's incentive to provide effort will be reduced, but teachers will still collect base salaries that compensate them for the cost of efficient effort levels. Teachers can achieve a similar result by convincing the testing agency to manipulate the content of the spring exam in a way that compresses scores.

Concerns about scale manipulation may seem far fetched to some, but the literature on the implementation of state accountability systems under NCLB contains suggestive evidence that several states inflated the growth in their reported proficiency rates by making assessments easier without making appropriate adjustments to how exams are scored or by introducing new assessments and equating the scales between the old and new assessments in ways that appear generous to the most recent cohorts of students.[8] Given these observations, it is natural to explore the optimal design of teacher incentives in a setting where the testing agency administers new forms of the assessment each period, and the education authority must employ incentive schemes that are scale invariant, i.e. schemes that rely only on ordinal information and are thus implemented without regard to the scaling of various assessment results. In order to develop intuition for our results, we first consider ordinal contests among pairs of teachers. We then examine tournaments that involve simultaneous competition among large numbers of teachers and show that such tournaments are essentially a pay for percentile scheme.

Consider a scheme where each teacher j competes against one other teacher and the results of this contest determine bonus pay for teacher j and her opponent. Teacher j does not know who her opponent will be when she makes her effort choices. She knows only that her opponent will be randomly chosen from the set of other teachers in the system and that her opponent will be facing the same compensation scheme that she faces. Let each teacher

[7] See Jacob (2005), Jacob and Levitt (2002), Klein et al (2000) and Koretz (2002).

[8] See Peterson and Hess (2006) and Cronin et al (2007). Further, in 2006, the state of Illinois saw dramatic and incredible increases in proficiency rates that were coincident with the introduction of a new assessment series. See ISBE (2006).

receive a base pay of X_0 per student, and at the end of the year, match teacher j with some other teacher j' and pay teacher j a bonus $(X_1 - X_0)$ for each student i whose score is higher than the corresponding student in teacher j''s class, i.e. if $s'_{ij} \geq s'_{ij'}$. Since scores increase monotonically with human capital, this rewards the teacher whose student has attained the higher level of human capital by the end of the year. The total compensation for teacher j is thus

$$NX_0 + (X_1 - X_0) \sum_{i=1}^{N} \mathbb{I}(a'_{ij} \geq a'_{ij'})$$

where $\mathbb{I}(A)$ is an indicator that equals 1 if event A is true and 0 otherwise. Because ordinal comparisons determine all payoffs, teacher behavior and teacher welfare are invariant to any re-scaling of the assessment results that preserves ordering.

For each $i \in 1, ..., N$, let us define a new variable $\nu_i = \varepsilon_{ij} - \varepsilon_{ij'}$ as the difference in the shock terms for students in the two classes whose initial human capital is a_i. Let $H(x) \equiv \Pr(\nu_i \leq x)$ denote the distribution of ν_i. We define $h(x) = dH(x)/dx$, and note that given our assumptions about $F(\cdot)$, $H(\cdot)$ is also unimodal, mean zero, and symmetric. Further. we assume that $H(.)$ is twice differentiable.

Since the initial achievement of the students who are compared to each other is identical, the maximization problem for teacher j is

$$\max_{\mathbf{e_j}, t_j} NX_0 + (X_1 - X_0) \sum_{i=1}^{N} H(\alpha(e_{ij} - e_{ij'}) + t_j - t_{j'}) - C(\mathbf{e_j}, t_j) - U_0$$

The first order conditions for each teacher are given by

$$\frac{\partial C(\mathbf{e_j}, t_j)}{\partial e_{ij}} = \alpha h(\alpha(e_{ij} - e_{ij'}) + t_j - t_{j'})(X_1 - X_0) \text{ for } i = 1, 2..N \tag{4.1}$$

$$\frac{\partial C(\mathbf{e_j}, t_j)}{\partial t_j} = \sum_{i=1}^{N} h(\alpha(e_{ij} - e_{ij'}) + t_j - t_{j'})(X_1 - X_0) \tag{4.2}$$

Consider setting the bonus $X_1 - X_0 = R/h(0)$, and suppose both teachers j and j' choose the same effort levels, i.e. $\mathbf{e_j} = \mathbf{e_{j'}}$ and $t_j = t_{j'}$. Then (4.1) and (4.2) become

$$\begin{aligned} \frac{\partial C(\mathbf{e_j}, t_j)}{\partial e_i} &= R\alpha \qquad \text{for } i = 1, ..., N \\ \frac{\partial C(\mathbf{e_j}, t_j)}{\partial t_j} &= RN \end{aligned}$$

Recall that these are the first order conditions for the planner's problem, and thus, the socially optimal effort levels $(\mathbf{e}^*, t^*)$ solve these first order conditions. Nonetheless, the fact that these levels satisfy teacher j's first order conditions is not enough to show that they are optimal responses to the effort decisions of the other teacher. In particular, since $H(\cdot)$ is neither strictly convex nor strictly concave everywhere, the fact that $\mathbf{e_j} = \mathbf{e}^*$ and $t_j = t^*$ satisfy the first order conditions does not imply that these effort choices are global best responses to teacher j' choosing the same effort levels.

Appendix A provides proofs for the following two propositions that summarize our main results for two teacher contests:

Proposition 1: *Let $\widetilde{\varepsilon}_{ij}$ denote a random variable with a symmetric unimodal density and mean zero, and let ε_{ij}=$\sigma\widetilde{\varepsilon}_{ij}$. There exists $\overline{\sigma}$ such that $\forall\sigma > \overline{\sigma}$, both teachers choosing the socially optimal effort levels $(\mathbf{e}^*, t^*)$ is a pure strategy Nash equilibrium of the two teacher contest.*

The intuition behind the variance restriction in this proposition is straightforward. In any given contest, both effort choices and chance play a role in determining the winner. When chance plays a small role, small prizes are optimal because large prizes would induce too much effort. In fact, the optimal bonus, $\frac{R}{h(0)}$, tends to zero as $\sigma \to 0$. Thus, the restriction on σ in Proposition 1 is needed to rule out cases where, given that the other teacher is choosing $(\mathbf{e}^*, t^*)$, the bonus is so small that teacher j's expected gain from responding with $(\mathbf{e}^*, t^*)$ as opposed to some lower effort level does not cover the incremental cost.[9] However, if the element of chance in these contests is important enough, a pure strategy Nash equilibrium exists which involves both teachers choosing the socially optimal effort vectors, $(\mathbf{e}^*, t^*)$, and Proposition 2 adds that this equilibrium is unique.

