NATIONAL CENTER ON
Performance Incentives

From Data to Bonuses:

A Case Study of the Issues Related to Awarding Teachers Pay on the Basis of Their Students' Progress

Daniel F. McCaffrey
Bing Han
J.R. Lockwood

Prepared for *Performance Incentives: Their Growing Impact on American K-12 Education* in Nashville,Tennessee on February 29, 2008

Working Paper 2008-14
February 2008

LED BY
VANDERBILT
PEABODY COLLEGE

IN COOPERATION WITH:
RAND
CORPORATION
Mizzou
University of Missouri - Columbia

The National Center on Performance Incentives (NCPI) is charged by the federal government with exercising leadership on performance incentives in education. Established in 2006 through a major research and development grant from the United States Department of Education's Institute of Education Sciences (IES), NCPI conducts scientific, comprehensive, and independent studies on the individual and institutional effects of performance incentives in education. A signature activity of the center is the conduct of two randomized field trials offering student achievement-related bonuses to teachers. The Center is committed to air and rigorous research in an effort to provide the field of education with reliable knowledge to guide policy and practice.

The Center is housed in the Learning Sciences Institute on the campus of Vanderbilt University's Peabody College. The Center's management under the Learning Sciences Institute, along with the National Center on School Choice, makes Vanderbilt the only higher education institution to house two federal research and development centers supported by the Institute of Education Services.

This working paper was supported by the National Center on Performance Incentives, which is funded by the United States Department of Education's Institute of Education Sciences (R30SA06034).
This is a draft version of a paper that will be presented at a national conference, *Performance Incentives; Their Growing Impact on American K-12 Education, in Nashville,* Tennessee on February 28-29, 2008. This paper has not been formally reviewed and should not be cited, quoted, reproduced, or retransmitted without RAND's permission. The views expressed in this paper do not necessarily reflect those of sponsoring agencies or individuals acknowledged. Any errors remainthe sole responsibility of the author.

Please visit **www.performanceincentives.org** to learn more about our program of research and recent publications.

From Data to Bonuses: A Case Study of the Issues Related to Awarding Teachers Pay on the Basis of Their Students' Progress

DANIEL F. MCCAFFREY
BING HAN
J.R. LOCKWOOD

1. INTRODUCTION

There is growing recognition within the education policy community of the need to reform how teachers are paid in order to improve the quality and performance of the teaching workforce (Committee for Economic Development, 2004; Hassel, 2002; Malanga, 2001; Odden & Kelley, 1996; Odden, Kelley, Heneman & Milanowski, 2001; Southern Regional Education Board, 2000). Traditionally, teachers have been paid using a fixed salary schedule that takes into account years of experience and education but does not consider teachers' performance. Currently there is a great interest in adding performance-based components to teacher salaries. This interest is driven by both state and local initiatives (including Florida, Minnesota, Texas, and Denver, Colorado among others) and the federal Teacher Incentive Fund (TIF) program.

A key component to the new wave of performance-based pay initiatives is the use of student achievement data to evaluate teacher performance. The testing requirements of the No Child Left Behind Act (NCLB) have resulted in greater numbers of students being tested than ever before. Moreover, annual testing in grades 3 to 8 and one grade in high school has yielded longitudinal data on students. At the same time as greater amounts of data are being collected, researchers have been developing and applying innovative statistical and econometric models to the longitudinal data to develop measures of an individual teacher's contributions to his or her students' learning (Sanders, Saxton, and Horn, 1997; Webster and Mendro, 1997; McCaffrey et al, 2004; Harris and Sass, 2006). Generally referred to as value-added models, this class of models has demonstrated significant variation among teachers in their performance and demonstrated that this variation among teachers is a significant source of variation in student outcomes (Rivkin, Hanushek, and Kain, 2005, Kane, Rockoff, and Staiger, 2006, Gordon, Kane and Staiger, 2006).

The class of value-added models includes a wide range of models with a common feature of using students' prior achievement to account for student inputs to their learning and to separate these inputs from the inputs of their teachers. Such controls are necessary because students are not uniformly assigned to classes and some classes include students with significantly lower levels of achievement and at significantly greater risk for poor achievement. The alternative specifications for value-added models are likely to yield estimates of teacher performance with different statistical properties in terms of the level of error in the estimates and amount of residual confounding between estimated teacher performance and the characteristics of the students in the teacher's class.

A few studies have compared subsets of the various value-added estimators of teacher performance (McCaffrey et al., 2004; Harris and Sass, 2006, Buddin et al., 2007) but few have considered the alternative explicitly in the context of using the estimates as the basis of performance-based pay. Moreover, there has been very little consideration of aspects of the process of generating performance measures and using them to award teacher bonuses or make other compensation decisions.

This paper directly addresses these issues. It describes the process of taking a large administrative database of student test scores and class assignments and turning it into bonus decisions for teachers. The choices to be made at each step of the process are illuminated with careful consideration of impact on the types of teachers who receive awards of the choice of performance measure and decision rule for awarding bonus. Middle school mathematics data from a large urban school system provide the case study of the process of awarding teachers a bonus on the basis of their students' progress. We begin by describing the case study data and then turn the discussion of using data like these to determine teacher compensation.

2. CASE STUDY DATA

In this paper we use a case study of generating performance measures for middle school mathematics teachers and simulation studies of alternative decision rules for awarding bonuses to illuminate some of the issues that arise in using student achievement to determine teacher compensation. The data are from a large urban school district in which roughly 50 percent of tested students are African-American, 36 percent are white, 11 percent are Hispanic, and about 3 percent are Asian or other ethnic group. We focus on middle school students and their teachers, using data for the district's traditional middle schools, including magnet schools but excluding special education and alternative schools. The district treats grade-levels five to eight as its traditional middle school grade-levels and we focus only on teachers of students in these grades. Our study population includes all 37,887 students who attended any of the district's middle schools in grades five to eight for at least part of one or more of the school years 2004-05, 2005-06, or 2006-07.

The teacher population of interest for this case study consists of every teacher who taught mathematics to the students in the student sample during the 2005-06 or 2006-07 school years. Using the administrative data, we identified 478 teachers in the 2006-07 school year and 476 teachers in the 2005-06 school year in our population.   Of these teachers, 338 were in both years for a total of 615 teachers in our case study.

The administrative data are collected from three sources: test score files, enrollment history files, and course enrollment files. The test score files include both testing data--the student test scores and grade-level of testing--and the background variables such as race, gender, and special education status. The enrollment history file contains enrollment related transactions that can be used to determine the dates that every student was enrolled in each middle school during each school year. The course files provide by school the cumulative record of all the courses in which a student was enrolled during the school year. The course data include course titles and teacher identifiers. The course file does not provide any information about the dates in which the student was enrolled in any course. The course file is used to link students to individual teachers.

The data sources were linked using student identifiers to create a database with one record for each student in the study population for each year of the study and each tested subject.[1] Each record includes a school identifier for the school a student attended that school year, the student's grade-level for the school year, student demographics such as race and gender, and special education status when available. Using rules described below, we also assigned to each record (defined by a student, year, and subject) an identifier for the teacher who was accountable for the student's learning in the specified subject for the given school year. The rules for assigning students to teachers resulted in some students being assigned no teacher for a given subject in a given school year even if the student was enrolled in the district's middle schools. In such cases, and in cases where the student was not enrolled in the district, the teacher links are set to a missing data value.

Students in the case study school district are tested each year on statewide assessment. In the 2003-04, 2004-05, 2005-06, and 2006-07 school years students in grades 3 to 8 were tested in mathematics, reading-English language arts (ELA), social studies and science. In 2002-03 students in grades 3, 5 and 8 were tested in mathematics and reading. The mathematics and ELA test scores are presented on a developmental scale with scores linked across grades and test years from 2003-04 forward. The social studies and science tests are not vertically linked or linked across school-years. These tests are scaled to have roughly the same mean and variance at each grade-level and year of testing.

[1] For students who were retained in a grade or skipped a grade the database contains a sequence of records for each series of grades in which the student was enrolled. For example, for a student who was enrolled in grade 6 in the 2005-06 school year and retained in grade 6 for a second year in the 2006-07 school year, the file includes two series of records for the student-- each with consecutive grades. The first has the student in grades 4, 5, 6, 7 in the 2002-03 to 2006-07 school years with missing values for the 2006-07 school year. The second sequence has the student in grades 3, 4, 5, 6 for the 2002-03 to 2006-07 school years with missing values for every year except the 2006-07 school year.

In our analyses we used both the raw mathematics scale scores and transformed values we created to improve the bivariate relationship between tests from adjacent years. We call our transformed values z-scores or rank-based z-scores. We first transformed each raw scale score to its percentile in the empirical distribution function of scores for a given grade-level and school year. The percentile of the empirical distribution function equals the rank of the score, divided by one plus the total number of students with scores in this grade-level and school year. We then transformed the percentile to the corresponding quantile from a normal distribution. In the base year of 2003-04, we transformed the percentiles using the quantiles of the standard normal distribution. In later years we allowed the mean and variance of the normal distribution to shift to account for possible trends in the scores relative to the baseline. See Appendix 1 for details on this transformation.

The database also includes teachers' scores on the Learning Mathematics for Teaching Project's Multiple Choice Measures of Mathematical Knowledge for Teaching (LMT) for a sample of 126 mathematics teachers teaching in the district during the 2007-08 school year. The LMT was developed by a team at the University of Michigan to measure the mathematical knowledge specific to teaching (Hill & Ball, 2004; Humphrey & Wechsler, 2007; Humphrey, Wechsler, & Hough, 2008a). It has been used successfully to document: teacher learning in a California professional development program (Hill & Ball, 2004); distributional problems in teaching knowledge in a nationally representative sample of middle school teachers (Hill, 2007); and differences between certification routes (Humphrey, Wechsler, & Hough, 2008a, 2008b). Finally, the instrument has been used to document a relationship between teachers' LMT scores and their students' achievement in first and third grade (Hill, Rowan, and Ball, 2005).

