American Economic Review 2012, 102(5): 1805–1831
http://dx.doi.org/10.1257/aer.102.5.1805

Pay for Percentile[†]

By Gadi Barlevy and Derek Neal*

We propose an incentive scheme for educators that links compensation to the ranks of their students within comparison sets. Under certain conditions, this scheme induces teachers to allocate socially optimal levels of effort. Moreover, because this scheme employs only ordinal information, it allows education authorities to employ completely new assessments at each testing date without ever having to equate various assessments. This removes incentives for teachers to teach to a particular assessment form and eliminates opportunities to influence reward pay by corrupting assessment scales. Education authorities can employ separate no-stakes assessment systems to track trends in scaled measures of student achievement. (*JEL* I21, I28, J33, J45)

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 policymakers 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 policymakers 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.

Holmstrom and Milgrom (1991) raise the possibility that assessment-based incentive systems for educators may induce teachers to take hidden actions that increase

* Barlevy: Economic Research Department, Federal Reserve Bank of Chicago, 230 S. LaSalle, Chicago, IL 60604 (e-mail: gbarlevy@frbchi.org); Neal: Department of Economics and the Committee on Education, University of Chicago, 1126 East 59th Street, Chicago, IL 60637 (e-mail: d-neal@uchicago.edu). We thank Fernando Alvarez, Julian Betts, Ann Bartel, Robert Gibbons, Richard Holden, John Kennan, Kevin Lang, Roger Myerson, Kevin Murphy, Canice Prendergast, Phil Reny, Doug Staiger, Chris Taber, and Azeem Shaikh for helpful comments and discussions. We thank Eric Nielsen, Richard Olson, and Armin Rick for excellent research assistance. 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.

[†] To view additional materials, visit the article page at http://dx.doi.org/10.1257/aer.102.5.1805.

assessment-based performance measures without creating comparable increases in student learning, and much empirical work on high-stakes testing systems provides evidence that this concern is well founded.[1] We explore how education authorities may limit the scope of these hidden actions by improving the design of assessment-based incentive systems.

We begin by noting that if the authority knows the mapping between the test score scale and the true expected value of student skill, it can implement efficient effort using an incentive scheme that pays teachers for the skills that their efforts help create. Some might contend that, in practice, social scientists cannot construct mappings between test scores and skill values and that such performance pay systems are therefore infeasible,[2] but we focus on other concerns about assessment scales that are more closely related to Holmstrom and Milgrom's original insights.

Even if policymakers 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 maintain this mapping across a series of assessments. In order to maintain the integrity of a scale, the authority typically needs to administer a series of assessments that contain significant overlap in terms of item format and content. Scores on a particular assessment often become inflated, however, when teachers have the opportunity and incentive to coach students concerning the format or items expected on that assessment. Many studies provide evidence of this type of coaching, and Neal (2012) reviews this literature. Stecher (2002) cites two particular districts where, in response to the introduction of high-stakes testing, teachers eliminated assignments that involved students composing original written work and instead asked students to practice finding mistakes in prepared passages. Stecher also reports how some teachers in Arizona responded to high-stakes testing by only assigning math story problems that followed a specific format used on the Iowa Test of Basic Skills. These pedagogical choices make sense as strategies for maximizing student scores on standardized tests, but they seem less than optimal ways of creating good writers or students who can use math tools to solve a broad range of applied problems.[3]

In order to deter teachers from coaching instead of teaching, education authorities may wish to limit the predictability of future assessments by employing a series of assessments that differ in terms of specific item content and form. But, in order to map results from each assessment into a common scale, the authority must equate the various assessment forms, and proper equating often requires common items that link the various forms.[4] If designers limit the number of common items, they

[1] Baker (1992) presents a related model of how principals provide incentives for agents who may take hidden actions without explicit reference to assessment-based incentive schemes in education.

[2] Cunha and Heckman (2008) do 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. See Ballou (2009) for more on difficulties of interpreting psychometric scales. Cawley, Heckman, and Vytlacil (1999) address the task of using psychometric scales in value-added pay for performance schemes.

[3] Other studies that provide either direct or indirect evidence of coaching in response to assessment-based incentives include Koretz and Barron (1998); Klein et al. (2000); Koretz (2002); Jacob (2005); Vigdor (2009); Glewwe, Ilias, and Kremer (2010); and Carrell and West (2010).

[4] A common alternative approach involves randomly assigning one of several alternate forms to students in a large population, and then developing equating procedures based on the fact that the distribution of achievement in the population receiving each form should be constant. In practice, this approach also invites coaching because, at the beginning of the second year of any test-based incentive program, educators know that each of their students will receive one of the forms used in the previous period.

advance the goal of preventing teachers from coaching students for specific questions or question formats, but they hinder the goal of properly equating 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.[5]

One way to address these issues is to consider incentive systems that rely only on the ordinal information contained in assessment results. If the mapping between assessment outcomes and teacher rewards is scale invariant, education officials do not need to place all assessments on a common scale, and they can implement performance pay schemes using a series of assessments that contain no repeated items and no common format.

We describe an ordinal 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.

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 0.5 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. A linear relationship between total bonus pay and the fraction of contests won works because all of the contests share an important symmetry. Each pits a student against a peer who has the same expected final achievement when both receive the same instruction and tutoring from their teachers.

Previous work on the use of contests as incentive mechanisms has typically dealt with environments where workers choose actions to produce a single product. Lazear and Rosen (1981) describe a contest between two agents who choose one effort level and each produce one output. Nalebuff and Stiglitz (1983) as well as Green and Stokey (1983) describe contests between many agents who still choose only one effort level and each produce one output, and in all of these papers, the principal is able to rank all contestants according to their total output. Our application

[5] 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).

involves contestants (teachers) who allocate effort among many tasks (individual tutoring and classroom-wide activities) that jointly create multiple outputs (human capital gains for each student). Because the authority can rank the performance of individual students relative to their peers, the authority can assign a set of performance rankings to each teacher that describes her relative performance concerning the training of each of her students. But, because ranks provide no information about the distances between any two students, the education authority cannot rank teachers according to some composite measure of total classroom performance. Our results imply that such a ranking is not necessary. Principals who hire agents to produce multiple forms of output can elicit efficient effort from these agents even if they are not able to rank them with respect to their total productivity. In the context of current education policy debates, our insights mean that education authorities can implement effective incentive schemes for teachers without ever forming aggregate statistical measures of what individual teachers have contributed to the learning of all students in their classrooms.

Since our system employs only the ordinal ranks implied by assessment outcomes, it provides no information about the secular growth of achievement in a school district over time. Some may view this as a serious shortcoming because public education authorities typically want to track the evolution of student achievement over time using a consistently scaled metric. We argue, however, that education authorities should treat the provision of incentives and the documenting of student progress as separate tasks. Our results show that education officials can provide effective incentives without ever placing high-stakes assessments on a common scale. Thus, authorities can use our scheme for incentive provision and then measure secular achievement growth using a separate low-stakes assessment system that is designed to promote proper equating among assessments. This two-system approach should produce better incentives and better measures of achievement growth.

After presenting our theoretical results, we discuss issues that may arise in the implementation of pay for percentile. Our scheme requires education authorities to estimate within-comparison set percentile scores for students. Wei and He (2006) and Betebenner (2009) have recently developed methods that allow researchers to estimate the distribution of assessment scores holding constant any set of baseline student characteristics. For any student, his actual score combined with an estimate of the distribution of scores given his characteristics yields an estimate of his percentile relative to his peers. These new methods are currently being used to implement or develop new measures of school performance in several states, and they can also be used to calculate our percentile performance index.

Although our system has several desirable properties, it does not address all of the hidden action concerns that scholars and policymakers have raised when evaluating assessment-based incentives schemes. Our system alone does nothing to address the various forms of outright cheating that test-based incentive systems often invite. Further, it does not address the concern that potentially important dimensions of student skill, e.g., creativity and curiosity, may not be included in curricular definitions.[6] Our scheme provides incentives for teachers to build student skills that can be

[6] Holmstrom and Milgrom (1991) also raised this concern.

assessed and ranked. To the extent that schools desire teachers to pursue objectives that cannot be assessed using standard testing procedures, the scheme we propose may be viewed as one component of a larger set of personnel practices that education officials may employ to direct teacher effort.