Proposition 2: *In the two teacher contest, whenever a pure strategy Nash equilibrium exists, it involves both teachers choosing the socially optimal effort levels $(\mathbf{e}^*, t^*)$.*

Taken together, our propositions imply that our tournament scheme can elicit efficient effort from teachers who compete against each other in seeded competitions. Thus, the efficiency properties that Lazear and Rosen (1981) derived for a setting in which two players make one effort choice and compete in a single contest carry over to settings in which two players make multiple effort choices and engage in many contests simultaneously. This is true even when some of these effort choices can affect the outcome of many contests.

Finally, to ensure that teachers are willing to participate in this scheme, we need to make sure that

$$NX_0 + \frac{RN}{2h(0)} - C(\mathbf{e}^*, t^*) \geq U_0$$

Given this constraint, the following compensation scheme minimizes the cost of providing efficient incentives

[9] Lazear and Rosen (1981) require a similar condition for existence in their single task, two person game.

$$X_0 = \frac{U_0 + C(\mathbf{e}^*, t^*)}{N} - \frac{R}{2h(0)}$$
$$X_1 = \frac{U_0 + C(\mathbf{e}^*, t^*)}{N} + \frac{R}{2h(0)}$$

Note that the authority needs only four pieces of information to implement this contest scheme. The authority needs to know each student's teacher, the ranks implied by s and s', and the ratio $\frac{R}{h(0)}$. Recall that R is the gross social return per student generated by one effective unit of classroom instruction. If we stipulate that the authority knows what effective instruction is worth to society but simply cannot observe whether or not effective instruction is being provided, $h(0)$ is the key piece of information that the authority requires.

Here, $h(0)$ is the derivative with respect to classroom instruction, t, of the probability that a given teacher wins one of our contests when both teachers are initially choosing the same effort vectors. It will be difficult for any authority to learn $h(0)$ precisely, but one can imagine experiments that could provide considerable information about $h(0)$. The key observation is that, for many different prize levels other than $\frac{R}{h(0)}$, there exists a symmetric Nash equilibrium among teachers in pure strategies. Thus, given our tournament mechanism and some initial choice for the prize structure, suppose the authority selected a random sample of students from the entire student population and then invited these students to a small number of weekend review classes taught by the authority and not by teachers. If our teachers share a common prior concerning the probability that any one student is selected to participate in these review classes, there will still exist a Nash equilibrium in which both teachers choose the same effort levels. However, given any symmetric equilibrium, the ex post probability that a particular student who received extra instruction will score better than a peer who did not receive extra instruction should increase. Let Δt be the length of the review session. The associated change in the probability of winning is $\Delta p \approx \frac{h(0)+h(\Delta t)}{2}\Delta t$. If we assume that the authority can perfectly monitor instruction quality during these experimental sessions and if we choose a Δt that is a trivial intervention relative to the range of shocks, ε, that affect achievement during the year, the sample mean of $\frac{\Delta p}{\Delta t}$ provides a useful approximation for $h(0)$.[10]

The two-teacher contest system described here elicits efficient effort from teachers and because it relies only on ordinal information, it can be implemented without equating scores scores from different assessment forms. Further, since equating is not required, our scheme allows the education authority to employ completely new assessment forms at each point in time and thereby remove any opportunity for teachers to coach students for specific questions or question formats based on previous assessments. Our scheme is also robust

[10]Our production technology implicitly normalizes the units of ε so that shocks to achievement can be thought of in terms of additions to or deletions from the hours of effective classroom instruction t students receive. Further, because R is the social value of a unit of effective instruction time, the prize $\frac{R}{h(0)}$ determined by this procedure is the same regardless of the units used to measure instruction time, e.g. seconds, minutes, hours.

against efforts to weaken performance incentives by corrupting assessment scales because the mapping between outcomes and reward pay is scale invariant.

5. Pay for Percentile

While our two-teacher contests address some ways that teachers can manipulate incentive pay systems, the fact that each teacher plays against a single opponent may raise concerns about a different type of manipulation. Recall, we assume that the opponent for teacher j is announced at the end of the year after students are tested. Thus, some teachers may respond to this system by lobbying school officials to be paired with a teacher whose students performed poorly. If one tried to avoid these lobbying efforts by announcing the pairs of contestants at the beginning of the year, then one would worry about collusion on low effort levels within pairs of contestants. We now turn to performance contests that involve large numbers of teachers competing anonymously against one another. We expect that collusion on low effort among teachers is less of a concern is this environment.

Suppose that each teacher now competes against K teachers who also have N students. Each teacher knows that K other teachers will be drawn randomly from the population of teachers with similar classrooms to serve as her contestants, but teachers make their effort choices without knowing whom they are competing against. We assume that teachers receive a base salary of X_0 per student and a constant bonus of $(X_1 - X_0)$ for each contest she wins.[11] In this setting, teacher j's problem is

$$\max_{\mathbf{e_j}, t_j} NX_0 + \sum_{k=1}^{K}\sum_{i=1}^{N} H(\alpha(e_{ij} - e_{ik}) + t_j - t_k)(X_1 - X_0) - C(\mathbf{e_j}, t_j) - U_0$$

The first order conditions are given by

$$\frac{\partial C(\mathbf{e_j}, t_j)}{\partial e_{ij}} = \sum_{k=1}^{K} \alpha h(\alpha(e_{ij} - e_{ik}) + t_j - t_k)(X_1 - X_0) \text{ for } i = 1, ..., N \tag{5.1}$$

$$\frac{\partial C(\mathbf{e_j}, t_j)}{\partial t_j} = \sum_{k=1}^{K}\sum_{i=1}^{N} h(\alpha(e_{ij} - e_{ik}) + t_j - t_k)(X_1 - X_0) \tag{5.2}$$

As before, suppose all teachers put in the same effort level, i.e. given any j, $t_j = t_k$ and $\mathbf{e_j} = \mathbf{e_k}$ for $k = 1, ..., K$. In this case, the right-hand side of (5.1) and (5.2) reduce to $\alpha K h(0)(X_1 - X_0)$ and $NKh(0)(X_1 - X_0)$, respectively. Thus, if we set $X_1 - X_0 = \frac{R}{Kh(0)}$ and assume that all teachers choose the socially optimal effort levels, the first order conditions

[11] A constant prize per contest is not essential for eliciting efficient effort, but we view it as natural given the symmetry of the contests.

for each teacher are satisfied. Further, Proposition 1 extends trivially to contests among $K > 2$ teachers. Given a similar restriction on the scale parameter σ from Proposition 1 and a prize $\frac{R}{Kh(0)}$ per student, there exists a pure strategy Nash equilibrium in which all teachers choose the socially optimal levels of effort.