Extensive equating and validation work has been done on the three forms of the instrument (Hill, Ball, Blunk, Goffney, & Rowan, 2007; Hill, et al., 2004; Schilling, Blunk, & Hill, 2007; Schilling & Hill, 2007). The instrument measures knowledge in two domains – content knowledge and knowledge of content and students – across three areas of mathematics. Those three areas– number concepts and operations, geometry, and patterns, functions, and algebra– make up much of the K-8 mathematics curriculum. Analyses of item response theory reliabilities indicate all three forms of the instrument had good to excellent reliability, ranging from 0.71 to 0.84 on subscales and 0.916 to 0.931 for whole forms. Coefficient alphas ranged from 0.450 to 0.821 on subscales and 0.845 to 0.888 for whole forms of the test (Hill, Schilling, and Ball, 2004).

The LMT scores in our data were based on a subset of items from the K-8 form. The items were chosen to be most relevant to middle school mathematics while limiting the survey burden on teachers. All middle school mathematics teachers in the district were invited to complete the survey and roughly 40 percent did so. An informal review of the items suggests that they focus on algebraic skills and reasoning more so than other mathematics concepts.

3. CREATING THE STUDENT ACHIEVEMENT DATABASE

The process of awarding teachers bonuses on the basis of their students' achievement begins with the task of creating a database of student achievement data linked to the students' teachers. The importance and inherent challenges of this first step in the process are often overlooked and it is taken for granted that clean data appropriate for analysis already exist and will be used for the estimation of effects and the final bonus decisions. In our experience, administrative data, even from good systems, requires extensive processing before it can be used to generate performance measures. The processing involves making numerous decisions that do not necessarily have right or wrong answers.

What Teachers to Include

One of the first tasks is to determine which teachers are eligible for bonuses and what courses are to be included when measuring a teacher's performance. Ideally, the policy maker might want every teacher to be eligible for a bonus and to award bonuses so that the "best" teachers receive bonuses regardless of what jobs the teachers have. This ideal is unlikely to be feasible for systems designed to award performance measured by student achievement tests. For many subjects, such testing does not exist or does not exist at some grades. Hence, assuming that administrative data will be used for bonuses requires that bonuses be restricted to teachers teaching tested subjects and grades. Also, if prior test scores are to be used in the estimation of effects, bonuses will need to be restricted to teachers in grades after the first grade of testing. Even in systems that combine achievement based on test scores and other measures, the achievement based component must be restricted to teachers of tested subjects and grades.

Identifying teachers to be measured via student performance can be more challenging than might be expected. When making the choice for determining which teachers to evaluate for specific teaching tasks such as a subject or a subject and grade-level, there are two considerations. First, for every teacher, the determination must be made about which subjects and grade-levels constitute a sufficiently substantial portion of the teacher's job responsibility and are sufficiently related to the standard achievement assessments so that his or her students' performance on the tests in this grade-level are appropriate for evaluating the teacher's job performance. Second, because objective standards of performance for teachers generally do not exist, most performance measures and all compensation decisions depend on the teacher's performance relative to some set of peers who are also accountable for their performance teaching a grade-level and subject. Hence, for each teacher the determination must be made as to which peer groups he or she should be assigned for establishing all teachers' performance levels and compensation. Such determinations might be made by school administrators, on the basis of rules applied to administrative data, or a combination of both sources (e.g., administrators are given guidelines for selecting teachers for evaluation via each tested subject and grade-level and their decisions are checked via applying the rules to administrative data or visa versa rules are applied to administrative data and the results are verified by school administrators).

Regardless of the means for determining which teachers' performance to measure using student outcomes on a given test and grade-level, several factors make such determinations challenging. First, some teachers teach very few students in a particular subject. Resource and special education teachers might teach very few students in any particular subject or from any particular grade. Similarly, teachers with special language skills might teach small numbers of students. The class sizes might be even smaller when students who are unlikely to complete are excluded. One must decide if a teacher with very few mathematics students meets the criteria for being a mathematics teacher for evaluation purposes. Should a teacher with very few mathematics students be accountable for this instruction when the teacher primarily teaches other subjects? Should such a teacher be used in evaluating the relative performance of other mathematics teachers?

Class size is not the only factor that can make choices of which teachers should be evaluated on a subject complicated. The range of courses being taught is also a complicating factor. Many secondary schools offer a complex array of courses to students even in the basic subject areas of mathematics or English language arts. In our middle school data, mathematics courses include special education courses, sheltered mathematics courses for English language learners, traditional grade level courses, pre-algebra, algebra and advanced courses, including geometry and algebra II. It seems quite clear that most policy makers and researchers would to want to consider teachers of all these courses as mathematics teachers and hold them accountable for student mathematics performance. It might be debatable, however, whether the standardized statewide assessments adequately measure the learning that the teacher is responsible for supporting. This is especially true for special education and advanced courses.

It is also less clear whether teachers of these diverse courses and grade-levels should be combined when measuring the relative performance of teachers or when making normative decisions for awarding compensation. That is, one must decide if the performance of a fifth grade teacher teaching mathematics in a self-contained classroom should be compared to the performance of a teacher teaching algebra to eighth graders. This decision will affect the determination of teacher population and must be carefully considered when developing performance measures because most performance measures and compensation systems rely on teachers' performance relative to a peer group to assess performance and determine compensation.

Educators, policy makers, and researchers do not generally have an objective standard for evaluating teacher performance. Typically, performance measures implicitly define a teacher's contribution to student learning by comparing student outcomes when taught by the teacher to counterfactual outcomes that would have occurred under alternative conditions. The implicit use of a counterfactual to define teacher contributions to learning results in performance measures being relative to a reference point that functions as the counterfactual. Common reference points are the average performance for students across teachers in a district or state. For example, models that estimate effects with linear regression models (such as the ANCOVA, regression residuals, multilevel mixed models and fixed effects approaches described in the following section) typically include school district-specific intercepts by grade-level that make the reference point the average performance of teachers in the district who are teaching students in the same grade-level and included in the teacher population according the rules established when creating the data. Methods that use data from the entire state to estimate expected performance (e.g., the lookup-table method described below or the Student Growth Percentiles measure used in Colorado, Betebenner, 2007) use the average performance across teachers from the entire state as the point of reference.[2]

Beyond the estimation of effects, compensation decisions also typically involve comparing performance measures among teachers combined into a peer group (e.g., all middle school mathematics teachers). For example, the XX system in Florida proposed awarding teachers in the top quintile of their peer group in the state. Other methods award teachers compared to the statistical population describing the general distribution of teacher performance (e.g., awarding bonuses to all teachers above the 80th percentile of an appropriately scaled normal distribution).

In any of these methods for measuring performance or awarding compensation, the choice of the peer groups can be significant. Changing the teachers included together in a peer group can change a teacher performance measure and how that measure is valued. Different groups can be used in estimation and compensation decisions, but the decision of which teachers to evaluate on a specific subject and include in the population needs to be carefully considered relative to the goals of the system and the performance measure to be used.

[2] Regression methods are all based on predicting students' performance under some alternative condition and comparing the students' performance to the expected value to determine the teacher's input. One method for estimating performance measures that does not require implicit comparisons to a standard in the estimation process is average gain score (i.e., the student's current year score less his prior year score). Gain scores estimate performance as the change in performance for a teacher's students without centering these gains against a standard. However, implicit in this estimation method is the assumption that in the absence of variation in teacher effectiveness, average gain scores would be equal across classrooms.

The choices of teachers to evaluate using a given test are even more complex for language arts tests because there are many language arts related classes that might not be considered the primary course work of students. For example, in our middle school database students were taking courses in drama and speech as well as standard language arts courses. Whether or not teachers in these courses should be considered language arts teachers is open to debate. In addition, many students in our middle school data had reading classes separate from language arts. Again, it is not clear if reading teachers and language arts teachers should be evaluated by a reading test, a language arts test, or both. Similarly, whether or not comparative samples should combine these teachers or be restricted to only reading and only language arts teachers is a choice with no clear answer but the potential to significantly affect inferences about individual teachers and compensation decisions. Another consideration for language arts and reading is the contribution of social studies teachers to these skills. Social studies teachers often emphasize reading and writing skills in their classes. Should these teachers be evaluated by their students' performance on these tests? Should these teachers be included in normative samples for determining the performance of teachers teaching reading and language arts courses? Similar and equally vexing problems arise when considering other subjects.

In the case study presented here, we focus on estimating the performance of teachers of courses labeled as mathematics in one of the district's standard middle schools. We did not restrict the courses included in this sample and we did not control for courses in our analysis. Also, we included teachers teaching any of the grades in the middle schools and combined data across grades for teachers who taught students from multiple grades. We did not restrict the sample by the number of students the teachers taught in mathematics or the courses taught, including special education courses.

Selecting students for evaluating teachers

After identifying which teachers will be evaluated for the performance of students from a given population on achievement in a selected subject, the students to be used in the evaluation must be selected. The challenge here arises because not all students are taught by a single teacher for the entire year for a given subject. It must be determined for which students a teacher should be held accountable. For example, most people would believe that a teacher probably had minimal effect on a student in his or her class for only a few days and most teachers would probably feel being held accountable for such a student's progress would be inappropriate.

The cut point for the length of time is a policy question. The cut point must meet general expectations about which students' outcomes might reasonably attributable to actions of the teacher. On the other hand, choosing to exclude students from teacher evaluations has the potential negative consequence of encouraging teachers to focus on other students. More restrictive rules for inclusion will create greater potential for negative consequences because more students will be at risk for exclusion and at risk for reduced attention for a larger portion of the school year. For example, if we exclude only students who were in a teacher's class for less than 2 weeks, then we will put few students at risk for decreased attention and the period of decreased attention will be just two weeks. If we exclude students who were in the teacher's class for less than 95 percent of the school year, we will exclude more students and a student could be at risk for less attention for up to 95 percent of the school year. However, including students for whom the teacher might have had only very minimal input could weaken teacher confidence in the evaluation system and potentially add error to the performance measures.

In the case study presented here, we decided that the student needed to be in the teacher's class for greater than 95 percent of the school year. Other rules might be to include only students who were enrolled continuously from the start of school until testing or were enrolled above a fraction of days in the months prior to testing.

Conversely, all students might be linked to their teachers but the time spent in the teacher's class might be factored into the estimation of teacher effects. There is little research on the relative contributions to learning of multiple teachers on students taught by more than one teacher in a school year. Weighting by time in the class might be one solution. This would add complexity to statistical analyses.