I. Basic Model

In this section, we illustrate our main results using a simple educational production function. In later sections, we consider a production technology that permits both peer effects and instructional spillovers within classrooms. Assume there are J classrooms, indexed by $j \in \{1, 2, \ldots, J\}$. Each classroom has one teacher, so j also indexes teachers. For now, 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. We discuss teacher heterogeneity in Section V.

Each classroom has N students, indexed by $i \in \{1, 2, \ldots, N\}$. Let a_{ij} denote the initial human capital of the ith student in the jth 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 $j \in \{1, 2, \ldots, J\}$. This does not mean, however, that our analysis applies only to an environment where all classes share a common baseline achievement distribution. The task of determining efficient effort for a school system that contains heterogeneous classes can be accomplished by determining efficient effort for each classroom type. 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.[7]

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. 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. Here, we treat t_j as a scalar. All of our results remain, however, if t_j is a vector of teacher activities—e.g., lesson planning, drafting assignments, etc.—that provide benefits directly to all students in the classroom. We assume the following educational production function:

$$a'_{ij} = g(a_i) + t_j + \alpha e_{ij} + \varepsilon_{ij}. \tag{1}$$

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. The

[7] 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. Given the optimal composition of classrooms, however, 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.

production of human capital is thus linear in teacher efforts and separable between the student's initial human capital and all other factors. Tutoring instruction, e_{ij}, is student-specific, and any effort spent on teaching student i will not directly affect any other student. Classroom teaching, t_j, benefits all students in the class. The function $g(\cdot)$ is increasing and $\alpha > 0$ measures the relative productivity of classroom teaching versus individual instruction, and neither the productivity of tutoring effort nor classroom instruction depend on a student's baseline achievement or the baseline achievement of his classmates.

The shocks ε_{ij} are mean zero, pairwise independent for any pair (i,j), and identically distributed according to a common continuous distribution $F(x) \equiv \Pr(\varepsilon_{ij} \leq x)$.

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

$$U_j = W_j - C(e_{1j}, \ldots, e_{Nj}, t_j), \tag{2}$$

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

$$C(\mathbf{e}_j, t_j) = C(\pi(\mathbf{e}_j), t_j).$$

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 socially 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 trade-offs between scale economies and congestion externalities at the center of this issue have been explored by others.[8] Our goal is to analyze the optimal provision of incentives given a fixed class size, N, so we suppress reference to N in the cost function.

Let R denote the social value of a unit of a'. Because all students receive the same benefit from classroom instruction, we can normalize units of time such that R can also be interpreted as the gross social return per student when one unit of teacher time is devoted to effective classroom instruction. Assume that each teacher has an outside option equal to U_0, which we assume is smaller than the maximum surplus each teacher can generate in the classroom. This assumption implies that teaching is socially valuable for all $j \in \{1, 2, \ldots, J\}$. An omniscient social planner would choose teacher effort levels in each class to maximize the following:

$$\max_{\mathbf{e}_j, t_j} E\left[R\sum_{i=1}^{N}\left[g(a_i) + t_j + \alpha e_{ij} + \varepsilon_{ij}\right] - C(\mathbf{e}_j, t_j)\right].$$

Since $C(\cdot)$ is strictly convex, first-order conditions are necessary and sufficient for an optimum. Given that 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

[8] See Lazear (2001), for example.

all j. Hence, the optimal effort levels dictated by the social planner, $e_1^*, \ldots, e_N^*$ and t^*, will solve the following system of equations:

$$\frac{\partial C(\mathbf{e}_j^*, t_j^*)}{\partial e_{ij}} = R\alpha \qquad \text{for } i \in \{1, 2, \ldots, N\}$$

$$\frac{\partial C(\mathbf{e}_j^*, t_j^*)}{\partial t_j} = RN.$$

Let e^* denote the socially optimal level of individual tutoring effort that is common to all students and $(\mathbf{e}^*, t^*)$ denote the socially optimal effort vector common to all classrooms. When we generalize the model to allow heterogeneity in returns from instruction, instructional spillovers, and peer effects, the optimal tutoring effort and classroom instruction for a given student varies with the baseline achievement of both the student and his classmates. The mechanism we propose below, however, can still elicit efficient instruction and tutoring from each teacher.

II. Performance Pay with Invertible Scales

Next, we consider the 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, both the authority and teachers only observe test scores that provide a 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 increasing function. For now, assume that this ranking is perfect. Below, we discuss how the possibility of measurement error in the achievement ranks implied by test scores affects our analysis.

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}. From the authority's perspective, however, 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, $g(a_{ij})$.[9]

If the authority knows both $m(\cdot)$ and $g(\cdot)$, it can elicit efficient effort while avoiding this waste by forming an unbiased estimator, V_{ij}, of teacher j's contribution to student i's human capital,

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

[9] 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. Absent such an auction mechanism, however, we expect any scheme that pays teachers for skills students possess independent of instruction would induce wasteful activities by teachers seeking assignments to high-achieving students.

and then paying teachers RV_{ij} per student. Further, even if the authority does not know $g(\cdot)$, it can still provide incentives for teachers based only on their contributions to student skill. For each student i, let the authority form a comparison group composed of all students with the same initial test score as student i at the beginning of the period; i.e., the ith students from all classrooms. 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 $i \in \{1, 2, \ldots, N\}$. If J is 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, $(\mathbf{e}^*, t^*)$.[10]

Because plim $\overline{a}_i' = g(a_i) + t^* + \alpha e^*$, the relative achievement of student i, $(a_{ij}' - \overline{a}_i')$, does not depend on $g(\cdot)$ or a_i in equilibrium. 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.

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.[11] Teachers associated with a negative classroom effect will receive below-average salaries and teachers with a positive classroom effect will receive above-average salaries. Average pay must cover teacher costs, $C(\mathbf{e}^*, t^*) + U_0$. Thus, the authority must take care to choose a base salary such that expected total compensation covers these costs.

III. Tournaments

The scheme described in Section II 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 hires a testing agency to provide $\mathbf{s}$ and $\mathbf{s}'$, the vectors of baseline and final test scores for all students. To implement the relative performance scheme we describe above, the authority must announce a mapping between the distribution of student test scores, $\mathbf{s}'$, and the distribution of reward pay given to the teachers of these students. But, once the authority announces how it will map $\mathbf{s}'$

[10] As an alternative, one can calculate performance relative to teacher-specific means that do not involve the scores of a teacher's own students; i.e., $\overline{a}_{ij}' = \frac{1}{J-1}\sum_{k \neq j} a_{ik}'$.

[11] 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.

into reward pay, it must guard against at least two ways that teachers may attempt to game this incentive system.

First, standard methods for placing the results of assessments given at different points in time on a common scale require the testing agency to administer a series of assessments that contain overlap in terms of item content and format. The presence of this overlap, however, allows teachers to use the specific items and format of one assessment to predict the items and format on future assessments, and the existing literature provides much evidence that when teachers have prior knowledge concerning the content of high-stakes assessments, they often inflate student assessment results by providing students with the answers to specific questions or having students practice taking tests that contain questions in a specific format.

Second, taking as given any set of procedures that a testing agency announces concerning how it will administer a series of assessments and place results from these assessments on a common scale, teachers face a strong incentive to lobby the testing agency to alter its scaling procedures in ways that weaken effort incentives. Concerns about scale manipulation may seem far-fetched to some, but the literature on the implementation of state accountability systems under NCLB contains 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.[12] Teachers facing the relative performance pay scheme described in Section II above benefit if they can secretly pressure the testing agency to correctly equate various assessments but then report scores that are artificially compressed. If teachers know that reported scores will be artificially compressed, each teacher faces weaker incentives, but teachers still collect the same expected salary.[13]

Online Appendix A fleshes out in more detail the ways that scale-dependent systems invite coaching and scale manipulation. Given these concerns, we explore the optimal design of teacher incentives, restricting attention to incentive schemes that are scale invariant; i.e., schemes that rely only on ordinal information and can thus be implemented without regard to scaling and without the presence of repeated test items or formats. 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 pay for percentile schemes. While there is a large previous literature on tournaments, the problem facing our education authority requires that we generalize existing models of economic contests to address settings where contestants produce several different outputs simultaneously; e.g., human capital for different students in the same class. Our analysis explores whether tournaments can be used to elicit efficient effort from workers who each produce many outputs simultaneously, including in settings where it is possible to rank worker performance with respect to particular outputs but not with respect to total output.

[12] See Peterson and Hess (2006), Cronin et al. (2007), and Neal (2010).