Now let $K = (J-1) \to \infty$ and let A'_i denote a terminal score chosen at random and uniformly from the set of all terminal scores $(a'_{i1}, ..., a'_{iJ})$. Since the distribution $(a'_{i1}, ...a_{i,j-1}, a_{i,j+1}, ..., a'_{iJ})$ converges to the distribution $(a'_{i1}, ...a_{i,j-1}, a_{ij}, a_{i,j+1}, ..., a'_{iJ})$ as $K \to \infty$, it follows that

$$\lim_{K\to\infty} \sum_{k=1}^{K} \frac{\mathbb{I}(a'_{ij} \geq a'_{ik})}{K} = \Pr(a'_{ij} \geq A'_i)$$

and the teacher's maximization problem reduces to

$$\max_{\mathbf{e_j}, t_j} NX_0 + \frac{R}{h(0)} \sum_{i=1}^{N} \Pr(a'_{ij} \geq A'_i) - C(\mathbf{e_j}, t_j) - U_0$$

This pay for percentile scheme is the limiting case of our simultaneous contests scheme as the number of teachers grows large. Thus, a system that pays teachers bonuses that are proportional to the sum of the within comparison set percentile scores of their students can elicit efficient effort from all teachers.

In our presentation so far, comparison sets contain students who share not only a common baseline achievement level but also by assumption share a common distribution of baseline achievement among their peers. However, given the separability we impose on the human capital production function in equation (2.1) and the symmetry we impose on the cost function, student i's comparison set need not be restricted to students with similar classmates. For any given student, we can form a comparison set for this student by choosing all students from other classrooms who have the same baseline achievement level regardless of the distributions of baseline achievement among their classmates. This result holds because the socially optimal allocation of effort $(\mathbf{e}^*, t^*)$ dictates the same level of classroom instruction and tutoring effort from all teachers to all students regardless of the baseline achievement of a given student or the distribution of baseline achievement in his class.

Therefore, given the production technology that we have assumed so far, pay for percentile can be implemented quite easily and transparently in any large school system. The education authority can form one comparison set for each distinct level of baseline achievement and then assign within comparison set percentiles based on the end of year assessment results. In the following section, we consider more general production functions, and in these environments, comparison sets must condition on classroom characteristics. We show that the existence of peer effects, instructional spillovers and other forces that we have not modelled to this point do not alter the efficiency properties of pay for percentile but simply complicate the task of constructing comparison sets.

6. Heterogeneous Gains from Instruction And Other Generalizations

We now generalize the benchmark model above and show that pay for percentile can be used to elicit socially efficient effort from teachers even when optimal effort for a given student varies with his baseline achievement or is affected by the distribution of baseline achievement among his classmates. Let $\mathbf{a}_j = (a_{1j}, ..., a_{Nj})$ denote the initial level of human capital of all students in teacher j's class, where $j \in \{1, ..., J\}$. We allow the production of human capital for each student i in class j to depend quite generally on his own baseline achievement, a_{ij}, the composition of baseline achievement within the class, $\mathbf{a_j}$, the tutoring he receives, e_{ij}, and the tutoring received by all students in his class, $\mathbf{e_j}$. We further allow students at different achievement levels to differ in their rates of learning given various levels of instruction and tutoring, and we allow both direct peer effects and instructional spillovers. Formally, the human capital of student i in classroom j is given by

$$a'_{ij} = g_i(\mathbf{a}_j, t_j, \mathbf{e}_j) + \varepsilon_{ij} \tag{6.1}$$

Because $g_i(\cdot, \cdot, \cdot)$ is indexed by i, this formulation allows different students in the same class to benefit differently from the same environmental inputs, i.e. from their other classmates, their classroom instruction, t_j, and any common level of individual tutoring, $e_{ij} = e_{i'j}$ for $i \neq i'$. Nonetheless, we place two restriction on $g_i(\cdot, \cdot, \cdot)$. It must be weakly concave, and it must depend on class identity, j, only through teacher efforts. Our concavity assumption places restrictions on forms that peer effects and instructional spillovers may take. Our assumption that j enters only through teacher effort choices implies that, for any two classrooms (j, j') with the same composition of baseline achievement, if the two teachers in question choose the same effort levels, i.e. $t_j = t_{j'}$ and $\mathbf{e_j} = \mathbf{e'_j}$, the expected human capital for any two students in different classrooms but with the same initial achievement, i.e. $a_{ij} = a_{ij'}$, will be same. Given this result and the fact that the marginal cost of devoting more effort to either of these students will be the same for both teachers, we can form comparison sets at the classroom level and guarantee that all contests are properly seeded.

For now, we will continue to assume the ε_{ij} are pairwise identically distributed across all pairs (i, j), although we comment below on how our scheme can be modified if the distribution of ε_{ij} is assumed to be identical only for students with similar baseline achievement. In section 2, given our separable form for $g_i(\cdot, \cdot, \cdot)$, we could interpret the units of ε_{ij} in terms of additions to or deletions from effective classroom instruction time. Given the more general formulation of $g_i(\cdot, \cdot, \cdot)$ here, this interpretation need no longer apply in all classrooms. Thus, the units of ε_{ij} can now only be interpreted as additions to or deletions from the stock of student skill.

We maintain our assumption that the cost of spending time teaching students does not depend on their identity, i.e. $C(\mathbf{e}_j, t_j)$ is symmetric with respect to the elements of $\mathbf{e}_j$ and does not depend on the achievement distribution of the students. Our results would not change if we allowed the cost of effort to depend on the baseline achievement distribution

in a class, i.e. $C(\mathbf{a_j}, \mathbf{e}_j, t_j)$, or to be asymmetric, as long as we maintain our assumption that $C(\cdot)$ is strictly convex and is the same for all teachers.