Another complication is students linked to multiple teachers of the same subject from the same school during the school year. For example, a student might be linked to both a special education and a traditional classroom teacher for mathematics, or the student might be linked to both a reading and language arts teacher. There are again two challenges with such situations. First, there is the challenge in determining how much time each teacher taught the student and, second, there are the considerations of for which students teachers should be accountable when students are being taught simultaneously by multiple teachers and how to apportion responsibility to the multiple teachers.

In our experience with our case study data and other databases, links to multiple teachers indicated different patterns of course taking. Links to multiple teachers might mean that a student changed courses mid-year, for example switching to a more advanced mathematics class or transferring between regular and special education classes. Multiple links can also imply that a student was taught simultaneously by two or more teachers, for example, students receiving reading instruction from one teacher and language arts instruction from another. We also found that it could mean course assignments in the early part of the school year were incorrect in the database and later corrected without removing the errors. Depending on rules for assigning teachers, the alternative meanings for the same data can have different implications for processing the data and the ambiguity might make it impossible to implement some desired rules.

Again, the correct way to apportion the contribution of multiple simultaneous teachers is not clear. If the data are very detailed then the time spent with each teacher might be available, but it is not clear if this truly accounts for contributions and if our experience with high quality administrative data is representative of other school districts' data, it is very likely that such detailed data will not exist. Moreover, detailed data might not be completely accurate. We found that when teachers reviewed rosters of students assigned to them using our administrative database that a significant proportion of teachers identified errors in assignments. Some of the errors included students not properly identified as leaving the school or students changing classes, often to special education, without this information being properly captured in the database. Our experiences are limited primarily to one school system's data and another district's data might be more (or less) accurate. However, we believe that careful evaluation of the quality of course data will be a necessary component of any compensation system.

Many of the problems associated with linking students to teachers, other than data errors, could be eliminated or mitigated if all the teachers were assigned students according to the proportion of total instruction in a subject the student received. The data would need to be accurate and provide details on time spent with each teacher. However, using this method to ameliorate the problems of linking students to teachers might actually create complexities for estimating teacher performance measures. Also, it might be challenging to explain to teachers, especially given the lack of evidence for the appropriate weighting of students taught by multiple teachers.

If the approach is taken that students will be linked to teachers only if they meet certain criteria (e.g., they have been in the teacher's class for a sufficient time and have not been taught by multiple teachers), then some students might be in the database but not link to a teacher for a given subject and school year. These students could be deleted from the data or retained in the data without a teacher link. Deleting these students can create problems for longitudinal analyses since their scores in prior years might be lost due to the deletion rule. Retaining the students requires assigning them a teacher. In our case study, we retained these students and assigned them to a teacher whose effect was set to zero in multivariate models. The effects of this modeling choice have had limited evaluation in the literature; however, McCaffrey et al. (2005) compared this method to alternatives in the context of multivariate mixed model fit in the Bayesian framework and found estimated teacher effects were robust to this decision.

In the current case study, we linked teachers to students only if the student linked to the teacher for over 95 percent of the school year and the student had only one mathematics teacher in the school for the year. In 2007, XX percent of the students enrolled at any time during the school year were not linked to a mathematics teacher. Of the students enrolled in a single school for over 95 percent of the school year only XX percent were not linked to a mathematics teacher. If we had applied our rules to English language arts, only XX percent of enrolled students would have had links and only XX percent of students enrolled in a single school for over 95 percent of the school year would have had teacher links.

Other factors in creating the student achievement database

Administrative data are often messy and challenging in part because their use for administrative purposes differs from use for analysis purposes or because student experiences are varied to meet their complex individual needs. Some of the challenges posed by the data include:

a. students with data from one administrative data set that contradict data in another data source, for example, grade-level in the test score files differs from the grade-level in the enrollment file for a given school year;

b. students who are missing data from one or more file, for example students might be in the enrollment file for a given year but not in the course listing and test score files or students might have records in the test score file but not in the enrollment files; and

c. students who have multiple records for a given year in one or more files, for example students with two different sets of test scores for the same school year.

These types of problems are familiar to any analyst who has used administrative data to address research problems. Analysts typically develop rules for cleaning data and assume that lost data or remaining errors have a minimal effect on their conclusions. Sensitivity analyses or additional modeling might be conducted when the proportion of records with errors is not trivial.

Such common data errors, however, might be more problematic when using data to estimate teacher performance and award compensation. Dropping student records can signal to teachers that certain students do not count and could have negative consequences for these students. Hence, practices to retain all the available data are likely to be important. Furthermore, teachers who are unsatisfied with the compensation decisions might question their performance measures. Any data of questionable accuracy might support their challenges to the compensation system and could undermine other teachers' confidence in the system, possibly eroding any motivating effects the system might have. Careful review of the data with knowledgeable district staff and teacher verification of data prior to assigning awards are likely to be necessary components of any compensation system. For our case study we had a district employee research questionable data and provide accurate values when possible. However, because our primary goal was to demonstrate issues and compare performance measures and bonuses, we also used some of the standard approaches to data cleaning for the examples presented in this paper. For example, we changed values when data were inconsistent and deleted test scores when grade-level from the enrollment file and test-score files disagreed.[3]

4. MEASURING TEACHER PERFORMANCE

Value-added modeling entails a great variety of statistical and econometric approaches for analyzing longitudinal test score data to estimate teacher effects. We employed 24 different approaches to our middle school mathematics data to estimate teacher effects for the 2005-06 and 2006-07 school years. In this section, we describe the various methods, provide some general summary statistics for the methods and then compare four methods in detail to highlight how different teachers might fare under the alternative methods. In the next section, we discuss the effects of alternative rules for awarding bonuses on the basis of these performance measures.

Before describing the individual estimation methods, we consider two general issues that must be considered when creating performance measures. These are the unit of measurement for the teacher and the frame of reference for the performance measures.

Teachers teach various groups of students defined by course, grade-level, school year, demographics and possibly other factors. When defining a teacher's performance, we need to determine how the performance across these varied groups will be used to determine compensation. Compensation might be based on average performance, or weighted average performance, performance on a subset of students (only students for the current school year or only students in mathematics courses) or a combinations of performance with multiple groups (e.g., the teacher receives a bonus if they exceed a threshold of performance for students in every course he or she teaches). Estimating performance on different subgroups of students may provide better information about variation in a teacher's performance but it can also introduce greater sensitivity to the idiosyncratic outcomes of specific students or sampling error. Thus there is a need to balance between these two goals of estimation. The correct balance will depend in part on the goals of the compensation system and the likely size of sampling error in the estimates. Given the generally large levels of sampling error, we suspect that rather coarse sub-sampling will be preferable in most cases.

Decisions about the how to measure teacher performance for teachers teaching in varied contexts, however, must be made in conjunction with decisions about which students and subjects a teacher is accountable for and for which populations of teachers a teacher's

[3] The analyses presented here are not the analyses used by the National Center on Performance Incentives in its POINT experiment or other work with individual teacher compensation. The approach used here was chosen to facilitate the case study and estimates were not used in any way to provide compensation to teachers.

performance will contribute to normative assessments of teachers. Such decisions must be based on considerations of what performances are comparable and what student outcomes the teacher is likely to have affected. But they must also be based on what inferences the available data and statistical methods are able to support. Once the decisions are made they determine how the data must be coded and how the statistical procedures of the performance measure must be implemented.

As noted above, because performance measurement and compensation are almost always determined on a relative basis it is important to consider the frame of reference when determining the implementation of a performance measure. The choice of teachers to include in the population in part determines the frame of reference for evaluating performance, but modeling choices can also change the normative group used for evaluating a teacher's performance. For example, including course indicators in a model implicitly restricts some comparisons to teachers teaching the same course. Again, the appropriate comparison depends on what performance the policy makers believe should be comparable and what estimation is feasible.

In our case study, we assumed that comparisons of middle school mathematics teachers would be relevant to many practitioners considering implementing such a compensation system and combined together all such teachers, although our models included a grade-level specific intercept so that students were compared to their counterfactual outcome across the average teacher in the district or state for their grade-level. In our case study, we also aggregated estimates to teacher by school-year. We assumed that compensation was an annual decision so that estimates would be made annually and would be relevant. To reduce sampling error we combined across grade-level, course and demographic groups. This also allows for all middle-school mathematics teachers to be compared as was our desired comparison.

In our estimation of teacher performance for a given year, we use only data from that year or prior school years. Data from future years is not used. For example, we do no use data from the 2006-07 school year when estimating the teacher effects from the 2005-06 school year. Some estimation methods (e.g., multivariate mixed models or fixed effects methods, see details below) could use future data by jointly modeling all the years of data and waiting to produce estimates for a given school year until future year data are available. We felt it was unrealistic to assume that compensation decisions would be delayed for more than an entire school year in order to include future data in the estimation of teacher performance. Hence, we do not use such data in our estimation.

4.1 Performance Measures Compared in the Case Study

As shown in Table 1, the set of performance measures we consider can be classified according to three factors: by method, scale of the data (raw scale score or z-scores), and statistical adjustment (whether or not estimates were adjusted via shrinkage). Shrinkage is a statistical adjustment to reduce the error in estimates; details are provided below. We consider eight different methods. The methods can be grouped into five categories: methods based on deviations from expectation estimated with a single score; methods based on gain scores methods; methods based on deviations from expectation estimated with multiple scores; methods based on multivariate mixed models; and methods using fixed effects. For each estimation method we derived performance measures using both raw scores and z-scores and both with and without shrinkage. The exceptions were the two multivariate mixed effects models which take extra computation time to compute via Monte Carlo simulation and the lookup table method which requires data from the state for estimation of performance measures. For the multivariate mixed models, we consider only estimates based on z-scores with shrinkage. When using z-scores for estimation both current and prior year z-scores were used in modeling. When using raw scores for modeling, both current and prior year raw scores were used. The state could not provide us with z-scores, so we only estimated lookup table methods with the raw score data. All estimation methods were used to estimate the performance of 2005-06 teachers and separately for the 2006-07 teachers.

Table 1 About Here

Performance measures based on deviations from expectation estimated with a single score

Methods based on deviations from expectation estimated with a single score method share the common feature that the teacher's effect is estimated as the average of this or her students' deviations from their expected scores given their prior year mathematics test scores. We use three such methods. The first, which we refer to as regression residuals, uses linear regression to estimate the expected outcome using the single prior year math score for students. For each student, we then calculated the regression residual as the difference between the student's observed score and his or her expected score from the linear regression model. The teacher's performance measure equals that average of these residuals across all the students in his or her classes regardless of grade-level. Students who were missing prior scores or without standard grade progressions were excluded from the estimation.