[13] There are other ways to artificially compress the distribution of scores. The testing agency can also manipulate the content of the exam without making appropriate adjustments to the procedures used to score the exams.

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 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 receive a base pay of X_0 per student, and at the end of the year, match teacher j with some other teacher k and pay teacher j a bonus $(X_1 - X_0)$ for each student i whose score is higher than the corresponding student in teacher k's class; i.e., if $s'_{ij} \geq s'_{ik}$. The total compensation for teacher j is thus

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

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 rescaling of the assessment results that preserves ordering.

For each $i \in 1, 2, \ldots, N$, let us define a new variable $\nu_i = \varepsilon_{ij} - \varepsilon_{ik}$ as the difference in the shock terms for students in the two classes whose initial human capital is a_i. If both teachers j and k choose the same effort levels, then $s'_{ij} > s'_{ik}$ if and only if $\nu_i > 0$. Let $H(x) \equiv \Pr(\nu_i \leq x)$ denote the distribution of ν_i. We assume $H(\cdot)$ is twice differentiable and define $h(x) = dH(x)/dx$. Since ε_{ij} and ε_{ik} are i.i.d, ν_i has mean zero, and $h(\cdot)$ is symmetric around zero. Moreover, $h(x)$ attains its maximum at $x = 0$.[14]

In our framework, ν_i is the only source of uncertainty in contests, and since test scores rank students perfectly according to their true achievement levels, ν_i reflects shocks to true achievement. Some readers may wonder how our analyses change if test scores measure true achievement levels with error. Incorporating measurement error makes our notation more cumbersome, but the presence of measurement error in test scores does not alter our basic results if we assume that the errors in measurement are drawn independently from the same distribution for all students. Suppose test scores are given by $s = m(a + \delta)$, where δ is a random variable with mean zero drawn independently for all students from a common distribution; i.e., each student's test score depends on his own human capital, but imperfections in the testing technology create idiosyncratic deviations between the student's true skill and the skill level implied by his test score. In this environment, when both teachers j and k choose the same effort levels, $s'_{ij} > s'_{ik}$ if and only if $\nu_i > 0$, where now

$$\nu_i \equiv \left[g(m^{-1}(s_{ij}) - \delta_{ij}) + \varepsilon_{ij} + \delta'_{ij}\right] - \left[g(m^{-1}(s_{ik}) - \delta_{ik}) + \varepsilon_{ik} + \delta'_{ik}\right].$$

As in the case without measurement error in test scores, ν_i is mean zero, and its density is symmetric around zero and maximal at zero. These are the properties of ν_i that we require when proving the results presented below. To simplify our

[14] The fact that the density of a random variable equal to the difference of two i.i.d. random variables peaks at zero is discussed in Vogt (1983).

exposition, we proceed under the assumption that test scores provide a perfect ranking of student achievement levels.[15]

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

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

The first-order conditions for each teacher are given by

$$(3) \quad \frac{\partial C(\mathbf{e}_j, t_j)}{\partial e_{ij}} = \alpha h(\alpha(e_{ij} - e_{ik}) + t_j - t_k)(X_1 - X_0) \qquad \text{for } i = 1, 2, \ldots, N$$

$$(4) \quad \frac{\partial C(\mathbf{e}_j, t_j)}{\partial t_j} = \sum_{i=1}^{N} h(\alpha(e_{ij} - e_{ik}) + t_j - t_k)(X_1 - X_0).$$

Assume that base pay X_0 is set at a level that induces all teachers to participate and consider setting the bonus $X_1 - X_0 = R/h(0)$. If both teachers j and k choose the same effort levels, i.e., $\mathbf{e}_j = \mathbf{e}_k$ and $t_j = t_k$, then equations (3) and (4) become

$$\frac{\partial C(\mathbf{e}_j, t_j)}{\partial e_i} = R\alpha \qquad \text{for } i \in \{1, 2, \ldots, N\}$$

$$\frac{\partial C(\mathbf{e}_j, t_j)}{\partial t_j} = RN.$$

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 global best 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 $t_j = t^*$ solves the first-order conditions does not imply that these effort choices are optimal responses to teacher k choosing the same effort levels.

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

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

[15] Note that if measurement error takes the form of a shock that is common to all test takers—e.g., the questions on a particular exam are either too difficult or too easy given how the exam is scored—it enters as a common component of each student's error term that differences out of ν_i. Our contest scheme, like others in the tournament literature, is thus robust to any type of common shock because these shocks do not affect contestant ranks.

As the previous literature on contests involving a single dimension of effort demonstrates, the variance restriction in this proposition is essential. In any given contest, both effort choices and chance play a role in determining the winner. When chance plays a small role, $(\mathbf{e}^*, t^*)$ will not be a best response to the other teacher choosing $(\mathbf{e}^*, t^*)$ unless prizes are small because, when chance plays a small role, the derivative of the probability of winning a given contest with respect to teacher effort is large. In fact, the bonus, $\frac{R}{h(0)}$, tends to zero as $\sigma \to 0$. The restriction on σ in Proposition 1 is needed to rule out cases where, given 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 effort cost. If the element of chance in these contests is important enough, however, a pure strategy Nash equilibrium exists that 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 a 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 each make one effort choice and produce a single output carry over to settings in which two players each make multiple effort choices that jointly determine the levels of multiple outputs.

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:

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

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

In order to choose this cost-minimizing level of base pay, X_0, the authority needs to know the teachers' outside option, U_0, and the cost of providing efficient effort, $C(\mathbf{e}^*, t^*)$. Given any base pay level that satisfies the teachers' participation constraint, however, the authority needs only four additional pieces of information to implement an efficient set of contests. It needs to know each student's teacher, the ranks implied by $\mathbf{s}$ and $\mathbf{s}'$, 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.[16]

The expression $h(0)$ equals the derivative with respect to classroom instruction, t, of the probability that a given teacher wins one of our contests when both teachers choose 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 $h(0)$. Suppose the authority initially chose some arbitrary prize level. If this prize level is high enough, there will exist a symmetric Nash equilibrium in which effort is positive, and this fact allows the authority to run experiments that provide information about $h(0)$ without ex ante knowledge of the optimal prize level.

For example, 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 exists a Nash equilibrium in which both teachers choose the same effort levels. Given any symmetric equilibrium, however, the ex post probability that a particular student who received extra instruction scores 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 that affect achievement during the year, ε, the sample mean of $\frac{\Delta p}{\Delta t}$ provides a useful approximation for $h(0)$.[17]

IV. Pay for Percentile

The two-person contest we describe allows education authorities to avoid the coaching and scale manipulation that scale-dependent incentive schemes invite, but 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 in this environment.

[16] In the case of no measurement error in test scores, the assumption that ε_{ij} are distributed identically and independently for all students ensures that $h(0)$ is common to all students. With measurement error in test scores and nonlinear $g(\cdot)$, however, $h(0)$ may not be the same for all comparison sets even if both ε_{ij} and u_{ij} are distributed identically and independently for all students. In this case, a variation of our scheme that involves different prizes for contests involving students with different baseline test scores can still elicit efficient effort from all teachers to all students. In Section V, we discuss a related result for the case where the distribution of shocks, ε_{ij}, differs among baseline achievement levels $i = 1, 2, \ldots, N$.

[17] Our production technology implicitly normalizes the units of ε so that shocks to achievement can be thought of as additions to or deletions from the hours of effective classroom instruction, t_j, that 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.

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. Suppose 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.[18] In this setting, teacher j's problem is

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

The first-order conditions are given by

$$(5) \quad \frac{\partial C(\mathbf{e}_j, t_j)}{\partial e_{ij}} = \sum_{k=1}^{K} \alpha h\big(\alpha(e_{ij} - e_{ik}) + t_j - t_k\big)(X_1 - X_0) \quad \text{for } i \in \{1, 2, \ldots, N\}$$

$$(6) \quad \frac{\partial C(\mathbf{e}_j, t_j)}{\partial t_j} = \sum_{k=1}^{K} \sum_{i=1}^{N} h\big(\alpha(e_{ij} - e_{ik}) + t_j - t_k\big)(X_1 - X_0).$$

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 $k \in \{1, 2, \ldots, K\}$. In this case, the right-hand sides of equations (5) and (6) 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 for each teacher are satisfied. Further, Proposition 1 extends trivially to contests against $K > 1$ 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 $K = J - 1$, so that each teacher competes against all other teachers. Further, let A_i' denote a terminal score chosen at random and uniformly from the set of all terminal scores $(a_{i1}', \ldots, a_{iJ}')$. Since the distributions of $(a_{i1}', \ldots, a_{i,j-1}', a_{i,j+1}', \ldots, a_{iJ}')$ and $(a_{i1}', \ldots, a_{i,j-1}', a_{ij}', a_{i,j+1}', \ldots, a_{iJ}')$ both converge to the same distribution 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

$$\max_{\mathbf{e}_j, t_j} NX_0 + \frac{R}{h(0)} \sum_{i=1}^{N} \Pr(a_{ij}' \geq A_i') - C(\mathbf{e}_j, t_j).$$

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

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.