For each class j, the optimal allocation of effort solves

$$\max_{\mathbf{e}_j, t_j} \sum_{i=1}^{N} R[g_i(\mathbf{a}_j, t_j, \mathbf{e}_j) + \varepsilon_{ij}] - C(\mathbf{e}_j, t_j) - U_0 \tag{6.2}$$

Since $g_i(\cdot,\cdot,\cdot)$ is assumed to be concave for all i and $C(\cdot)$ is strictly convex, this problem is strictly concave, and the first-order conditions are both necessary and sufficient for an optimum. These are given for all j by

$$\begin{aligned} \frac{\partial C(\mathbf{e_j}, t_j)}{\partial e_{ij}} &= R \sum_{m=1}^{N} \frac{\partial g_m(\mathbf{a}_j, t_j, \mathbf{e}_j)}{\partial e_{ij}} \qquad \text{for } i = 1, ..., N \\ \frac{\partial C(\mathbf{e_j}, t_j)}{\partial t_j} &= R \sum_{m=1}^{N} \frac{\partial g_m(\mathbf{a}_j, t_j, \mathbf{e}_j)}{\partial t_j} \end{aligned}$$

For any composition of baseline achievement, there will be a unique $\left(\mathbf{e_j^*}, t_j^*\right)$ that solves these equations. However, this vector will differ for classes with different compositions, $\mathbf{a_j}$, and the tutoring effort, e_{ij}, for each student will generally differ across students in the same class if the students have different initial achievement.

We now argue that the pay for percentile scheme we described above will continue to elicit socially optimal effort vectors from all teachers. The bonus scheme is the same as before, and again, each student will be compared to all students with the same baseline achievement who belong to one of the K other classrooms in his comparison set.[12]

Assume that we offer each teacher j a base pay of X_0 per student, and a bonus $X_1 - X_0 = \frac{R}{Kh(0)}$ for each student in any comparison class $k = 1, 2..K$ who scores below his counterpart in teacher j's class on the spring assessment. Teacher j's problem can be expressed as follows:

$$\max_{\mathbf{e_j}, t_j} NX_0 + (X_1 - X_0) \sum_{k=1}^{K} \sum_{i=1}^{N} H(g_i(\mathbf{a}_j, t_j, \mathbf{e_j}) - g_i(\mathbf{a}_k, t_k, \mathbf{e_k})) - C(\mathbf{e_j}, t_j)$$

The first order conditions for teacher j are

[12] As we note at the end of the previous section, this composition restriction on comparison sets is now binding. We noted in the previous section that when $g_i(\cdot,\cdot,\cdot)$ is separable in a_i and teacher's effort, the comparison set for student i may contain any student with the same baseline achievement regardless of the composition of baseline achievement in this student's class.

$$\frac{\partial C(\mathbf{e_j}, t_j)}{\partial e_{ij}} = (X_1 - X_0) \sum_{k=1}^{K} \sum_{m=1}^{N} \frac{\partial g_m(\mathbf{a}_j, t_j, \mathbf{e}_j)}{\partial e_{ij}} h(g_m(\mathbf{a}_j, t_j, \mathbf{e_j}) - g_m(\mathbf{a}_k, t_k, \mathbf{e_k})) \qquad \text{for } i = 1, ..., N$$

$$\frac{\partial C(\mathbf{e_j}, t_j)}{\partial t_j} = (X_1 - X_0) \sum_{k=1}^{K} \sum_{m=1}^{N} \frac{\partial g_m(\mathbf{a}_j, t_j, \mathbf{e}_j)}{\partial t_j} h(g_m(\mathbf{a}_j, t_j, \mathbf{e_j}) - g_m(\mathbf{a}_k, t_k, \mathbf{e_k}))$$

If all teachers provide the same effort levels, these first order conditions collapse to the planner's first order conditions. If we assume that other teachers are choosing the socially optimal levels of effort, then for large enough σ, these first-order conditions are necessary and sufficient for an optimal response. Thus, there exists a Nash equilibrium such that all teachers choose the first best effort levels in response to a common prize structure. However, base pay is not common across all classrooms. Because socially efficient effort levels vary with classroom composition, the level of base pay required to satisfy the teachers' participation constraints will be a function of the specific distribution of baseline achievement that defines a comparison set or a set of competing classrooms.

Here, pay for percentile amounts to competition among teachers within leagues defined by classroom type. These leagues offer properly seeded contests even in the presence of peer effects and heterogeneity in student learning rates. Further, because the competition involves all students in each classroom, teachers internalize the consequences of instructional spillovers.

In practice, it may be impossible to form large comparison sets containing classrooms with identical distributions of baseline achievement. Nevertheless, it may still be possible to implement our system using a large set of quantile regression models that allow researchers to create, for any set of baseline student and classroom characteristics, a set of predicted scores associated with each percentile in the conditional distribution of scores. Given a predicted score distribution for each individual student that conditions on his own baseline achievement and the distribution of baseline achievement among his classmates, education authorities can assign a conditional percentile score to each student and then form percentile performance indices at the classroom level.[13]

Note that, even with our more general formulation for $g_i(\cdot, \cdot, \cdot)$, the optimal prize structure does not vary with baseline achievement and does not depend on the functional form of $g_i(\cdot, \cdot, \cdot)$. This result hinges on our assumption that the distribution of ε_{ij} and thus $h(0)$ do not vary among students, and this assumption is not testable. If we permit returns to effective instruction to vary with baseline achievement, it is straightforward to show that, if the populations of students at two different baseline achievement levels differ with respect to both $h(0)$ and the rate at which they acquire new skill given effective instruction, these differences cannot be separately identified from experiments like those described at the end

[13] See Briggs and Betebenner (2009) for an example of how these conditional percentile scores can be calculated in practice.

of section 4 above.[14] If $h(0)$ differs with baseline achievement, the authority can still elicit efficient effort by using a pay for percentile scheme that offers different prizes for winning contests that involve students of different baseline achievement levels. However, education authorities cannot implement such schemes without prior information concerning how $h(0)$ or $g(\cdot,\cdot,\cdot)$ differs among students.

The key implication of our analyses is that performance pay for educators should be based on ordinal rankings of student outcomes within properly chosen comparison sets. Whether or not the optimal mapping between rankings and reward money is constant over all comparison sets or all students within a comparison set, the decision to tie rewards to relative performance measures that are scale invariant allows the authority to combat teaching to the test behaviors by using a new form of the assessment in each period while also eliminating incentives to corrupt the system by manipulating the scales used to report assessment results.