We calculated both raw averages and shrunken averages. Shrunken averages are sometimes referred to as empirical Bayes or Stein estimates (Carlin and Louis, 2000). Theoretical results show that the average squared error (the square of the difference between an estimate, teacher performance measure, and the true quantity of interest, true teacher performance) can be minimized across an ensemble of estimates (e.g., teachers) if the individual estimates are combined with the average estimate to produce shrunken estimates. For example if $\hat{\theta}_i$ denotes the performance measure for teacher *i*, and $\overline{\theta}$ denotes the average estimate across all teachers, then the shrunken estimate for this teacher is $c_i \hat{\theta}_i + (1 - c_i)\overline{\theta}$. In our case, $\overline{\theta} = 0$ for all our estimates so we are shrinking the estimate toward zero. The shrinkage factor, c_i, depends on the ratio of the estimate of the sampling error or noise in the estimate for teacher *i* to the variability among the estimated effects for all the teachers adjusted to account for noise--the greater the ratio of noise to variability among teachers, the smaller the value of c_i. The primary effect of shrinkage is to shrink the estimates for teachers with very few students (less than 10) and it had minimal impact on teachers with 20 or more students. Details on this estimation method are in the Appendix.

Also in the class of performance measures based deviations from expectation estimated with a single score is what we call the ANCOVA estimate. Like regression residual, we estimate a student's expected achievement using a linear model. However, in this model we control for a student's teacher in 2007 when estimating the regression coefficient for prior mathematics achievement. This is the linear model that corresponds to the traditional analysis of covariance (ANCOVA, Snedecor and Cochran, 1980), so we call it the ANCOVA method. If true teacher effectiveness is correlated with the prior achievement of students, (e.g., more effective teachers are assigned to classes with students who achieved higher in the past), then regression adjustments without controlling for class assignment can result in bias estimation of the coefficient for prior achievement and could potentially over adjust scores by attributing some of the true teacher performance to the student's prior scores. Ballou, Sanders, & Wright (2005) provide additional details on this potential bias. The ANCOVA method avoids this potential bias.

By default the linear model fit for the ANCOVA yields estimated teacher effects. However, the default estimates from many statistical software packages including SAS and Stata will not be appropriate as measures of teacher performance. The ANCOVA model is over-parameterized because the teacher effect for every teacher and the overall mean cannot be uniquely estimated. Most statistical packages account for over-parameterization by setting one of the teacher effects to zero. All other teacher effect estimates are then differences between the holdout teacher (the teacher whose effect is forced to zero) and other teachers. Estimates are sensitive to the choice of holdout teacher and can be extremely unstable across years. The alternative is to constrain the estimated effects to sum to zero. Estimates are relative to the overall mean and are stable across years. Our estimates use this constraint.

The final performance measure in the performance measures based deviations from expectation estimated with a single score are lookup-table estimates. Again we estimate a student's expected performance using the single prior year mathematics achievement score. However, rather than use a linear regression model to predict the expected score from the prior score, we use the average current score for all students in the state with this exact prior score as the estimate of the expected score. For example, we identified all eighth graders in the 2006-07 school year in the state who scored 503 as seventh graders on the state mathematics test in 2005-06. We averaged the eighth grade scores for these students (510.5) and this is the expected score for a student who 480 as a seventh grader. Even at the state level some scores were very rare and the average scores were noisy. We removed this noise by smoothing across prior scores within a grade-level and forcing the expected scores to increase monotonically with prior scores. That is, we forced the expected score associated with a lower prior score to be lower than the expected score associated with a higher prior score. This smoothing resulted in very minor changes to the raw averages for almost all prior scores. The final estimates of expected scores create a lookup table were the rows are the prior scores and the column value is the expected score.

For each student, we found the expected score from the lookup table and calculated the difference between the student's observed and expected score. The teacher's performance measure is the average difference between observed and expected scores for his or her students. We average scores across all grade-levels. We again created shrunk and raw estimates. Because the state could not provide us with z-scores, we were unable to produce estimates with z-scores.

One advantage of this method is that it does not rely on linearity so nonlinearity in the bivariate relationships between current and prior year scores is not a problem for this estimator. However, z-scores could potentially reduce the problem of heteroskedasticity which also exists in the data and results in greater variability among students with extremely high or low prior scores. Another advantage of this method is that its reference is the state, so teachers in a single district are not directly competing against each other. Under this method all the teachers in a district could have positive or negative performance measures. For the case study presented here, we were not concerned with competing teachers in the same district and most of the other estimators had the district as the point of references, so we forced the average effect for this method to be zero.

One feature common to all three of the methods is that students who are missing the prior year mathematics score are excluded from estimation. Implicitly teachers are not held accountable for these students. This could create incentives for teachers to give less attention to these students and result in negative consequences for the system.

Another common feature of these methods is that they do not require scores to be on a single developmental scale for the methods to be interpretable. The prior score is used to predict estimation and there is no assumption that it must have a particular relationship to the current score other than it predicts likely performance and deviations from that prediction should not depend systematically on factors other than the current year teacher. This last requirement is unlikely to hold in practice because of the limited information about students contained in a single prior test score. The errors created by the failure of this assumption are discussed below.

Performance measures based on gain scores

A commonly considered method for estimating teacher performance is the average gain score for the teacher's students. A student's gain score equals the difference between her current year score and her score from the previous year on the same subject. We created both raw and shrunk estimates.

Gain scores have a meaningful interpretation only when scores are linked across grades to be on a single scale, so that a score of 500 at any grade-level can be interpreted as corresponding to the same level of achievement. There is considerable debate about whether linking of tests can provide data that support measuring gains (c.f., Martineau, 2006). These concerns should be carefully considered for any system that would use this performance measure.

As an alternative to gain scores on the raw scale score data, we also estimated gain scores using z-scores. These correspond to students' change in their rank or place in the distribution of scores. A single developmental scale is no longer required but the gain no longer measures growth, but rather growth relative to the distribution of growth.

For this measure we estimate both shrunk and raw estimates for both the scale-score and z-score gains.

Performance measures based on deviations from expectation estimated with multiple scores

One of the limitations of performance measures based on deviations from expectation estimate with a single score is the limited information about a student's performance provided by a single score due to the measurement error in the test and the inherent year-to-year fluctuation in a student's performance. One approach to overcome this limitation is to use additional data when predicting expected scores. We created such a performance measure by using all the available prior scores in mathematics and other subjects to predict student achievement in mathematics in 2006-07 and then again using separate models for the 2005-06 data. We controlled for classrooms when estimating the intercepts and slopes, so this model is a multivariate or multiple covariate extension of the ANCOVA method and we refer to this performance measure as the Multivariate ANCOVA measure.

This multivariate ANCOVA not only provides a potentially better means for controlling for student inputs but this method also potentially reduces the noise in the estimated performance measure using more background data for estimating expected scores compared to the ANCOVA method. It also uses more students by including students with any prior achievement data and not only those with prior year mathematics scores.

Given that many students did not have complete data from all the prior testing, we could not exclude students with incomplete data. Rather we used all the prior score data available for each student. For instance, if a student only has scores from 2005-06 we used those scores; but if a student had data from 2005-06, 2004-05 and 2003-04, then we would use of all of this data in our estimation. Because different students have different patterns of data we needed to fit separate models for each pattern of response. However, we wanted to control for teachers with a single teacher control, within grade-level. To accomplish this estimation, we created indicator variables for a student's pattern of responses and included in the model terms for the interaction of these indicators and the available prior year test scores. A separate model was fit for each by grade-level and year of outcome testing (2005-06 and 2006-07). We fit models by grade-level because of the complexity of estimating different models by response pattern.

We enforced sum to zero constraints for estimated teacher effects within grade-level by calculating adjusted residual values for each student that control for prior score and overall intercept but not the teacher.

There were several potential methods to combine values across grade-levels for each teacher. One method combined the adjusted residual and then averaged these. Another method averaged the values by grade-level and then averaged the averages. A third method averaged the residuals by grade-level and then created a weighted average of the grade-level values where the weights equal the precision (the reciprocal of the square of standard error due to sampling error) of the mean for each year. Preliminary analyses suggested the third method provided the measures with the best statistical properties so we report the estimates based on that method. Differences across these alternative methods for combining across grade-levels are small, so conclusions we make based the chosen performance measures hold for the alternatives.

For shrunken estimate we again had multiple options. The performance measure we report in this paper shrunk the averages by grade-level and then combined them with precision weighting.

We repeat the estimation using raw scale scores as outcomes and inputs and using z-scores as outcomes and inputs.

Performance measures based on multivariate mixed effects models

Multivariate mixed effects models develop a statistical model for the entire vector of student scores. The model includes random terms for teacher inputs to those scores and allows multiple scores from the same student to be correlated via an unspecified correlation structure. Details on these models are found in McCaffrey et al. (2003) and Lockwood et al. (2007). In our estimation we model only mathematics scores. For a limited example we modeled both mathematics and ELA scores and found this resulted in trivial differences in estimated teacher effects.

Following Lockwood et al. (2007) we fit the models using a Bayesian framework and the posterior mean as our estimate of a teacher's performance and the posterior standard deviation as the measure of error in that estimate. For simplicity we refer to the posterior standard deviation as the standard error of the estimate, although technically this is incorrect, but the difference is inconsequential.

One complexity to modeling outcomes via multivariate mixed effects models arises from students' membership in different classes each year. The model must then account for the effect of prior year teachers on students' later outcomes. McCaffrey et al. (2004) and Lockwood et al. (2007) develop a model that allows prior year teacher inputs to contribute additively to current year scores but weighted by persistence parameters estimated from the data. Lockwood et al. (2007) call this the variable persistence model and we use this model to estimate performance measures. Alternatively, Sanders, Saxton and Horn (1997) and Ballou, Sanders, and Wright (2005) suggest a model where prior teacher effects contribute to students' outcomes additively without any weighting. In terms of the variable persistence model, the persistence parameters are assumed to equal 1. Following Sanders et al. (1997) we refer to this as the layered model and also use it to estimate performance measures.[4]

We fit these models separately by cohorts defined by grade-level in 2006-07 and grade-level in 2005-06 to estimate teacher effects separately for these two school years. For each teacher we combine the estimates from the different grade-level using precision weighted averages.