Green and Stokey (1983) and Nalebuff and Stiglitz (1983) demonstrate that purely ordinal contests involving many workers can elicit efficient effort from all workers in a variety of production environments. Our contribution to this literature is to demonstrate that, when workers produce many different outputs simultaneously, the employer does not need to create a cardinal measure of total output per worker or even be able to rank total output per worker. If the principal can rank workers with respect to their contributions to each type of output, these rankings can be transformed into an efficient payment system. This result holds even though classroom instruction, t, affects the learning of all students in a given classroom, and we show in Section V that this result also holds in settings where the tutoring that one student receives affects the learning of his classmates.[19]

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. Given the separability we impose on the human capital production function in equation (2.1) and the symmetry we impose on the cost function, however, student i's comparison set need not be restricted to students with similar classmates. For any given student, we can form a comparison set 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.

Thus, given the production technology that we have assumed so far, pay for percentile can be implemented quite easily and transparently. 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 show that the existence of peer effects, instructional spillovers, and observable differences in classroom resources do not alter the efficiency properties of pay for percentile but do complicate the task of constructing comparison sets.

[19] One can imagine many settings where employers may find it more difficult to rank workers according to the value of their total output than to rank them separately with regard to the values of many different types of output they produce. Consider a consulting firm that wants J consultants to simultaneously produce N different outputs by devoting effort to T distinct tasks. The N outputs may include changes in the value of the firm's brand name, changes in the value of its client and referral lists, etc., and the number of tasks T that contribute to these outputs may be greater or less than N. Our results imply that the firm does not need to be able to rank its J consultants according to their total contribution to the firm, and it never needs to quantify any aspects of the realized performance differences among its consultants. The firm only needs to rank consultants' individual performances with respect to each of the N distinct forms of output. In this section, we have assumed that the mappings between teacher (consultant) effort and human capital (output) created are the same for all students (types of output). In Section V, we consider settings in which there are heterogenous mappings between worker effort choices and the production of different types of output, and we show that pay for percentile works in this environment as well.

V. Heterogeneous Returns, Heterogeneous Teachers, and Other Generalizations

Thus far, we have considered a simple environment in which the education authority desires the same level of classroom instruction and tutoring from each teacher for all students. In this section, we consider several generalizations. First, we consider a more general technology that incorporates student heterogeneity in gains from instruction, peer effects among students, and instructional spillovers. We show that pay for percentile can be used to elicit socially efficient effort from teachers in this setting, even though the optimal efforts allocated to individual students may vary with their baseline achievements and the distribution of baseline achievement among their peers. We then introduce heterogeneous teachers. Here, socially efficient effort allocations vary with teacher talent, holding constant the composition of students in a classroom. Given various assumptions about the information that teachers possess concerning their own talents and the talents of other teachers, we discuss strategies for implementing variations of our basic pay for percentile scheme that can elicit efficient effort from teachers of all ability levels.

A. *Heterogeneous Returns, Peer Effects, and Spillovers*

Let $\mathbf{a}_j = (a_{1j}, \ldots, a_{Nj})$ denote the initial levels of human capital of all students in teacher j's class, where $j \in \{1, 2, \ldots, J\}$. We allow the production of human capital for each student i in class j to depend 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, $\mathbf{e}_j$. In the terminology commonly employed in the educational production function literature, we allow heterogeneous rates of learning, direct peer effects, and instructional spillovers. Formally, the human capital of student i in classroom j is given by

$$a'_{ij} = g_i(\mathbf{a}_j, t_j, \mathbf{e}_j) + \varepsilon_{ij}. \tag{7}$$

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; e.g., from tutoring, lectures, interactions with classmates, etc. Nonetheless, we place three restrictions on $g_i(\cdot, \cdot, \cdot)$: the first derivatives of $g_i(\cdot, \cdot, \cdot)$ with respect to each dimension of effort are finite everywhere; $g_i(\cdot, \cdot, \cdot)$ is weakly concave; and $g_i(\cdot, \cdot, \cdot)$ depends on class identity, j, only through teacher efforts. Our concavity assumption places restrictions on the 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, k) with the same composition of baseline achievement, if the two teachers in question choose the same effort levels, i.e., $t_j = t_k$ and $\mathbf{e}_j = \mathbf{e}_k$, the expected human capital levels for any two students in different classrooms who share the same initial achievement are the same. Given this property and the fact that the cost of effort is 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} can vary across students with different baseline

achievement levels. In Section I, 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} are now interpreted as additions to or deletions from the stock of student human capital.

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 with respect to levels of individual tutoring effort, 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

$$\max_{\mathbf{e}_j, t_j} E\left[R\sum_{i=1}^{N}\left[g_i(\mathbf{a}_j, t_j, \mathbf{e}_j) + \varepsilon_{ij}\right] - C(\mathbf{e}_j, t_j)\right]. \tag{8}$$

Since $g_i(\cdot,\cdot,\cdot)$ is 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

$$\frac{\partial C(\mathbf{e}_j, t_j)}{\partial e_{ij}} = R\sum_{m=1}^{N}\frac{\partial g_m(\mathbf{a}_j, t_j, \mathbf{e}_j)}{\partial e_{ij}} \qquad \text{for } i \in \{1, 2, \ldots, N\}$$

$$\frac{\partial C(\mathbf{e}_j, t_j)}{\partial t_j} = R\sum_{m=1}^{N}\frac{\partial g_m(\mathbf{a}_j, t_j, \mathbf{e}_j)}{\partial t_j}.$$

For any composition of baseline achievement, there will be a unique $(\mathbf{e}_j^*, t_j^*)$ that solves these equations. This vector will differ for classes with different compositions, $\mathbf{a}_j$, however, and the tutoring effort, e_{ij}, for each student may differ across students in the same class.

We now argue that the same 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. In contrast to the case above where the human capital production function was separable and returns to instruction were identical for all students, it is now essential that all classrooms in this comparison set share a common distribution of baseline achievement.

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

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

The first-order conditions for teacher j are

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

$$\forall i \in \{1, 2, \ldots, N\}$$

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

If all teachers provide the same effort levels, and $X_1 - X_0 = \frac{R}{Kh(0)}$, these first-order conditions collapse to the planner's first-order conditions, and the proof of Proposition 1 in online Appendix B establishes that there exists a Nash equilibrium such that all teachers choose the first best effort levels in response. Efficiency requires that all teachers earn bonus pay at the same rate regardless of the baseline achievement of their students, $\mathbf{a}_j$. The level of base pay required to satisfy the teachers' participation constraints will be a function of the specific distribution of baseline achievement in her classroom, however, since the socially efficient effort levels induced by pay for percentile vary with classroom composition.

Even with production technologies that depend on the level of baseline achievement, $g_i(\cdot, \cdot, \cdot)$, the optimal prize structure does not vary with baseline achievement because we have assumed that the distribution of ε_{ij}, and thus $h(0)$, does not vary among students. It is not possible to test this assumption without placing restrictions on how the production function, $g_i(\cdot, \cdot, \cdot)$, may vary with achievement level i.[20] Further, even if one assumes that $h(0)$ is the same for all pairwise student contests, the process of discovering $h(0)$ given the more general technology $g_i(\cdot, \cdot, \cdot)$ is more involved than the process described in Section III. Here, R is no longer the value of instruction time provided to any given student; R must now be interpreted as the value of the skills created when one unit of instruction time is devoted to a particular type of student in a specific type of classroom. Thus, attempts to learn $h(0)$ based on experiments like those described in Section III must be restricted to a specific set of students who all have the same baseline achievement level, the same peers, and some known baseline level of instruction and tutoring. Finally, if $h(0)$ is not the same for given baseline achievement levels, the authority can still elicit efficient effort using a pay for percentile scheme if it knows how $h_i(0)$ varies among baseline achievement levels. In this setting the authority can elicit efficient effort by offering prizes, $\frac{R}{Kh_i(0)}$, that are specific to each of the levels of baseline achievement that define types of student contests.