7. Lessons for Policy Makers

In the previous sections, we describe a simultaneous contest mechanism that can elicit efficient effort from teachers and is robust to certain types of manipulation. In this section, we shift our attention to existing performance pay systems and analyze them in light of the lessons learned from our model. Table 1 summarizes a number of existing pay for performance schemes that are currently in operation.

Our model yields several insights that are important but have not been fully recognized in current policy debates. To begin, our analyses highlight the value of competition as a means of revealing what efficient achievement targets should be. With the exception of TAP, all the systems described in Table 1 permit subjective judgements by principals or education officials concerning ex ante achievement targets for individual students or ex post measurement of teacher performance. We will not develop an explicit model of the negotiations and lobbying efforts that are part of these processes, but we note that one possible outcome is that many or all teachers will receive bonus payments for providing less than efficient effort level. The Performance Related Pay (PRP) system in England directed teachers who were applying for a performance bonus as follows:

[14] In the experiment we describe at the end of section 4, students are given a small amount of extra instruction, and the experiment identifies $h(0)\frac{\partial a_{ij}^{'}}{\partial t_j}$. Because the separable production function employed in Section 4 imposes the normalization, $\frac{\partial a_{ij}^{'}}{\partial t_j}=1$, $h(0)$ is identified. This normalization implies that the shocks that affect $a_{ij}^{'}$ can be interpreted as additions to or deletions from the effective instruction that students receive. However, if both rates of student gains from instruction and the distribution of shocks differ at each point in the baseline achievement distribution, no such normalization is possible, and $h(0)$ is not identified because many different combinations of $h(0)$ and $\frac{\partial a_{ij}^{'}}{\partial t_j}$ will provide ways to rationalize any particular increase in winning percentage that might be induced when a particular type of student receives extra instruction.

"Please summarise evidence that as a result of your teaching your pupils achieve well relative to their prior attainment, making progress as good or better than similar pupils nationally."[15]

Ex post, roughly 77% of teachers with enough experience to be eligible for the bonus received the bonus. Unless the benefits of teaching experience in England are quite extraordinary, this outcome is strong suggestive evidence that education officials adopted a rather lenient interpretation of what "as good or better" means.

Table 1
Recent Pay for Performance Systems in Education

Name	Place	Description
ProComp	Denver	Teachers and principals negotiate achievement targets for individual students.
QComp	Minnesota	Schools develop their own plans for measuring teacher contributions to students' achievement.
TAP	14 States	Statistical VAM method produces teacher performance indices.
MAP	Florida	Districts choose their own method for measuring teacher contribution to achievement.
PRP	England	Teachers submit applications for bonus pay and provide documentation of better than average performance in promoting student achievement.

Notes: Each system employs additional measures of teacher performance that are not directly tied to student assessment results. The descriptions presented here describe how performance statistics derived from test scores are calculated.

In contrast, our pay for percentile system involves endogenous performance thresholds determined by competition among large numbers of contestants, and every dollar of bonus pay won by one teacher is a dollar of bonus pay lost by another teacher. Forcing teachers to compete within our scheme or any other relative performance pay schemes for a fixed amount of bonus money minimizes the possibility that performance pay systems simply become a means of bestowing pay raises on teachers regardless of whether or not they improve their performance.[16] Pay for relative performance systems allows education authorities to implement incentive pay without adopting any procedures that specify ex ante achievement

[15]See Atkinson et al (2009) and Wragg et al (2001). If one uses the pass rate from the Wragg et al (2001) sample to extrapolate the success rate for the population, the application rate found in Atkinson et al (2009) implies that 77.6 percent of eligible teachers received the bonus.

[16]This point is most easily understood in the context of systems that involve subjective judgements concerning achievement targets. However, similar problems may arise when education authorities employ statistical models to develop such targets. In any system that does not involve direct competition for a fixed amount of bonus pay, there will be incentives for teachers to lobby those who develop tests and models of achievement growth to design measurement systems in a manner that makes it easier for teachers to earn bonuses.

targets for students because by focusing on relative student performance within comparison sets, the authority avoids the need to forecast $g_i(\mathbf{a_j}, \mathbf{e}_j^*, t_j^*)$, i.e. expected spring achievement for student i given efficient teacher effort.

Among the entries in Table 1, the Value-Added Model (VAM) approach is the only scheme based on an objective mapping between student test scores and relative performance measures for teachers. The VAM approach embodies the assumption, $s = a$, i.e. the units of a given psychometric scale s are at least proportional to the correct units for expressing the social value of student skill. Advocates of VAM are quick to point out that, given this assumption, VAM models produce performance indices that allow education authorities to both implement relative performance pay schemes and make judgements concerning whether or not the rate of achievement growth is increasing or decreasing over time both at an aggregate level and within a classroom or school.[17] In contrast, our pay for percentile scheme employs only ordinal information and thus provides no information about changes in rates of achievement growth over time. However, we contend that this limit on pay for percentile's ambitions is the key reason to prefer it over VAM approaches.

Donald Campbell (1976) famously claimed that government performance statistics are always corrupted when high stakes are attached to them, and our analyses indicate that Campbell's observation may reflect the perils of trying to accomplish two objectives with one set of performance measures. Systems that try to both provide incentives for teachers and track the evolution of student achievement and educational productivity over time are likely to do neither well because assessment procedures that enhance an education authority's capacity to measure achievement growth consistently over time introduce opportunities for teachers to game assessment-based incentive systems. Systems that track achievement must place results from a series of assessments on a common scale, and the equating process that creates this common scale will not be credible unless the assessments in question contain common items. The existence of these common items invites teachers to coach students based on the specific items and format found in the last assessment. Further, any system that links rewards to cardinal measures of achievement growth may introduce political pressures that corrupt equating procedures and compromise our understanding of how student achievement is changing over time.

In contrast, if education authorities employ our pay for percentile scheme, they can provide incentives for teachers while employing new assessments at each point in time that need not be equated and therefore need not contain common items. The absence of common items makes it difficult for teachers to coach students for particular questions or question formats, and our scheme provides no incentive for anyone to manipulate the scales used to report results since the distribution of reward pay is based only on ranks and is thus scale invariant. If education authorities desire measures of changes in achievement growth

[17] Some VAM practitioners employ functional form assumptions that allow them to also produce universal rankings of teacher performance and make judgements about the relative performance of two teachers even if the baseline achievement distributions in their classes do not overlap. In contrast, our pay for percentile scheme takes seriously the notion that teaching academically disadvantaged students may be a different job than than teaching honors classes and provides a context-specific measure of how well teachers are doing the job they actually have.

over time, they can deploy a second assessment system that is scale dependent but with no stakes attached and with only random samples of schools within the system involved. By separating the tasks of incentive provision and output measurement, education authorities are likely to do both tasks better.