Because the models rely on multivariate normality and are time consuming to fit through Monte Carlo methods, we created performance measures only with z-scores. Shrinkage is implicit in mixed model estimation; hence we only produced shrunken estimates.

Performance measures based on fixed effects

A common econometric approach to remove possible bias when estimating program effects using longitudinal data is to estimate changes in outcomes within unit (e.g., a student) by controlling for the unit's average level of outcomes via "fixed effects" or indicator variables for each unit. If we think of the current year teacher as treatment applied to a student, the fixed effects approach applied to estimating teacher performance entails fitting a linear model to multiple years of student data with indicator variables for students (to remove the student's fixed effects and estimate the teacher's performance

[4] Lockwood et al. (2007) refer to this as the complete persistence model.

using the difference between the student's performance in the teacher's class and his or her average performance) and indicator variables for the "treatments" or the teachers. Our implementation of fixed effects using fixed effects with level scores is appropriate if teacher effects do not persist at all across years. In the economics literature, fixed effects on gain scores are also applied, which we did not apply because of our relatively short data series on some students.

We implemented this approach using both raw and z-scores. We estimated separate effects by cohort defined student grade-levels in 2006-07 and again by grade-levels in 2005-06 to support estimation of performance measures separately for these two school years. For each teacher, we combined the estimates across grade-levels using precision weighted average. We created both raw and shrunken estimates. For the shrunken estimates we shrunk the estimates by cohort and then combined the shrunken estimates.

Fixed effects estimation is challenging with large samples because the models must account for many students. Statistical software has methods to handle this challenge but the default estimates for teacher effects from standard software cannot be used as performance measures because the estimates are relative to a holdout teacher rather than based on the sum-to-zero constraint. In our estimation we forced the sum-to-zero constraint; when we did not the estimates were unstable and yielded uninterpretable results where teacher effects across grade-levels within a year or across years had negative or very low correlations.

Fixed effects are a generalization of gain scores that use all the students' scores rather than just current and prior year scores. Only students with no prior mathematics achievement scores are excluded from estimation.

As with gain scores, fixed effects methods implicitly assume that scores must be on a scale where differencing scores across grades is sensible. Scores on a truly developmental scale would meet this requirement. In addition, the theory supporting the use of fixed effects requires that the commonality among the multiple scores from a single student depends on a single factor that contributes additively to every score. When this assumption fails to hold, fixed effects estimators can perform less well than multivariate mixed models (Lockwood and McCaffrey, 2007). In the case of z-scores, fixed effects require that differences of rank placement rescaled are meaningful and that a single additive factor explains the correlation in students' rescaled ranks. There is no obvious psychometric justification to support these assumptions so the support for using z-scores in this context will be primarily empirical.

4.2 Comparison and evaluations of performance measures

Table 2a presents various summary statistics for each of the 24 performance measures. These indices provide useful information on how the various measures behave and how they differ but they do not necessarily provide accurate comparisons of the statistical properties of the performance measures. Estimated teacher effects consist of three components: the true teacher performance, error that is correlated with other factors such as student inputs, which we call "bias", and errors that are uncorrelated with other factors, which we call noise or sampling error. To compare the performance measures on the basis of their statistical properties, we ideally would evaluate the contribution of each component to the variability of estimated effects. This would provide a meaningful comparison because the errors in most compensation rules and the consequences of errors in the performance measures depend on the relative shares of total variability attributable to each of these factors. However, we cannot develop indices that uniquely estimate these contributions to the variability of each factor – each measure we consider possibly confounds the effects of at least two factors for the following reasons:

1. We do not know the true level of a teacher's performance.
2. Even though we have a proxy for the true level of performance, the LMT, this measure is imperfect. It is not necessarily scaled on the same scale as teacher performance based on achievement test outcomes and, more importantly, it appears that the skills measured by the LMT do not measure all the skills we might want to capture with a performance measure. The LMT is concentrated on algebraic skills and tends to be lower for teachers teaching special education and lower achieving students, even when our test based performance measures indicate these teachers are high performing.
3. It appears that true teacher performance is correlated with the potential level of achievement for the students in their classes. For instance, the LMT is negatively correlated with the percent minority students in the teacher's class in 2007-07 (-0.28) and positively correlated with the average of the students' average prior z-scores (.36), where the student average prior z-scores equals the average of the mathematics z-scores from years prior to the 2006-07 school year. These correlation coefficients remain large even when we exclude teachers who teach special education or advanced classes such as algebra.
4. Class size is positively correlated with average prior score. This is in part due to special educations classes tending to have smaller numbers of students.
5. Variables such as average prior scores can be correlated with both bias and noise. Hence attempts to use such measures to determine the relative strength of bias can be distorted by the correlation with the noise. For example, because gain scores subtract the prior score from the current score, classroom average gain scores are negatively correlated with classroom average prior year score. Consequently, using prior scores to assess bias can underestimate the bias in gain scores.

Although none of the summary statistics in Table 2aprovide a perfect index for ranking or comparing the measures on bias or noise, the various summary statistics provide indications of the relative performance of the measures. In particular, identifying outliers among the performance measures for any index in Table 2 suggests performance measures with either large bias or noise. Combining this evidence with general statistical theory and previous research offers some conclusions about the relative performance of the measures.

Indicators of Signal

We combine the first set of statistics from Table 2 into a single table because they all provide some information about the strength of the signal in the performance measure. The first column of Table 2a provides the correlation between the performance measures across the two school years. This correlation provides an indication of the consistency of information that the various measures will provide about teachers. There is considerable variation from 0.04 for raw gain scores to almost 0.6 for multivariate ANCOVA on raw scores with shrinkage. This indicates that gain scores have considerable noise relative to the consistent information they provide about teachers. Regression residuals also have a large amount noise relative to the consistent information. Multivariate mixed models, fixed effects with shrinkage and ANCOVA with shrinkage all have high levels of consistent information relative to the noise. Shrinkage increases the correlation by reducing the noise relative to the consistent information about teachers.

Table 2a about here

Consistent information about a teacher can be composed of both accurate information about the teacher or it can be error that is stable across time. For example, if a performance measure conflates differences in student inputs with true teacher effects and student inputs are correlated over time, (i.e., the teacher teaches the same types of students year after year), then this could inflate the correlation coefficient compared to a measure with equal noise and signal but no bias. Consequently a performance measure with greater overall error might actually have a higher cross-year correlation than a measure with less error. The correlation between the 2006-07 and 2005-06 average prior mean z-scores for teachers' classes is about 0.85. Hence, measures that conflate student inputs with true performance could have very high cross-year correlation even if they have large overall errors. However, the cross-year correlation does provide an indication of stability over time in the estimates and clearly compensation systems based on estimates with very low correlation across years (such as gain scores) would be unlikely to be productive.

The next column in Table 2a provides the correlation between the performance measures and the LMT. Given that the LMT and each performance measure estimates a teacher's ability to promote student learning, we would expect moderately high correlation. Moreover, differences between performance measures and LMT might provide information about the contributions of different dimensions of teacher effectiveness on student outcomes. The correlations range from about 0.03 for gain scores up to 0.34 for ANCOVA methods on z-scores with shrinkage. Estimates based on multivariate mixed models are also among the methods with the highest correlation with LMT. Again, this demonstrates limited relative signal for the gain score method. Given that both the performance measures and LMT have measurement error a correlation of 0.34 suggests a strong correlation in the signal from these alternative measures of teacher performance. However, student assignments make students' inputs correlated with teacher signal as measured by LMT. Thus the LMT might be correlated with both the bias and the signal in a teacher performance measure. Given other summary statistics which suggest large bias in the ANCOVA method, we expect the high correlation with LMT in part reflects the contribution of bias to the performance measure.

The next three columns in Table 2a provide information for calibrating the contribution of teachers to student outcomes. These columns provide a means of calibrating the size of teachers' contributions to student learning. Large values indicate variation in teacher performance is somewhat stable across time, since we are using the estimate from the prior cohort, and that teachers are a substantial source of the variability in student scores. Because of possible bias and correlation between true teacher performance and their students' potential to learn, the correlation without controlling for prior scores can be misleading. The effect of controlling for prior scores on the incremental R^2 for teacher effects provides some evidence of the potential contribution of bias in the estimate. Large decreases in the incremental R^2 suggest a performance measure has a large bias that could distort inferences about teachers.

The first of these columns presents the incremental increase to the R^2 resulting from adding the 2005-06 teacher effects estimates to a student-level model for 2006-07 mathematics scores that already included grade-level means. The next column contains the incremental increase to the R^2 resulting from adding the 2005-06 teacher effects estimates to a student-level model for students' 2006-07 mathematics score model that included grade-level means and student prior scores and the third column is the incremental increase to the R^2 from adding estimated teacher effects to a model that already included grade-level means, prior scores and school fixed effects. In general, the estimates all suggest teachers account for about an additional 2 to 3 percent of the variance in scores after controlling for prior achievement. Multivariate ANCOVA, mixed models and fixed effects all provide similar estimates. ANCOVA methods also provide a similar estimate, but the large drop in the incremental R^2 between models with and without prior scores for regression residual, ANCOVA, and lookup tables, suggests these estimates might have more significant bias than the other methods (see further discussion on bias below). Failure to control for student background variables with these methods could lead to significant overstatement of the contributions of teachers to student achievement.

Although not shown in the table, we also estimated the incremental R^2 for adding estimated teacher effects from 2005-06 to a classroom-level model for 2006-07 average scores that included average prior scores. Average prior scores accounted for about 85 percent of the variance in classroom mean scores. Across measures prior year teacher effects tended to account for about 10 percent of the variance in classroom average scores. Estimated teacher effects often account for about 2/3 of the variance not explained by student inputs.

The final column in Table 2a presents the share of the variance among 2006-07 teacher performance measures that is between--rather than within--schools. This statistic provides very little useful information for evaluating the statistical properties of the measures but it does provide information about how effective teachers are distributed among schools. There is considerable variability of estimators for teachers from the same school. Given that noise is within schools and not between them, this statistic might underestimate the schools' share of variance for true teacher performance. However, even if 50 percent of the variance in the estimated performance measures is noise, schools would still only account for about 20 percent of the variance among teachers.