[20] In the experiment we described at the end of Section III, students are given a small amount of extra instruction, and the experiment identifies $h(0)(\partial a'_{ij}/\partial t_j)$, which equals $h(0)$ given the linear technology assumed in Section I. If we ran separate experiments of this type for each baseline achievement level, we could not use variation in the product $h_i(0)(\partial a'_{ij}/\partial t_j)$ to test the hypothesis $h_i(0) = h(0)$ for all i without knowing how $\partial a'_{ij}/\partial t_j$ varies with i.

B. *Heterogeneous Teachers*

To this point, we have analyzed contests among equally talented teachers. We now assume that teachers differ in their abilities to produce human capital and explore the properties of pay for percentile given several different assumptions concerning the dispersion of information about these productivity differences.

Consider the following description of teacher heterogeneity: if a student with baseline achievement a_i is paired with teacher j, the student's human capital at the end of year is

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

where λ_j is a teacher-specific productivity term drawn from a distribution with mean 1 and support $\left[\underline{\lambda}, \overline{\lambda}\right]$ with $0 < \underline{\lambda} < \overline{\lambda} < \infty$. Here, the expected skill level of students is increasing in the talent of their teachers, holding constant the time that teachers devote to classroom instruction and the tutoring of individual students. Thus, socially efficient levels of classroom instruction and tutoring are increasing functions of teacher productivity, λ_j.

Given this setup, consider the case where the education authority observes the talent of each teacher. In this case, the authority can elicit efficient effort from all teachers by simply requiring each teacher to compete only against other teachers who not only work with classrooms of comparable students, $\mathbf{a}_j = \mathbf{a}_k$, but who also enjoy the same level of talent, $\lambda_j = \lambda_k$. Note that, even though the efficient level of effort varies with teacher talent, the authority can induce all types to provide efficient effort using a common prize rate, $\frac{R}{h(0)}$, as long as the authority seeds contests using information on both student characteristics and teacher types. This observation indicates that pay for percentile should, in practice, require teachers to compete against other teachers who not only work with similar students but who also share similar levels of experience and comparable levels of support in the form of teachers' aides, classroom computers, and other resources.

Even if education authorities sort teachers into leagues based on observed correlates of teacher productivity, however, residual differences in productivity will likely remain. Given these differences, what teachers know about their own ability levels relative to others in their leagues is the key factor that determines how the authority should amend pay for percentile to maintain efficiency. If all teachers are uncertain about how they compare to their competitors but share symmetric beliefs about their relative productivity levels, then in any symmetric equilibrium of our pay for percentile game, the teachers' uncertainty about their own talents and the talents of other teachers simply enters our model as a symmetric, contest-specific component of the shock terms that determine the outcomes of the contests we describe.

Consider a classroom type defined by a specific vector of baseline achievement levels $\mathbf{a}$, and assume that teacher j competes against teacher k with $\mathbf{a}_j = \mathbf{a}_k = \mathbf{a}$. If both choose effort levels $(\mathbf{e}^*, t^*)$, the winner of the contest i, for all $i = 1, 2, \ldots, N$, is determined by the sign of the realization of the random variable

$$u_i = (\lambda_j - \lambda_k) g_i(\mathbf{a}, t^*, \mathbf{e}^*) + (\varepsilon_{ij} - \varepsilon_{ik}).$$

Although u_i depends on the effort choices of both teachers, it has a symmetric distribution whose density peaks at zero when both teachers choose the same effort levels, and a pay for percentile scheme still exists that elicits efficient effort from all teachers. The optimal bonus rate, however, is now greater than $\frac{R}{h(0)}$. Here, unknown differences in teacher talent are an additional source of uncertainty in contest outcomes, and if one teacher deviates from the equilibrium effort levels, the associated rate of change in the probability of winning a particular contest is now less than $h(0)$. Still, pay for percentile can be used to elicit efficient effort from heterogeneous teachers as long as the authority observes teacher characteristics such that, conditional on these characteristics, teachers share a common prior about their talent and the talents of their competitors.

If instead individual teachers possess private, asymmetric information about their own talents relative to other teachers in their league, then pay for percentile with a uniform bonus rate for all leagues will not generally elicit efficient effort from all teachers. In this case, the authority needs to learn the hidden type of each teacher and then elicit a different efficient level of effort from each type. Our simple mechanism alone cannot do both jobs at once. We noted above that a common pay for percentile prize rate induces efficient effort from all ability types when the authority observes differences in teacher ability and can thus force teachers to sort into homogeneous leagues. It is obvious, however, that when the authority chooses a common prize rate for all leagues, it is providing no incentives for teachers to sort among leagues.[21]

Intuitively, one should be able to use differences in bonus rates across leagues to induce more talented teachers to sort into leagues with higher prize spreads, but the education authority must simultaneously employ an additional instrument that restores efficient incentive provision within leagues. The existing literature on asymmetric tournaments with private information contains two mechanisms that follow this approach.

O'Keeffe, Viscusi, and Zeckhauser (1984) analyze an environment involving workers who are either of high or low ability. They propose a mechanism that involves two leagues that vary in base pay, prize levels, and the extent to which chance affects the outcomes of contests given symmetric effort choices by contestants. In the context of our pay for percentile scheme, their results suggest that by simultaneously reducing the noise in the process that determines contest outcomes, raising base pay and lowering bonus rates, education officials can design a league that attracts only the low-ability types, and these types will choose efficient effort levels while competing against other teachers of low ability. The high-ability types select the league that involves lower base pay, higher prizes, and a larger role of chance as a determinant of contest outcomes, and they also choose the efficient level of effort for their talent level.

Bhattacharya and Guasch (1988) also propose a scheme where workers select the compensation rules they face, but all contestants compete only against the agents who choose the compensation scheme designed to attract agents of the lowest-ability type. Here, leagues that offer larger prizes for winning contests also offer lower levels of base pay, and Bhattacharya and Guasch establish conditions under

[21] Lazear and Rosen (1981) made this observation in their seminal paper on economic contests.

which their menu scheme induces workers to sort into the contract designed for their ability types and supply efficient effort levels given their types. This scheme works because low base pay serves as an entry fee into the high-prize contests that low-ability workers are not willing to pay. One can imagine a similar variation on pay for percentile in which all teachers choose a combination of base pay and bonus rate from a menu, and then compete ex post against only those teachers in their comparison sets who choose the lowest bonus rate.

Both of these schemes require more information to implement than the simple pay for percentile scheme we describe above, and they also require additional restrictions on the model environment. In order to implement the Bhattacharya and Guasch scheme, the authority must know the education production function and the entire distribution that generates shocks to student achievement. Further, the Bhattacharya and Guasch results only hold given joint restrictions on the shape of this shock distribution and the range of differences in teacher talent. Finally, because many of the shocks that affect student learning are beyond the control of teachers and education authorities, it may not be possible to minimize the role of chance in determining contest outcomes by an amount sufficient to create the self-selection patterns that O'Keeffe, Viscusi, and Zeckhauser (1984) describe. As we noted earlier, a certain amount of randomness in contest outcomes is required to sustain Nash equilibria in these games.

Even if these menu approaches are infeasible, however, the authority may still be able to improve the effort allocations of heterogeneous teachers who participate in a pay for percentile scheme with a common reward structure. In the case where teachers possess private information about their own talent levels, we can show that teachers of all ability types respond to our pay for percentile scheme by supplying less than the efficient effort levels for their types. Thus, if the equilibrium effort choices of all types are increasing in the bonus rate, the authority can move all types closer to efficient effort levels simply by raising the bonus rate above $\frac{R}{h(0)}$.[22]

So far, we have discussed private information that teachers possess ex ante. Since our model is static, we cannot address how private information revealed to teachers about their students' progress during the year may affect their effort over time within a school year. It is possible that teacher j may not allocate efficient effort to some students during the spring if portions of ε_{ij} are revealed to teacher j during the fall. Still, ε_{ij} captures shocks to total achievement growth over the entire year, conditional on the baseline human capital of student i and the experience of teacher j, and it is not clear that private signals early in the school year are important predictors of how particular students will rank at the end of the year within their comparison sets given particular levels of teacher effort. In addition, in order to make precise statements concerning how one might modify our scheme to address the presence of such private information, we would need a dynamic model of information revelation during the school year, which is beyond the scope of this paper.