8. Conclusion

Designing a set of assessments and statistical procedures that will not only allow policy makers to measure secular achievement growth over time but also isolate the contribution of educators and schools to this growth is a daunting task in the best of circumstances. Further, when the results of this endeavor determine rewards and punishments for teachers and principals, some educators respond by taking actions that artificially inflate measures of student learning. These actions may include coaching students for assessments as well as lobbying testing agencies concerning how results from different assessments are equated. The high stakes testing literature provides much evidence that teachers coach students for high stakes assessments in ways that inflate assessment results relative to student subject mastery, and the literature on NCLB involves significant debate concerning the integrity of proficiency standards. For example, in 2006 in Illinois, the percentage of eighth graders deemed proficient in math under NCLB jumped from 54.3 to 78.2 in one year. This improvement dwarfs gains typically observed in other years and in other states. Because this enormous gain was coincident with the introduction of a new series of assessments that were scored on a new scale and then equated to previous tests, the entire episode raises suspicions about the comparability of proficiency standards across assessment forms.[18]

We propose an incentive scheme that makes no attempt to compare levels of achievement or achievement growth over time. Because our scheme does not require that education authorities equate results from different assessments, it can be implemented using a sequence of assessments that contain no common items and random variation in item formats. Thus, our system is robust to many forms of teaching to the test that have plagued existing performance pay and accountability systems, and since our scheme does not require testing agencies to equate assessments, it is completely robust to political pressure concerning the scaling of assessment results.

Our key insight is that properly seeded contests where winners are determined by the rank of student outcomes can provide incentives for efficient effort. Thus, the ordinal content of assessment results provides the information that education authorities need in order to elicit socially efficient effort from teachers. Policy makers may still want to know how achievement levels are evolving over time or how the contribution of schools to achievement is evolving over time. However, they will do a better job of providing credible answers

[18]See ISBE 2006. The new system came online when the state began testing in grades other than grades 3, 5, and 8. The introduction of the new assessment system resulted in significant jumps in both reading and math proficiency in all three grades, but the eighth grade math results are the most suspicious. Cronin et al (2007) contends that other states have inflated proficiency results by compromising the comparability of assessment scales over time.

to these questions if they address them using a separate measurement system that has no impact on the distribution of rewards and sanctions among teachers and principals.

We are advocating competition based on ranks as the basis for incentive pay systems that are immune to specific corruption activities that plague existing performance pay and accountability systems, but several details concerning how to organize such competition remain for future research. First and foremost, teachers who teach in the same school should not compete against each other. This type of direct competition could undermine useful cooperation among teachers. Further, although we have discussed all our results in terms of competition among individual teachers, education authorities may wish to implement our scheme at the school or school-grade level as a means of providing incentives for effective cooperation.[19]

In addition, because our scheme is designed to elicit effort from teachers who share the same cost of providing effective effort, it may be desirable to have teachers compete only against other teachers with similar levels of experience, similar levels of support in terms of teacher's aids, and similar access to computers and other resources.[20] While more work remains concerning the ideal means of organizing the contests we describe above, our results demonstrate that education authorities can enjoy important efficiency gains from building incentive pay systems for teachers that are based on the ordinal outcomes of properly seeded contests and that are completely distinct from any assessment systems used to measure the progress of students or secular trends in student achievement.

[19]This approach is particularly attractive if one believes that peer monitoring within teams is effective. New York City's accountability system currently includes a component that ranks school performance within leagues defined by student characteristics.

[20]The task of developing a scheme that addresses unobservable differences in teacher talent remains for future research. We have not yet characterized the optimal system for both screening teachers and providing effort incentives based only on the ordinal information in assessments.

REFERENCES

Balou, Dale. "Test Scaling and Value-Added Measurement," NCPI Working paper 2008-23, December, 2008.

Briggs, Derek and Damian Betebenner, "Is Growth in Student Achievement Scale Dependent," mimeo. April, 2009.

Campbell, Donald T. "Assessing the Impact of Planned Social Change," Occasional Working Paper 8 (Hanover, N.H.: Dartmouth College, Public Affairs Center, December, 1976).

Cawley, John, Heckman, James, and Edward Vytlacil, "On Policies to Reward the Value Added of Educators," *The Review of Economics and Statistics* 81:4 (Nov 1999): 720-727.

Cronin, John, Dahlin, Michael, Adkins, Deborah, and G. Gage Kingsbury. "The Proficiency Illusion." Thomas B. Fordham Institute, October 2007.

Cunha, Flavio and James Heckman. "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Journal of Human Resources* 43 (Fall 2008): 739-780.

Gootman, Elissa. "In Brooklyn, Low Grade for a School of Successes." *New York Times*, September 12, 2008.

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

Illinois State Board of Education. *2006 Illinois State Report Card*, 2006.

Jacob, Brian. "Accountability Incentives and Behavior: The Impact of High Stakes Testing in the Chicago Public Schools," *Journal of Public Economics* 89:5 (2005), 761-796.

Jacob, Brian and Steven Levitt. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics* 118:3 (2003) 843-877.

Klein, Stephen, Hamilton, Laura, McCaffrey, Daniel, and Brian Stecher. "What Do Test Scores in Texas Tell Us?" Rand Issue Paper 202 (2000).

Koretz, Daniel M. "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity," *Journal of Human Resources* 37:4 (2002) 752-777.

Lazear, Edward. "Educational Production." *Quarterly Journal of Economics* 16:3 (Aug 2001) 777-803.

Lazear, Edward and Sherwin Rosen. "Rank Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy* 89:5 (Oct 1981): 841-864.

Peterson, Paul and Frederick Hess. "Keeping an Eye on State Standards: A Race to the Bottom," *Education Next* (Summer 2006) 28-29.