Indicators of Bias

The summary statistics or indices presented in Table 2b were chosen to provide information about the relative contribution of bias to each performance measure. The first column presents the most commonly used indicator of bias, the correlation between the measure and student prior achievement. In our case we use the average of all students' prior year z-scores as a measure of prior achievement. High values are generally considered an indicator of strong bias relative to signal and noise, indicative of the potential for bias to strongly distort performance measures and compensation. Although this correlation can also reflect the correlation between true teacher effects and student inputs (e.g., even an unbiased estimate with little noise would have modest correlation with students' prior achievement), very high values relative to other measures clearly identify performance measures with relatively large bias. In particular, the various ANCOVA based methods all have high correlation with students' prior achievement and we know that the noise in these estimates must be comparable or larger than that of similar methods such as regression residuals or multivariate ANCOVA methods which have substantially smaller correlation with average prior achievement.

Table 2b about here

The large bias in the ANCOVA methods is likely to arise from two sources. First the use of a single prior score to adjust for student background is insufficient to account for student inputs resulting in errors that are positively correlated with high potential for achievement. The estimates based on raw scores have additional error because of the nonlinearity in the bivariate relationship between current and prior scores. The errors are smaller for regression residual methods possibly because the adjustments are greater or possibly because of the greater noise in the estimates. Lookup tables do not assume linearity and have somewhat lower correlation with prior scores than the ANCOVA methods. Multivariate ANCOVA methods with raw scores have high correlation with prior achievement measures because of the violation of the assumed linearity. Gain scores have negative correlation because of the spurious negative correlation between gain scores and prior scores. This negative correlation is inflated by the z-score transformation which forces a negative correlation between prior scores and change in scores by forcing the variance across years to be roughly constant. A similar negative correlation with the noise could be the source of the negative correlation between fixed effects with z-scores and the prior achievement measures resulting from the negative correlation created by the z-score transformation.

As a means of separating the two potential sources of correlation between performance measures and prior achievement, we fit a linear regression model with teacher performance as the outcome variable and percent minority (as a control for inputs) and LMT as predictors. The next two columns of Table 2b present the standardized estimated coefficients for this model for each estimator. By controlling for LMT, we remove some (hopefully much) of the correlation between true teacher performance and student potential for high achievement, so that the coefficient on minority status again indicates the existence of systematic errors. Again regression residuals and ANCOVA method show clear signs of bias. The coefficient on LMT provides a measure on the strength of the true teacher performance signal in the estimator. Gain score methods perform substantially worse than all the other methods on this metric, in part because of the relatively high level of noise in gain scores.

As a means of understanding the potential impact of systematic errors in the estimated performance measure we created two indices. To calculate these indices, we first transformed both the performance measure and the LMT to rank based z-scores and then we differenced these values for every teacher. This difference serves as an estimate of discrepancy between a teacher's performance measured by the performance measure and his or her performance measured by a common standard of the LMT. We used z-score transforms to measure this discrepancy because the performance measures and the LMT are not all on same scale. We then estimated the correlation between the discrepancy (the difference between the transformed performance measure and LMT) and the average mean prior z-score for the teacher's students and the percent of minority students in the teacher's classes. The measures provide an indication of whether errors in the performance measure tend to favor teachers teaching certain students more than others. Strong correlations would tend to suggest estimates with large bias. Again the measures indicate poor performance for regression residuals, ANCOVA, and gain score methods. Fixed effects with z-scores methods also perform poorly on this metric. This results form the weak negative (positive) correlation between teacher performance measures for these methods and prior achievement (minority status). However, it is not clear how to interpret this correlation.

Indicators of Noise

The two columns in Table 2c provide a measure of the relative noise of each performance measure. The first column contains the proportion of teachers with small classes (less than 10 students) ranked in the extreme (top or bottom deciles) of the performance measures for the entire sample of teachers. The second column contains the proportion of teachers with large classes (20 or more students) ranked in the extreme (top or bottom deciles) of the performance measures for the entire sample of teachers (the estimator combines performance measures from 2005-06 and 2006-07). Performance measures with large amounts of noise relative to other sources of variability should place a greater proportion of teachers with small classes in the extreme deciles of the sample and relatively fewer teachers with large classes in the extremes of the distribution. As with other measures of the relative contributions of noise to the various performance measures, gain scores appear to be the noisiest estimates and regression residual based measures perform poorly on this metric with over 50 percent of teachers with small classes in the extreme tails of the sample (when no shrinkage is used). Regression residuals, lookup tables, ANCOVA, and multivariate ANCOVA also perform poorly on this metric. In general shrinkage substantially reduces the noise and the proportion of teachers with small classes with performance measures in the tails of the distribution. The proportion of teachers with small classes in the extreme is particularly small for multivariate mixed model based measures and fixed effects on z-score based measures with shrinkage. This might be an indication that these methods over shrink teachers with small classes or it might indicate that teachers with small classes are less likely to be in the extremes of the distribution.

Implications for Differences in Performance Measures

Although we cannot provide a few statistics to completely rank the relative performance of the alternative measures, we can conclude that some measures appear to perform better than others. To understand the implications of these differences in performance we choose four measures to investigate further:

Gain scores on scale scores without shrinkage
ANCOVA on scale scores without shrinkage
Multivariate mixed models fit to z-scores (naturally include shrinkage)
Fixed effects models fit to z-scores with shrinkage

These four methods cover the domain space of the measures in many dimensions. Gain scores and ANCOVA are simple to implement, transparent, and widely used. In practice many people equate these methods with value-added modeling. Simple gain scores are more widely used by practitioners and many economic applications have used simple ANCOVA-like approaches (for example, Aaronson, Barrow, and Sander, 2003, Kane, Rockoff, and Staiger, 2006). Both these methods are relatively poor performing on one or more of the indices in Table 2 – the gain score measures appear to have most noise relative to other sources of variability in the estimator and the ANCOVA measures appear to have the largest noise component.

Both these methods are highly correlated with other easy to implement and transparent methods like the lookup table and regression residual methods. For example, the correlation between lookup table measures and the ANCOVA method measures is .96 for estimates without shrinkage and .94 for estimates with shrinkage. Performance measures based on regression residuals also are highly correlated with ANCOVA measures (.97 and .96 for estimates with and without shrinkage respectively).

Multivariate mixed models and fixed effects represent the opposite end of the complexity spectrum. Both methods are advocated by methodologists because of their better theoretical properties, but they have not been embraced by practitioners because they are harder to implement and harder to understand. In particular, they are often dismissed as potential performance measures in the context of pay-for-performance because they are deemed too complex and likely to alienate rather than motivate teachers.

Multivariate mixed effects models model the joint vector of a student's scores and implicitly use regression adjustment to account for student prior achievement. As the number of tests becomes large, these methods can yield nearly unbiased estimates of teacher performance even when students in different classes have different potential for high achievement. According to the indices in Table 2, multivariate mixed model measures appear to have small noise but some bias that is clearly relatively smaller than the bias in ANCOVA, regression residuals and lookup tables.

Multivariate ANCOVA methods are similar to multivariate mixed models in that they use regression on multiple prior scores to account for variation in student potential for high achievement. However, whereas the multivariate mixed models use implicit regression to adjust for prior achievement, multivariate ANCOVA uses explicit regression. In our implementation, the multivariate ANCOVA models use more prior scores than in the mixed models, and the multivariate ANCOVA models allow the model to depend on the observed data pattern. Such adjustments are possible with the multivariate mixed models but they increase the computational burden of the model. Moreover, limited exploratory modeling with the current data indicated that including additional test scores did not have a measurable effect on the estimates of teacher performance. In other datasets, we also found that mixed model estimates of teacher effects were robust to more complex models for missing data (Lockwood and McCaffrey, 2005). Hence the additional flexibility in the multivariate ANCOVA methods compared to our current formulation of multivariate mixed models is probably of little consequence. Another difference in the methods is the explicit modeling of prior teacher effects in the mixed models. However, even with all the differences between the methods the correlation between performance measures based on multivariate ANCOVA on z-scores with shrinkage and the estimates based on the variable persistence model is 0.95.

Similarly the variable persistent model provides very similar estimates to those from the layered or complete persistence model (correlation coefficient of 0.98). Hence, the performance measures based on the variable persistence model provide a good proxy for other multivariate estimators using regression-like approaches to control for student prior achievement.

Fixed effects methods also use the full vectors of longitudinal data when estimating teacher performance but do so in a distinctly different manner than the mixed models. Fixed effects methods use the differences between each student's achievement in a given year and his or her average achievement across years to isolate teachers' contributions to student learning. Fixed effect methods also are challenging to compute especially for large samples with large numbers of teachers and students and special care is require to account for the over identification of the model that includes effects for all students and teachers. The model must be parameterized to enforce a sum to zero constraint on the estimated teacher performance measures. If this parameterization is used then estimates are stable across time and appear interpretable. If this parameterization is not used the estimates are not interpretable and can behave erratically across time. Enforcing this sum to zero constraint is particularly challenging when the sample contains large numbers of teachers and students.

Our empirical analyses suggest that fixed effects on z-scores have little bias and small relative noise – for example the cross-year correlation between the performance measures from 2005-06 and 2006-07 is large. Fixed effects on z-scores yield estimates that are negatively correlated with the LMT conditional on student prior achievement. It is not clear if this demonstrates a problem with the performance measure – the only other performance measure to demonstrate this property was gain scores with z-scores which clearly has a negative bias between prior achievement and the student's potential for high achievement resulting from differencing z-scores. However, numerous empirical analyses could not confirm a similar bias for fixed effects.[5]

[5] Gain scores difference of adjacent scores and fixed effects subtract the mean of prior achievement; this could reduce the negative correlation between prior achievement and the differences used in estimating teacher effects. Moreover we found strong positive correlation between gain scores and minority status, even within class but we did not find such correlation between differences in current year and average performance. The negative coefficient for LMT in the model for fixed effects based performance measures might represent limitations of the LMT measure. It focuses on

Any bias in the performance measures based on fixed effects with z-scores seems likely to be small by the various measures in Table 2c and the relative noise in the estimate also appears to be small. Hence fixed effects methods serve as an alternative to the other multivariate methods and also provide estimated performance measures with desirable statistical properties.