Finally, we have also ignored the dynamic issues that arise from the fact that teachers will teach many different classrooms of students over their careers. Access to repeated measures of teacher performance may provide education authorities with additional instruments that facilitate incentive provision. In a single-period setting,

[22] Harbring and Lunser (2008) describe one particular asymmetric contest model in which the effort of all types is an increasing function of prize levels. Further, they provide experimental evidence that supports this prediction.

when heterogeneous teachers compete against each other in our pay for percentile scheme, less-able teachers typically lose a disproportionate share of their contests. If this result remains in a dynamic setting, the contests we describe would be a vehicle for learning about teacher talent over time. As teachers accumulate histories of percentile performance indices that describe their performance over their careers, education authorities may be able to usefully link retention decisions, tenure decisions, decisions about changes in base pay levels, and promises concerning future pension benefits to these performance histories. We plan to explore these issues in future research.

VI. Implementation Issues

As we stress in the previous section, pay for percentile requires that contests among teachers are properly seeded. In practice, it may be impossible to form large comparison sets containing classrooms with common measured resources, equally experienced teachers, and identical distributions of baseline student achievement. Thus, education authorities who implement pay for percentile must adopt some method that employs information on the performance of similar students and teachers to estimate the percentile scores that would have been realized in comparison sets defined by matching on identical students and teachers. Recent work in education statistics by Wei and He (2006) and Betebenner (2009) has produced estimation methods and software that allow education officials to estimate, for given sets of baseline conditioning variables, sets of predicted scores that describe the percentiles in the corresponding conditional distributions of scores.[23] Given a predicted score distribution for each individual student that conditions on the characteristics of each student's teacher and peers, education authorities can assign a conditional percentile score to each student. Given these estimated percentiles, one can easily estimate percentile performance indices in systems with large student populations.[24]

Colorado, Indiana, and Massachusetts currently estimate conditional percentile scores for many of their students. These conditional percentile scores are also known as "student growth percentiles," and they serve as key inputs into new metrics of school performance that are being used for accountability purposes under NCLB.[25] Briggs and Betebenner (2009) and others in the education statistics research community who advocate scale-invariant performance metrics are not concerned with the optimal design of incentive contracts. Instead, they are interested in constructing ex post measures of educator performance that are robust to decisions about how assessments results are scaled. We focus on ordinal statistics because we want to avoid the coaching and scale manipulation behaviors that scale-dependent systems invite. These differences in motivation notwithstanding, the Betebenner (2008)

[23] The Betebenner software employs Q separate quantile regression models to create Q predicted quantiles in the conditional distribution of scores for each student. The software then uses interpolation methods to create a predicted conditional distribution of scores for each student. In exploratory work with data from the Chicago Public Schools, Neal has found that this method produces credible results for choices of Q around 30.

[24] See https://sites.google.com/site/dereknealresearch/home/pay-for-percentile-software for free software that uses a variation of Betebenner's method to calculate PPI measures.

[25] According to Damian Betebenner, who developed the software package commonly used to estimate student growth percentiles, more than 15 other state education authorities are currently considering how they can implement estimation of student growth percentiles for students in their states. The software is available at http://cran.r-project.org/web/packages/SGP/index.html.

Table 1—Recent Pay for Performance Systems in Education

Name	Place	Description of performance metric
ABC	NC	State sets achievement targets for each school based on a model of school achievement growth.
MAP	FL	Districts choose their own method for measuring teacher contribution to achievement.
POINT	TN	VAM analyses on historical achievement data yield value-added performance targets.
PRP	England	Teachers submit applications for bonus pay and provide documentation of better than average performance in promoting student achievement.
ProComp	Denver	Teachers and principals negotiate achievement targets for individual students.
TAP	14 states	Relative performance metrics come from a value-added model (VAM).
QComp	MN	Schools develop their own plans for measuring teacher contributions to students' achievement.

Notes: Each system employs additional measures of teacher performance that are not directly tied to student assessment results. Here, we describe performance statistics derived from test scores.

software contains the basic tools required to estimate our percentile performance indices. Because this approach involves estimating entire distributions of final test scores given fixed sets of baseline characteristics, the practical implementation of pay for percentile should work best in school systems that involve large student populations; e.g., entire states or very large, unified districts.

We are highlighting the potential benefits of competition based on ranks as the basis for incentive pay systems in education, 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.[26]

VII. Lessons for Policymakers

In Sections I through IV, we analyzed incentive provision for teachers and highlighted the benefits of the relative performance incentive schemes as well as the additional benefits of seeded contests determined by ordinal ranks. Here, we employ the same framework we used to derive these results to discuss the advantages and deficiencies of teacher incentive schemes that are already in place. Table 1 lists a number of recent pay for performance systems employed in public schools.[27] Each attaches bonus pay for teachers to some assessment-based metric of classroom or school performance. The table provides a brief description of the performance metrics used in these systems. Here, we discuss these metrics in light of the lessons learned from our model.

[26] 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.

[27] Here, we restrict attention to performance pay schemes that are either ongoing policies or were in place recently. See Ladd (1999); Lavy (2002, 2009) for evaluations of experimental incentive pay programs. Ladd employs data from Dallas, Texas, and Lavy uses data from Israel.

Beginning in Section II, our analyses highlight the value of peer comparisons as a means of revealing what efficient achievement targets should be. With the exception of Teacher Advancement Program (TAP) and the possible exception of Performance-Related-Pay (PRP),[28] all the systems described in Table 1 involve setting achievement targets for individual students or collections of students and rewarding teachers for meeting these targets. The targets and rewards in these systems can conceivably be chosen so that teachers respond to such systems by choosing efficient effort. The education authority cannot choose these targets correctly, however, without knowing the educational production function, $g_i(\cdot,\cdot,\cdot)$, and the scaling of assessments, $m(\cdot)$. Because the education authority may not observe the production function directly, teachers may seek to corrupt the process that the authority uses to determine achievement targets. Further, the scales used by testing agencies to report results to the education authority may also be vulnerable to manipulation.

In contrast, a scheme that pays for performance relative to appropriately chosen peers rather than for performance relative to specific thresholds is less vulnerable to corruption and manipulation. By benchmarking performance relative to peer performance, the education authority avoids the need to forecast $E[m(g_i(\mathbf{a}_j, \mathbf{e}_j^*, t_j^*) + \varepsilon_{ij})]$; i.e., the expected end-of-year score for student i given efficient teacher effort, which it needs in order to set appropriate thresholds. Moreover, relative performance schemes do not create opportunities for influence activities that either lower performance thresholds or inflate the scales used to report assessment results and thereby transform performance pay schemes into programs that primarily increase effective base pay for all teachers. Total reward pay is fixed in relative performance systems, and performance thresholds are endogenously determined through peer comparisons.[29]

Among the entries in Table 1, the Value-Added Model (VAM) approach contained in the TAP is the only scheme based on an objective mapping between student test scores and relative performance measures for teachers.[30] Our percentile performance indices provide summary measures of how often the students in a given classroom perform better than comparable students in other schools, while VAM models measure the distance between the average achievement of students in a given classroom and the average achievement one would expect from these students if they were randomly assigned to different classrooms.

Both schemes produce relative performance indices for teachers, but the VAM approach is more ambitious. Because VAM treats the units of a given psychometric

[28] PRP instructed teachers to document that their students were making progress "as good or better" than their peers. PRP involved no system for producing objective relative performance measures, however, and ex post, the high rate of bonus awards raised questions about the leniency of education officials in setting standards for "as good or better" performance. See Atkinson et al. (2009) and Wragg et al. (2001).

[29] A previous literature examined other benefits of performance pay schemes that involve competition for a fixed amount of reward funds. Malcomson (1984) explains that relative performance contracts are enforceable in circumstances where standards-based performance pay is not. Even in settings where standards are not verifiable, the delivery of bonus payments is verifiable, and firms have an incentive ex post to make these payments to workers who performed best in a relative sense. Carmichael (1983) also demonstrates that relative performance schemes can be used as devices that allow firms to credibly commit to provide efficient levels of complementary inputs for their workers.