Appendix A: Proofs

Proposition 1: *Let $\widetilde{\varepsilon}_{ij}$ denote a random variable with a symmetric unimodal density and mean zero, and let $\varepsilon_{ij}=\sigma\widetilde{\varepsilon}_{ij}$. There exists $\overline{\sigma}$ such that $\forall\sigma>\overline{\sigma}$, both teachers choosing the socially optimal effort levels $(\mathbf{e}^*,t^*)$ is a pure strategy Nash equilibrium of the two teacher contest.*

Proof of Proposition 1: Define $\tilde{\nu}_{ii}=\widetilde{\varepsilon}_{ij}-\widetilde{\varepsilon}_{ij'}$, and let $\widetilde{H}(x)\equiv\Pr(\widetilde{v}_i\leq x)$. Then, $\widetilde{H}(x/\sigma)=H(x)$. Similarly, we have $h(x)\equiv\dfrac{dH(x)}{dx}=\dfrac{1}{\sigma}\widetilde{h}(x/\sigma)$. Note that

$$h(0)=\frac{1}{\sigma}\widetilde{h}(0)$$

and

$$\begin{aligned}H(\alpha e_i-\alpha e^*+t-t^*) &= \int_{-\infty}^{\alpha e_i-\alpha e^*+t-t^*}h(x)dx\\ &= \int_{-\infty}^{\alpha e_i-\alpha e^*+t-t^*}\frac{1}{\sigma}\widetilde{h}(x/\sigma)dx\end{aligned}$$

The teacher's objective function is given by

$$\max_{\mathbf{e}_j,t_j}NX_0+(X_1-X_0)\sum_{i=1}^{N}H(\alpha e_{ij}-\alpha e^*+t_j-t^*)-C(\mathbf{e}_j,t_j)-U_0$$

If we set $X_1-X_0=R/h(0)$, and use the fact that $\frac{h(x)}{h(0)}=\frac{\widetilde{h}(x/\sigma)}{\widetilde{h}(0)}$ this objective function reduces to

$$\max_{\mathbf{e}_j,t_j}NX_0+R\sum_{i=1}^{N}\left[\int_{-\infty}^{\alpha e_{ij}-\alpha e^*+t_j-t^*}\frac{\widetilde{h}(x/\sigma)}{\widetilde{h}(0)}dx\right]-C(\mathbf{e}_j,t_j)-U_0 \tag{8.1}$$

We first argue that the solution to this problem is bounded in a way that does not depend on σ. Observe that the objective function (8.1) is nonnegative at $(\mathbf{e},t)=(\mathbf{0},0)$. Next, since $\widetilde{h}(\cdot)$ is unimodal with a peak at zero, it follows that $\dfrac{\widetilde{h}(x/\sigma)}{\widetilde{h}(0)}\leq 1$ for all x and so

$$\begin{aligned}\int_{-\infty}^{\alpha e_{ij}-\alpha e^*+t_j-t^*}\frac{\widetilde{h}(x/\sigma)}{\widetilde{h}(0)}dx &= \int_{-\infty}^{-\alpha e^*-t^*}\frac{\widetilde{h}(x/\sigma)}{\widetilde{h}(0)}dx+\int_{-\alpha e^*-t^*}^{\alpha e_{ij}-\alpha e^*+t_j-t^*}\frac{\widetilde{h}(x/\sigma)}{\widetilde{h}(0)}dx\\ &\leq \int_{-\infty}^{-\alpha e^*-t^*}\frac{\widetilde{h}(x/\sigma)}{\widetilde{h}(0)}dx+\alpha e_{ij}+t_j\end{aligned}$$

The objective function in (8.1) is thus bounded above by

$$NX_0+NR\int_{-\infty}^{-\alpha e^*-t^*}\frac{\widetilde{h}(x/\sigma)}{\widetilde{h}(0)}dx+R\sum_{i=1}^{N}(\alpha e_{ij}+t_j)-C(\mathbf{e}_j,t_j)-U_0 \tag{8.2}$$

Next, define the set $U = \left\{ \mathbf{u} \in \mathbb{R}_+^{N+1} : \sum_{i=1}^{N+1} u_i^2 = 1 \right\}$. Any vector $(\mathbf{e}_j, t_j)$ can be uniquely expressed as $\lambda \mathbf{u}$ for some $\lambda \geq 0$ and some $\mathbf{u} \in U$. Given our assumptions on $C(\cdot, \cdot)$, for any vector $\mathbf{u}$ it must be the case that $C(\lambda \mathbf{u})$ is increasing and convex in λ and satisfies the limit $\lim_{\lambda \to \infty} \frac{\partial C(\lambda \mathbf{u})}{\partial \lambda} = \infty$. Since $\lambda R \sum_{i=1}^{N} (\alpha e_{ij} + t_j)$ is linear in λ, for any $\mathbf{u} \in U$ there exists a finite cutoff $\lambda^* (\mathbf{u})$ such that the expression in (8.2) evaluated at $(\mathbf{e}_j, t_j) = \lambda \mathbf{u}$ is negative for all $\lambda > \lambda^* (\mathbf{u})$. Since U is compact, $\lambda^* = \sup \{ \lambda^* (\mathbf{u}) : \mathbf{u} \in U \}$ is well defined and finite. It follows that the solution to (8.1) lies in the bounded set $[0, \lambda^*]^{N+1}$.

Next, we argue that there exists a $\overline{\sigma}$ such that for $\sigma > \overline{\sigma}$, the Hessian matrix of second order partial derivatives for this objective function is negative definite over the bounded set $[0, \lambda^*]^{N+1}$. Define $\pi(t, e_1, ..., e_N) \equiv R \sum_{i=1}^{N} \left[\int_{-\infty}^{\alpha e_{ij} - \alpha e^* + t - t^*} \frac{\widetilde{h}(x/\sigma)}{\widetilde{h}(0)} dx \right]$. Then the Hessian matrix is the sum of two matrices, $\mathbf{\Pi} - \mathbf{C}$, where

$$\mathbf{C} \equiv \begin{bmatrix} \frac{\partial^2 C}{\partial e_1^2} & \frac{\partial^2 C}{\partial e_2 \partial e_1} & \cdots & \frac{\partial^2 C}{\partial t \partial e_1} \\ \frac{\partial^2 C}{\partial e_1 \partial e_2} & \frac{\partial^2 C}{\partial e_2^2} & \cdots & \frac{\partial^2 C}{\partial t \partial e_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 C}{\partial e_1 \partial t} & \frac{\partial^2 C}{\partial e_2 \partial t} & \cdots & \frac{\partial^2 C}{\partial t^2} \end{bmatrix}$$

and

$$\mathbf{\Pi} \equiv \begin{bmatrix} \frac{\partial^2 \pi}{\partial e_1^2} & \frac{\partial^2 \pi}{\partial e_2 \partial e_1} & \cdots & \frac{\partial^2 \pi}{\partial t \partial e_1} \\ \frac{\partial^2 \pi}{\partial e_1 \partial e_2} & \frac{\partial^2 \pi}{\partial e_2^2} & \cdots & \frac{\partial^2 \pi}{\partial t \partial e_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \pi}{\partial e_1 \partial t} & \frac{\partial^2 \pi}{\partial e_2 \partial t} & \cdots & \frac{\partial^2 \pi}{\partial t^2} \end{bmatrix}$$