These four estimators (gain scores, ANCOVA, mixed models and fixed effects) therefore give us examples of measures with some of the worst statistical properties but which are simple to implement and measures with clearly better properties but which are complex to estimate and comprehend for teachers and others. On the basis of our comparisons so far, it is not clear what implications the differences among the measures might have for evaluating teachers. To explore the potential implications, we compare each of the performance measures to all of others (6 pairwise comparisons overall) by comparing the teachers whose performance each method finds as significantly better than average (zero) using standard teacher by teacher t-tests (e.g., the teacher is significantly better than average if the lower limit of a 95% confidence interval exceeds zero[6]). Although the performance measures can be used in many other ways (see the discussion below for alternative decision rules), this comparison is concrete and provides straightforward data on the measures. It is clear from our findings and other explorations we conducted that the results would be similar if we used alternative classification schemes and that these results also provide a general picture of how the performance measures differentially value teachers teaching different types of students.

Table 3 provides a summary of the teachers whose performance is classified as significantly better than average by each of the four selected performance measures. The table presents the percentage of teachers in different groups that receive this classification and the Kappa statistic (Fleiss, 1981) as a measure of agreement in the classification across the two school years. The Kappa statistic equals the percent of agreement in classification adjusted for chance agreement.

Some general patterns are clear from the table. First the bias identified in the ANCOVA methods results in more teachers whose performance is classified as significantly better than average with this method than with any of the others. As would be expected given the bias, the ANCOVA strongly favors teachers teaching higher achieving students. Only 6 percent of teachers teaching classes with average prior achievement less than one standard deviation below the mean of all classes receive this designation, whereas 70

algebraic skills that might not capture positive teacher performance for teachers teaching lower achieving students such as minority or special education students.

[6] For the multivariate mixed models estimates the test is formally based on a credible interval since they use the posterior mean and standard deviation.

percent of teachers teaching classes with the highest average prior achievement receive this designation. Substantially more teachers in the group teaching the highest performing students are designated above average with the ANCOVA method than with any of the other performance measures. It appears that the ANCOVA method identifies additional teachers of high performing students as performing above average that other methods do not. A related consequence of the bias in this measure is the high proportion of teachers who teach only advanced classes being designated as above average performers. This is a small group of teachers but nearly all of them would be designated as superior by this method whereas other methods might identify as few as 13 percent of these teachers as superior. Because teachers' course assignments are stable across time, ANCOVA would consistently favor teachers of advanced classes and higher achieving students across years (Kappa is 0.55).

Performance measures based on gain scores behave in a manner opposite to those based on the ANCOVA measure. The relatively large noise results in a very low Kappa (0.24) and large proportions of teachers with small classes being classified as above average performers. The tendency of gain scores to be negatively correlated with students' prior achievement results in roughly equal percentages of teachers being classified as above average performers regardless of their students' prior achievement. Also the proportion of special education teachers classified as superior performers is high (17 percent) compared to all the other measures.

Performance measures based on fixed effects for z-scores with shrinkage identify similar types of teachers as above average performers as gain scores. However, the fixed effects measure is more stable across years with a Kappa of 0.50. Performance measures based on mixed models tend to favor teachers of students with high prior achievement but not to the extent that the ANCOVA method does. Moreover, it does not falsely identify the large number of teachers that ANCOVA does. For example, the performance of only 40 percent of teachers teaching students with the highest prior achievement is designated as significantly better than average compared to ANCOVA. However, 63 percent of teachers teaching only advanced classes are designated as superior performers.

Direct comparisons of the methods reflect these overall trends. As shown in Table 4, performance measures based on fixed effects generally favor teachers of lower achieving students than the measures based on mixed models or ANCOVA as shown by the differences among the teachers where the two methods disagree. No such differences exist with gain scores and fixed effects. However, fixed effect and mixed model methods are highly correlated and disagree on just 10 percent of teachers when the data from the two school years are combined. Moreover the differences between these two methods are not persistent. Very few of the teachers where the methods disagree in 2005-06 have the same disagreement 2006-07. At the margins the two methods favor slightly different teachers but due to noise and true year to year variability in performance the individual teachers at the margins are not the same every year. Thus a system with fixed effects will tend to favor teachers of lower performing classes but it will not treat individual teachers differently than a system based on the mixed model performance measures.

Fixed effects and ANCOVA methods disagree on 20 percent of cases and the discrepancies between the two measures persist across years. Over 40 percent of the teachers classified as above average by the ANCOVA measures but not the fixed effects measures had the same difference in classifications in both school years. Hence, using ANCOVA-based measures would tend to systematically favor individual teachers of low risk students compared to students of high risk students as well as creating a system that favored such teachers.

The differences between mixed model measures and ANCOVA measures are similar to but less dramatic than the differences between fixed effect measures and ANCOVA measures. The differences between mixed model measures and gains score measures are similar to the differences between mixed models measures and fixed effect measures. The differences between gain score measures and ANCOVA measures are also similar to the differences between fixed effect measures and ANCOVA measures.

It is difficult to determine if the difference between fixed effects and mixed models reflects bias in the fixed effect based measure or the mixed models based measure. As discussed previously, some of the measures in Table 2 suggest mixed model measures might be biased in favor of teachers teaching students with higher prior achievement. Alternatively, fixed effect measures show patterns that are similar gain score measures, which we suspect are biased in favor of teaching students at risk for low achievement. Table 4 shows that when the two measures differ the fixed effects methods tend to identify teachers with lower LMT scores as above average performers than the mixed model measures. This might indicate a bias in the fixed effects method or as noted above it might indicate a limitation in the LMT measure. Hence, we cannot clearly conclude that either method is necessarily superior to the other. However, we can conclude that the methods provide very similar inferences and would lead to a similar bonus decision under most rules. A system based on fixed effects measures would tend to be more favorable to teachers of students at risk for low performance and should tend to err for the better if not the best performing teachers in this group. A system based on mixed model measures would tend to be more favorable to teachers of students at low risk for low performance and again would tend to over reward teachers near but not at the top of performers in this group. It would also tend to under reward high performing teachers teaching at risk students. Given the struggle to staff schools with high poverty and high minority populations with the best teachers, as a policy choice fixed effects measures are preferred if we cannot rule out potential bias in each approach because fixed effects methods should be less likely to discourage good teachers from teaching at risk students. However, as discussed in the next section, choices about bonus systems are complex and involve many factors and many potential tradeoffs.

5. AWARDING TEACHER BONUSES

The goal of a compensation system is not to create performance measures but to use those performance measures to compensate teachers. This requires creating a decision rule for awarding bonuses on the basis of performance measures. Again, many choices need to be made when creating a decision rule for assigning awards. To date there has been little specific consideration of what those choices are or what criteria can be used to make these choices. In this section we discuss a framework for considering those choices and explore implications for teachers of using alternative rules with performance measures with different levels of noise and bias.

Performance-based compensation systems are proposed as a means of motivating teachers to higher levels of performance and enticing better teachers to join and remain in the profession (Springer and Podgursky, 2007, Buddin et al, 2007). Various aspects of a compensation system might influence teacher behavior. Standard results (cite) suggest that the expected bonus for a given level of effort will influence teachers' decisions about the level of investment in their performance and changes in the expected value of the bonus for a change in performance and effort will influence teachers' decisions about whether to increase or decrease their efforts.[7] However, other factors such as the uncertainty in the bonus amount, uncertainty about the relationship between effort and performance, and perceptions of fairness of the system and the performance measure might also contribute to teachers' decisions. Noise and bias in the performance measure might affect any of these factors. The decision rule might also affect the distribution of possible bonuses and can interact with the features of the performance measure.

Figure 1 demonstrates the complex relationship between performance measures and eight decision rules in determining the expected compensation for teachers. The decision rules are:

1. Receive the full bonus if the performance measure exceeds a threshold (in this case the 80th percentile of an assumed reference distribution for true teacher performance—a standard normal distribution with variance equal to the (estimated) variability in true teacher performance);

[7] Increasing the expected bonus by simply increasing the bonus amount given any decision rule will not discriminate among the alternative rules for awarding compensation and would need to be weighed against the overall cost the system can bear, which we do not know. Hence, we assume that the maximum bonus value is 1 under any system and study choices among alternative award procedures under this constraint.

2. Receive the full bonus if the lower limit of the confidence interval for the performance measure exceeds the threshold;
3. Receive the full bonus if the lower limit of the confidence interval for the performance measure exceeds zero, where the performance of the average teacher equals zero;
4. The bonus equals the posterior probability that the performance measure exceeds the threshold, with an noninformative prior;
5. Receive the full bonus if the shrunken performance measure exceeds the threshold;
6. Receive the full bonus if the lower limit of the confidence interval for the shrunken performance measure exceeds the threshold;
7. Receive the full bonus if the lower limit of the confidence interval for the shrunken performance measure exceeds zero; and
8. The bonus equals the posterior probability that the performance measure exceeds the threshold, with an informative prior for the performance (in this case the prior that corresponds to using the shrinkage estimator).

Figure 1 shows the expected bonus for three hypothetical unbiased performance measures based on the ratio of the noise to the true variability of teachers; noise levels are one percent (solid line), 10 percent (dashed line), and 25 percent (dotted line). We might consider these different performance measures with different noise levels or the same performance measure for teachers with differing class sizes. If we knew the true performance, the expected bonus would be a step function at true performance equals 0.84.

For performance measures with little noise, the expected compensation resembles that which would occur if we knew true performance. As the noise increases, expected bonus decreases for teachers who are truly above the threshold and increases for those who are below it. Thus, if the threshold was chosen to have a desired effect on performance, then noise in the performance will weaken that effect by over- or under-compensating teachers.

Different decision rules have substantial effects on the expected bonus. When the measure is noisy, requiring the teacher to be significantly greater than the threshold shifts the location of the error away from teachers who do not truly exceed the threshold to exclusively those teachers who do. This reflects the fact that requiring teachers to be significantly above the threshold is the classification rule that minimizes errors from rewarding of teachers who are not truly above the of threshold (Type I errors in standard hypothesis testing) for a given level of failure to reward teachers who are truly above the threshold (Type II) errors. Using a 95 percent confidence interval when making decisions, corresponds to valuing the cost of a Type I error 19 times more than the cost of making a Type II error. Consequently, systems that use this rule make almost no Type I errors. Rewarding teachers when the performance measure alone exceeds the threshold (i.e., decision rule 1) is optimal if Type I and Type II errors are equally costly.