[30] The Project on Incentives in Teaching (POINT) plan involves value-added targets that are derived from VAM analyses of achievement distributions from previous years, and thus POINT pays for performance relative to past peer performance. POINT does not involve current competition for a fixed amount of bonus funds, however. While the targets set for this experimental programming are quite demanding, it is possible that such demanding standards would be compromised over time in an ongoing program.

scale as a welfare index, VAM indices provide a universal ranking of classrooms according to performance and also provide measures of the sizes of performance gaps between classrooms.[31] In contrast, our percentile performance indices do not permit such comparisons. If teacher A has a higher percentile performance index than teacher B and the two teachers work with comparable students, we know that teacher A's students performed better than their peers more often than teacher B's students, but we do not know whether or not the total value of human capital created in teacher A's classroom actually exceeds the total value created in teacher B's classroom. Further, if teacher A and teacher B work with students who are very different in terms of their baseline characteristics, their percentile performance indices provide no information about the relative performance of these teachers because their students do not belong to common comparison sets. As we note in Section V, a key lesson from our approach is that education authorities can provide effective incentives for educators without ever forming cardinal measures of total teacher performance at the classroom level or even an ordinal performance ranking over all teachers.

Cardinal measures of relative performance are a required component of the TAP approach and related approaches to performance pay for teachers because these systems attempt to first measure the contributions of educators to achievement growth and then reward teachers according to these contributions. Campbell (1976) famously claimed that government performance statistics are always corrupted when high stakes are attached to them, and the contrast between our percentile performance indices and VAM indices suggests that Campbell's Law[32] is a warning about the perils of trying to accomplish two objectives with one performance measure.

Systems like TAP that try to both provide incentives for teachers and produce cardinal measures of educational productivity are likely to do neither well because, as we have noted several times, 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. Further, because it is reasonable to expect heterogeneity in the extent of these coaching activities across classrooms, there is no guarantee that scale-dependent measures of relative performance even provide reliable ex post rankings of true performance over classrooms.

Scale-invariant schemes like pay for percentile provide no information about secular changes in student achievement levels over time. If education authorities desire measures of secular changes in achievement or achievement growth 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 taking these scaled tests. Because educators face no direct incentives to manipulate the results of this

[31] Some VAM practitioners employ functional form assumptions that allow them to produce universal rankings of teacher performance and make judgments 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 teaching honors classes and provides a context-specific measure of how well teachers are doing the job they actually have.

[32] Campbell (1976, p. 49) concluded a review of studies concerning performance in government agencies this way: "I come to the following pessimistic laws (at least for the US 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."

second assessment system, it will produce more reliable information about trends in student achievement.

VIII. Conclusion

The current paradigm in education policy focuses on ways to build accountability and incentive schemes for public educators that produce scaled measures of student performance and educator performance while providing performance incentives for educators. Our results suggest that education authorities can improve performance measurement and incentive provision by addressing these goals separately.

We show that ordinal comparisons between each of a teacher's students and the student's peers provide the information required for efficient performance pay. Each of these comparisons determines the outcome of a symmetric contest, and a policy that pays bonuses proportional to the fraction of total contests won can elicit efficient effort from teachers. The fact that education authorities can implement an effective scale-invariant incentive scheme is important for two reasons. First, if education officials do not need to place results for different assessments on a common scale, they are free to employ assessments without repeated items and common formats, and much research demonstrates that, while repeated items and common formats make scale integrity possible in theory, these features also invite the coaching behaviors that undermine scale integrity in practice. Second, scale integrity is difficult to verify and therefore an obvious target for political manipulation.

Education authorities who desire consistently scaled measures of student or teacher performance should create these measures based on the results of a separate assessment series that involves no stakes for teachers or principals. Measurement systems that involve no stakes for educators provide no incentives for educators to take actions that contaminate the resulting metrics.

REFERENCES

Atkinson, Adele, Simon Burgess, Bronwyn Croxson, Paul Gregg, Carol Propper, Helen Slater, and Deborah Wilson. 2009. "Evaluating the Impact of Performance-Related Pay for Teachers in England." *Labour Economics* 16 (3): 251–61.

Baker, George. 1992. "Incentives, Contracts, and Performance Measurement." *Journal of Political Economy* 100 (3): 598–614.

Ballou, Dale. 2009. "Test Scaling and Value-Added Measurement." *Education Finance and Policy* 4 (4): 351–83.

Betebenner, Damian W. 2009. "Norm and Criterion-Referenced Student Growth." *Educational Measurement: Issues and Practice* 28 (4): 42–51.

Bhattacharya, Sudipto, and J. Luis Guasch. 1988. "Heterogeneity, Tournaments, and Hierarchies." *Journal of Political Economy* 96 (4): 867–81.

Briggs, Derek, and Damian Betebenner. 2009. "Is Growth in Student Achievement Scale-Dependent?" Paper presented at the annual meetings of the National Council on Measurement in Education, San Diego, CA.

Campbell, Donald T. 1976. "Assessing the Impact of Planned Social Change." Dartmouth College Public Affairs Center Occasional Working Paper 8.

Carmichael, H. Lorne. 1983. "The Agent-Agents Problem: Payment by Relative Output." *Journal of Labor Economics* 1 (1): 50–65.

Carrell, Scott E., and James E. West. 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *Journal of Political Economics* 118 (3): 409–32.

Cawley, John, James Heckman, and Edward Vytlacil. 1999. "On Policies to Reward the Value Added of Educators." *Review of Economics and Statistics* 81 (4): 720–28.

Cronin, John, Michael Dahlin, Deborath Adkins, and G. Gage Kingsbury. 2007. *The Proficiency Illusion*. Washington, DC: Thomas B. Fordham Institute.

Cunha, Flavio, and James Heckman. 2008. "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Journal of Human Resources* 43 (4): 738–82.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. "Teacher Incentives." *American Economic Journal: Applied Economics* 2 (3): 205–27.

Green, Jerry R., and Nancy L. Stokey. 1983. "A Comparison of Tournaments and Contracts." *Journal of Political Economy* 91 (3): 349–64.

Harbring, Christine, and Gabriele K. Lunser. 2008. "On the Competition of Asymmetric Agents." *German Economic Review* 9 (3): 373–95.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership and Job Design." *Journal of Law, Economics and Organization* 7 (1): 24–52.

Jacob, Brian. 2005. "Accountability Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics* 89 (5–6): 761–96.

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

Koretz, Daniel M. 2002. "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity." *Journal of Human Resources* 37 (4): 752–77.

Koretz, D. M., and S. I. Barron. 1998. *The Validity of Gains on the Kentucky Instructional Results Information System.* (*KIRIS*). Washington, DC: RAND.

Ladd, Helen F. 1999. "The Dallas School Accountability and Incentive Program: an Evaluation of its Impact on Student Outcomes." *Economics of Education Review* 18 (1): 1–16.

Lavy, Victor. 2002. "Evaluating the Effect of Teacher's Group Performance Incentives on Pupil Achievement." *Journal of Political Economy* 110 (6): 1286–1317.

Lavy, Victor. 2009. "Performance Pay and Teacher's Effort, Productivity, and Grading Ethics." *American Economic Review* 99 (5): 1979–2011.

Lazear, Edward. 2001. "Educational Production." *Quarterly Journal of Economics* 116 (3): 777–803.

Lazear, Edward, and Sherwin Rosen. 1981. "Rank Order Tournaments as Optimum Labor Contracts." *Journal of Political Economy* 89 (5): 841–64.

Malcomson, James M. 1984. "Work Incentives, Hierarchy and Internal Labor Markets." *Journal of Polical Economy* 92 (3): 486–507.

Nalebuff, Barry J., and Joseph E. Stiglitz. 1983. "Prizes and Incentives: Towards a General Theory of Compensation and Competition." *Bell Journal of Economics* 14 (1): 21–43.

Neal, Derek. 2010. "Aiming for Efficiency Rather than Proficiency." *Journal of Economic Perspectives* 24 (3): 119–32.

Neal, Derek. 2012. "The Design of Performance Pay in Education." In *Handbook of Economics of Education. Vol. 4*, edited by Eric Hanushek, Steve Machin, and Ludger Woessmann, 495–550. Amsterdam: Elsevier.

O'Keeffe Mary, Kip W. Viscusi, and Richard J. Zeckhauser. 1984. "Economic Contests: Comparative Reward Schemes." *Journal of Labor Economics* 2 (1): 27–56.

Peterson, Paul, and Frederick Hess. 2006. "Keeping an Eye on State Standards: A Race to the Bottom." *Education Next* 6 (3): 28–29.