Since the function $C(\cdot)$ is strictly convex, $-\mathbf{C}$ must be a negative definite matrix. Turning to $\mathbf{\Pi}$, since we assume that $H(\cdot)$ is twice differentiable, it follows that $\widetilde{H}(\cdot)$ is also twice differentiable, and so

$$\mathbf{\Pi} = \frac{R}{\sigma \widetilde{h}(0)} \times \begin{bmatrix} \alpha^2 \widetilde{h}'(\frac{\alpha e_1 - \alpha e^* + t - t^*}{\sigma}) & 0 & \cdots & \alpha \widetilde{h}'(\frac{\alpha e_1 - \alpha e^* + t - t^*}{\sigma}) \\ 0 & \alpha^2 \widetilde{h}'(\frac{\alpha e_2 - \alpha e^* + t - t^*}{\sigma}) & \cdots & \alpha \widetilde{h}'(\frac{\alpha e_2 - \alpha e^* + t - t^*}{\sigma}) \\ \vdots & \vdots & \ddots & \vdots \\ \alpha \widetilde{h}'(\frac{\alpha e_1 - \alpha e^* + t - t^*}{\sigma}) & \alpha \widetilde{h}\prime(\frac{\alpha e_2 - \alpha e^* + t - t^*}{\sigma}) & \cdots & \sum_{i=1}^{N} \widetilde{h}'(\frac{\alpha e_i - \alpha e^* + t - t^*}{\sigma}) \end{bmatrix}$$

For a fixed x, all of the elements in $\mathbf{\Pi}$ converge to positive multiples of $\widetilde{h}'(0)$ as $\sigma \to \infty$. Since $\widetilde{h}'(0) = 0$, we have $\mathbf{\Pi} \to \mathbf{0}$ uniformly as $\sigma \to \infty$ within the bounded set $[0, \lambda^*]^{N+1}$. Since $\mathbf{C}$ is positive definite and $\mathbf{\Pi} \to 0$, it follows that there exists a $\overline{\sigma}$ such that for all $\sigma > \overline{\sigma}$, the matrix $\mathbf{\Pi} - \mathbf{C}$ is negative definite for all values of $(\mathbf{e}_j, t_j) \in [0, \lambda^*]^{N+1}$. Hence, the objective function is strictly concave in the region that contains the global optimum, ensuring the first-order conditions are both necessary and sufficient to define a global maximum. ■

Proposition 2: *In the two teacher contest, whenever a pure strategy Nash equilibrium exists, it involves both teachers choosing the socially optimal effort levels* $(\mathbf{e}^*, t^*)$.

Proof of Proposition 2:

We begin our proof by establishing the following Lemma:

Lemma: Suppose $C(\cdot)$ is a convex differentiable function which satisfies standard boundary conditions concerning the limits of the marginal costs of each dimension of effort as effort on each dimension goes to 0 or ∞. Then for any positive real numbers $a_1, ..., a_N$ and b, there is a unique solution to the system of equations

$$\begin{aligned} \frac{\partial C(e_1, ..., e_N, t)}{\partial e_i} &= a_i \qquad\qquad \text{for } i = 1, ..., N \\ \frac{\partial C(e_1, ..., e_N, t)}{\partial t} &= b \end{aligned}$$

Proof: Define a function $bt + \sum_{i=1}^{N} a_i e_i - C(e_1, .., , e_N, t)$. Since $C(\cdot)$ is strictly convex, this function is strictly concave, and as such has a unique maximum. The boundary conditions, together with the assumption that $a_1, ..., a_N$ and b are positive, ensure that this maximum must be at an interior point. Because the function is strictly concave, this interior maximum and the solution to the above equations is unique, as claimed. ■

Armed with this lemma, we can demonstrate that any pure strategy Nash equilibrium of the two teacher contest involves both teachers choosing the socially optimal effort levels. Note that, given any pure strategy Nash equilibrium, both teacher's choices will satisfy the first order conditions for a best response to the other teacher's actions. Further, since h(.) is symmetric, we know that given the effort choices of j and j',

$$h(\alpha(e_{ij} - e_{ij'}) + t_j - t_{j'}) = h(\alpha(e_{ij'} - e_{ij}) + t_{j'} - t_j)$$

In combination, these observations imply that any Nash equilibrium strategies, $(\mathbf{e_j}, t_j)$ and $(\mathbf{e_{j'}}, t_{j'})$, must satisfy

$$\begin{aligned} h(0)\frac{\partial C(\mathbf{e_j}, t_j)}{\partial e_{ij}} &= R\alpha h(\alpha(e_{ij} - e_{ij'}) + t_j - t_{j'}) \\ &= R\alpha h(\alpha(e_{ij'} - e_{ij}) + t_{j'} - t_j) = h(0)\frac{\partial C(\mathbf{e_{j'}}, t_{j'})}{\partial e_{ij'}} \end{aligned}$$

and

$$h(0)\frac{\partial C(\mathbf{e_j}, t_j)}{\partial t_j} = R\alpha h(\alpha(e_{ij} - e_{ij'}) + t_j - t_{j'})$$
$$= R\alpha h(\alpha(e_{ij'} - e_{ij}) + t_{j'} - t_j) = h(0)\frac{\partial C(\mathbf{e_{j'}}, t_{j'})}{\partial t_{j'}}$$

Our lemma implies that these equations cannot be satisfied unless $e_{ij} = e_{ij'} = e^*$ for all $i = 1, ..., N$ and that $t_j = t_{j'} = t^*$. The only pure-strategy equilibrium possible in our two teacher contests is one where teachers invest the classroom instruction effort and common level of tutoring that are socially optimal. ■