Stecher, Brian. 2002. "Consequences of Large-Scale, High-Stakes Testing on School and Classroom Practice." In *Making Sense of Test-Based Accountability in Education*, edited by Laura S. Hamilton, Brian Stecher, and Stephen P. Klein, 79–100. Washington, DC: RAND.

Vigdor, Jacob L. 2009. "Teacher Salary Bonuses in North Carolina." In *Performance Incentives: Their Growing Impact on American K–12 Education*, edited by Matthew G. Springer, 227–50. Washington, DC: Brookings Institution.

Vogt, H. 1983. "Unimodality in Differences." *Metrika* 30 (1): 165–70.

Wei, Ying, and Xuming He. 2006. "Conditional Growth Charts." *Annals of Statistics* 34 (5): 2069–97.

Wragg, E., G. Haynes, C. Wragg, R. Chamberlin. 2001. "Performance Related Pay: The Views and Experiences of 1,000 Primary and Secondary Headteachers." University of Exeter School of Education Teachers' Incentives Pay Project Occasional Paper 1.

This article has been cited by:

1. Mark A. Paige, Audrey Amrein-Beardsley. 2020. "Houston, We Have a Lawsuit": A Cautionary Tale for the Implementation of Value-Added Models for High-Stakes Employment Decisions. *Educational Researcher* **49**:5, 350-359. [Crossref]
2. Richard Murphy, Felix Weinhardt. 2020. Top of the Class: The Importance of Ordinal Rank. *The Review of Economic Studies* **25**. . [Crossref]
3. Fang Chang, Huan Wang, Yaqiong Qu, Qiang Zheng, Prashant Loyalka, Sean Sylvia, Yaojiang Shi, Sarah-Eve Dill, Scott Rozelle. 2020. The impact of pay-for-percentile incentive on low-achieving students in rural China. *Economics of Education Review* **75**, 101954. [Crossref]
4. Andy Brownback, Sally Sadoff. 2019. Improving College Instruction Through Incentives. *Journal of Political Economy* . [Crossref]
5. Muharrem Yeşilırmak. 2019. Bonus pay for teachers, spatial sorting, and student achievement. *European Journal of Political Economy* **59**, 129-158. [Crossref]
6. Prashant Loyalka, Sean Sylvia, Chengfang Liu, James Chu, Yaojiang Shi. 2019. Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievement. *Journal of Labor Economics* **37**:3, 621-662. [Crossref]
7. Saturnin Dandala. 2019. Human resource policy and teacher appraisal in Ontario in the era of professional accountability. *Management in Education* **33**:1, 5-10. [Crossref]
8. Nirav Mehta. 2019. Measuring quality for use in incentive schemes: The case of "shrinkage" estimators. *Quantitative Economics* **10**:4, 1537-1577. [Crossref]
9. Nirav Mehta. 2018. The potential output gains from using optimal teacher incentives: An illustrative calibration of a hidden action model. *Economics of Education Review* **66**, 67-72. [Crossref]
10. Stephani L. Wrabel, Andrew Saultz, Morgan S. Polikoff, Andrew McEachin, Matthew Duque. 2018. The Politics of Elementary and Secondary Education Act Waivers. *Educational Policy* **32**:1, 117-140. [Crossref]
11. Andy Brownback, Sally Sadoff. 2018. Improving College Instruction Through Incentives. *SSRN Electronic Journal* . [Crossref]
12. Jeffrey Penney. 2017. Test Score Measurement and the Black-White Test Score Gap. *The Review of Economics and Statistics* **99**:4, 652-656. [Crossref]
13. Ben Ost, Anuj Gangopadhyaya, Jeffrey C. Schiman. 2017. Comparing standard deviation effects across contexts. *Education Economics* **25**:3, 251-265. [Crossref]
14. Weerachart T. Kilenthong, Gabriel A. Madeira. 2017. Observability and endogenous organizations. *Economic Theory* **63**:3, 587-619. [Crossref]
15. Naureen Karachiwalla, Albert Park. 2017. Promotion incentives in the public sector: Evidence from Chinese schools. *Journal of Public Economics* **146**, 109-128. [Crossref]
16. R.G. Fryer. The Production of Human Capital in Developed Countries 95-322. [Crossref]
17. K. Muralidharan. Field Experiments in Education in Developing Countries 323-385. [Crossref]
18. Isaac M. Mbiti. 2016. The Need for Accountability in Education in Developing Countries. *Journal of Economic Perspectives* **30**:3, 109-132. [Abstract] [View PDF article] [PDF with links]
19. Brian Jacob, Jesse Rothstein. 2016. The Measurement of Student Ability in Modern Assessment Systems. *Journal of Economic Perspectives* **30**:3, 85-108. [Citation] [View PDF article] [PDF with links]
20. Mark Ehlert, Cory Koedel, Eric Parsons, Michael Podgursky. 2016. Selecting Growth Measures for Use in School Evaluation Systems. *Educational Policy* **30**:3, 465-500. [Crossref]

21. Erwin Ooghe, Erik Schokkaert. 2016. School accountability: can we reward schools and avoid pupil selection?. *Social Choice and Welfare* **46**:2, 359-387. [Crossref]

22. P. Glewwe, K. Muralidharan. Improving Education Outcomes in Developing Countries 653-743. [Crossref]

23. Cory Koedel, Jiaxi Li. 2016. THE EFFICIENCY IMPLICATIONS OF USING PROPORTIONAL EVALUATIONS TO SHAPE THE TEACHING WORKFORCE. *Contemporary Economic Policy* **34**:1, 47-62. [Crossref]

24. Sarah R. Cohodes. 2016. Teaching to the Student: Charter School Effectiveness in Spite of Perverse Incentives. *Education Finance and Policy* **11**:1, 1-42. [Crossref]

25. Prashant Kumar Loyalka, Sean Sylvia, Chengfang Liu, James Chu, Yaojiang Shi. 2016. Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievement. *SSRN Electronic Journal* . [Crossref]

26. Cory Koedel, Kata Mihaly, Jonah E. Rockoff. 2015. Value-added modeling: A review. *Economics of Education Review* **47**, 180-195. [Crossref]

27. Scott A. Imberman, Michael F. Lovenheim. 2015. Incentive Strength and Teacher Productivity: Evidence from a Group-Based Teacher Incentive Pay System. *Review of Economics and Statistics* **97**:2, 364-386. [Crossref]

28. Daniel Koretz. 2015. Adapting Educational Measurement to the Demands of Test-Based Accountability. *Measurement: Interdisciplinary Research and Perspectives* **13**:1, 1-25. [Crossref]

29. David Thissen. 2015. Failing Tests: Commentary on "Adapting Educational Measurement to the Demands of Test-Based Accountability". *Measurement: Interdisciplinary Research and Perspectives* **13**:1, 49-52. [Crossref]

30. Naureen Karachiwalla, Albert Park. 2015. Promotion Incentives in the Public Sector: Evidence from Chinese Schools. *SSRN Electronic Journal* . [Crossref]

31. Mark Ehlert, Cory Koedel, Eric Parsons, Michael J. Podgursky. 2014. The Sensitivity of Value-Added Estimates to Specification Adjustments: Evidence From School- and Teacher-Level Models in Missouri. *Statistics and Public Policy* **1**:1, 19-27. [Crossref]

32. Dan Goldhaber, Joe Walch, Brian Gabele. 2014. Does the Model Matter? Exploring the Relationship Between Different Student Achievement-Based Teacher Assessments. *Statistics and Public Policy* **1**:1, 28-39. [Crossref]

33. Raj Chetty, John N. Friedman, Jonah E. Rockoff. 2014. Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review* **104**:9, 2593-2632. [Abstract] [View PDF article] [PDF with links]

34. Raj Chetty, John N. Friedman, Jonah E. Rockoff. 2014. Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review* **104**:9, 2633-2679. [Abstract] [View PDF article] [PDF with links]

35. Yongzheng Liu, Yongsheng Xu. 2014. 'Pay for Percentile': Rawlsian vs. Utilitarian. *SSRN Electronic Journal* . [Crossref]

36. Morgan S. Polikoff, Andrew J. McEachin, Stephani L. Wrabel, Matthew Duque. 2014. The Waive of the Future? School Accountability in the Waiver Era. *Educational Researcher* **43**:1, 45-54. [Crossref]

37. Derek Neal. 2013. The Consequences of Using one Assessment System to Pursue two Objectives. *The Journal of Economic Education* **44**:4, 339-352. [Crossref]