Using shrinkage also places greater value on Type I errors. The shift in the bonus curves is less pronounced than using confidence interval methods, but the method clearly reduces the number of Type I errors--a cost of greater Type II errors. Systems often focus on teachers who are significantly greater than average (zero in this case). If this is a proxy for identifying teachers who truly exceed a threshold above the average, then the approach introduces Type I errors and reduces Type II errors relative to other methods that explicitly control for noise in the estimated performance measures (shrinkage estimates and requiring the estimates to be significantly above the threshold).

It is not obvious how these two types of errors in classifying teachers or rewarding compensation should be valued. We are unaware of research that addresses the economic and psychological responses to different types of errors in this context. Repeatedly failing to award a teacher who is truly exceptional might be discouraging and lead him or her to leave teaching or apply less effort. Conversely, repeatedly rewarding poor performing teachers might make them complacent and fail to motivate them to improve or leave the profession. How strong each signal will be is unknown at this time. However, we do know that more teachers are at risk for Type I errors since only a small fraction of teachers' performances truly exceed the threshold.

Paying teachers continuously on the basis of the probability that their performance truly exceeds the threshold is the payout that minimizes the average squared difference between true compensation and the compensation we would award if we knew true performance. Hence it surprising that this method does not appear superior in terms of expected bonuses to just rewarding teachers if their performance measure exceeds the threshold. However, if we look at the variability in possible bonus awards (Figure 2), this measure greatly reduces the uncertainty compared to all the other methods. This method reduces uncertainty by removing the discreteness in the payouts; with this method a small error in the performance measure can never lead to a complete loss in the bonus. However, removing the uncertainty might remove the incentive for teachers to improve performance because every year they receive a relatively stable bonus. With other decision rules the only way to reduce uncertainty and retain a high bonus is to greatly improve performance.

The discussion s to this point has assumed unbiased performance measures. Bias will not change the results but will change the location within the distribution of performance where the results apply, effectively treating the true performance plus bias as the true performance in terms of various errors. Of particular concern, bias makes errors in the bonuses differential among teachers of equal quality but teaching different types of students. Figure 3 demonstrates this effect. We assumed that, conditional on the teachers' LMT scores, the relationship between percent minority students and ANCOVA-based performance measures on raw scores without shrinkage described the bias in the estimate as the function of the percent minority students in the teacher's classes. We then used this model to calculate the expected bonus across all teachers at different levels of percent minority classes ranging form zero to one. Figure 3 displays the results. The expected bonus drops precipitously as the percent of minority students in the classes increases, dropping by nearly 50 percentage points as classes increase from zero to 100 percent minority students. Clearly, if receiving a bonus is motivational, teachers with large percentages of minority students in their classes will lack motivation. Also, at such low expected values of the bonus, there would be little uncertainty in the bonus as well. Hence large bias could be very disruptive to a performance-based pay system.

The results presented here do not clearly suggest one method for awarding bonuses. They identify different types of errors that need to be evaluated and considered in the context of awarding compensation to teachers. We need to understand the effects of overpaying rather than underpaying teachers. We also need to understand the relative value of certainty versus higher expected payout. We also need to understand how teachers will use bonus awards and data on performance to evaluate their expected bonuses under alternative investments in their teaching.

6. CONCLUSIONS

This paper discusses the complex process of rewarding teacher bonuses on the basis of student achievement data from administrative databases. The first stage of the process involves preparing the administrative data to support the estimation of teacher performance measures. The teachers to be included in the sample must be identified. These must be teachers for whom the evaluated subjects (i.e., mathematics or reading) and grade levels constitute a sufficiently large portion of their job responsibility to be used in determining pay. Also, the teachers in the sample must be appropriate to function as the reference group for evaluating other teachers' performance. Most value-added estimates of teacher performance implicitly estimate a teacher's performance relative to the average of the pool of teachers included in the analysis sample. Hence, a decision to include a teacher or group of teachers (e.g., special education teachers) establishes the reference for measuring the performance of all teachers in the group.

The students whose performance for which each teacher will be held accountable must also determined. In particular, the decision must be made as to the number of days a student must be in a teacher's classroom for his or her learning to be reasonably attributable to the teacher. Identifying students who do not "count" toward performance might result in negative incentives where teachers focus attention away from these students toward other students. The greater the number of days in a classroom required for a student to be used in performance measurement, the greater the risk for negative incentives and negative consequences for a student. Alternatively, using students with very little time in the classroom might undermine the credibility of the performance measurement system and any cutoffs must find a balance between these competing demands.

Another challenge is modeling the data from students linked to multiple teachers for the same subject during the school year. In our case study, we could not determine if enrollment of a student in multiple courses at the same school implied the student was enrolled in the courses sequentially or simultaneously. This indeterminacy makes assigning the proportion of instruction provided by each teacher difficult. We suspect that many administrative data systems will not provide accurate data on course enrollment throughout the school year and more detailed data may be required to support estimating teacher performance.

We looked at numerous methods for estimating teacher performance. Each of the estimates is potentially biased so that teachers of equal quality but teaching different types of students will systematically have different estimates of performance, (e.g., a performance measure will tend to be too high for teachers of students who have high prior achievement). Moreover, we cannot rule out the possibility that every performance measure is either biased or incomplete. Hence, while we expect that the measures are biased, we cannot estimate the contribution of bias to the variability in any measure. Similarly, all the measures contain sampling error or noise, but we cannot fully characterize the size of the noise in comparable ways across methods because of the potential differences in bias and alternative scalings of performance that exist across the measures.

However, by comparing measures on several indices, it is clear to us that some performance measures are outliers on one or more indices in ways that are consistent with the measure having relatively large bias or noise. For example, there is a strong correlation between students' prior achievement or student demographics and the ANCOVA-based performance measures. Similar results hold for related methods that also use a single prior score to predict student performance and deviations for this predicted performance to measure teachers. On the basis of this empirical result and statistical theory, it is clear these methods for estimating performance measures yield relatively large bias and tend to favor teachers teaching students who are at low risk for low achievement.

Gain score methods tend to produce relatively noisier estimates than other methods, as demonstrated by the substantially larger percentage of teachers with small classes whose performance is ranked in the top or bottom deciles of the distribution using this method compared to others. Estimates with relatively greater noise will tend to exaggerate the impact of differential class sizes on the noise in estimated performance. Hence, the finding that over 50 percent of teachers with classes of less than 10 children rank in the top or bottom deciles of the distribution, whereas about 20 percent or fewer teachers of small classes rank in the extremes for many measures, strongly suggests the presence of large noise in the gain scores methods without shrinkage.

Gain scores are negatively correlated with students' prior achievement and demographic variables generally associated with higher scores, even within classes. As a result, the performance measure based on gain scores on raw and z-scores appears to show bias that favors teachers of students at greater risk for low achievement. That is, for two teachers of true equal performance level, the teacher teaching students with low prior performance or with a greater percentage of minority students would tend to score higher on gain score performance measures. This feature of the performance measure tends to decrease the overall variability of estimated performance measures for this method compared to what they would be if bias did not exist. Bias in other methods that favors teachers in low risk classes tends to inflate variability relative to how they would perform without bias.

Multivariate mixed model methods and fixed effects methods with shrinkage tend to provide estimates that appear to have relatively less noise and relatively less bias. Performance measures from both methods tend to have strong cross-year correlation within teacher, weak correlation with students' prior achievement, and relatively few teachers with small classes ranked in the extremes of the sample. The two methods have strong positive correlation (0.86) and agree on the classification of teachers as performing significantly above average or not on nearly 90 percent of teachers. However, the multivariate mixed models do appear somewhat to favor teachers teaching students with higher prior performance and fixed effects based on z-scores does appear to have a small bias toward teachers of classes with greater risk for low achievement. The bias in the fixed effects method is demonstrated in part by a negative partial correlation with the LMT, controlling for the classroom average prior achievement, and the fact that when this method identifies teachers as above average and the mixed model method does not, those teachers tend to teach classes with greater percent minority students and the teachers score lower on the LMT than others. However, the LMT as implemented might not fully measure teaching skills related to more basic mathematics and might have a bias of its own. Hence, the size of the bias in fixed effects and mixed models is somewhat difficult to fully determine; however, the empirical evidence suggests it is likely to be small for both measures.

Noise and bias can distort decision rules for awarding bonuses. Using performance measures with greater noise to award bonuses can reduce the expected bonuses for the truly best-performing teachers and increase the expected bonuses for the truly worst-performing teachers. Using performance measures with bias to award bonuses makes the expected bonuses for teachers of equal quality who teach in different contexts very different. For example, using ANCOVA-based measures based on raw scale scores without shrinkage would result in bonuses for 57 percent of the teachers teaching the classes of students with the highest average prior scores. Using fixed effects-based measures with z-scores and shrinkage would result in just 20 percent of these teachers receiving a bonus.

Alternative decision rules correspond to valuing different types of errors differentially. Methods that require estimates to be significantly over a threshold or use shrinkage correspond to valuing errors that mistakenly reward low-performing teachers as significantly more costly than errors that mistakenly fail to reward high-performing teachers. Using statistical significance testing at the 0.05-level to determine if a teacher's performance exceeds a threshold and he or she receives a bonus is appropriate if incorrectly rewarding undeserving teachers is 19 times more costly than failing to reward deserving teachers. However, it is unclear how teachers respond to rewards, so there seems to be little reason to value these two types of error so differently. If a threshold is chosen for a truly meaningful purpose, there is probably little current support for only rewarding teachers if their performance is significantly greater than the threshold using the traditional 0.05 level of significance.

Creating a performance-based pay system is challenging. Administrative data required to create estimates must be cleaned and processed before the performance measures can be estimated. Various alternative performance measures and rules assigning bonuses on the basis of those measures exist. Different measures and rules will lead to different teachers consistently receiving additional compensation. The choices that can be made during the process can directly affect whether or not some teachers receive bonuses. However, little is known about how these differences will affect teachers' behaviors and the overall quality of the teacher labor force.

To determine the best choices to yield truly efficient systems, we need to conduct research that provides information about how teachers will respond to bonus systems. We need to determine how teachers with various true levels of performance and teaching in different contexts will respond to different expected levels of bonus and different levels of uncertainty in the bonus. We need to study how these responses interact with other features of performance measures and bonus decision rules. For example, will teachers' decisions on how to respond to the payout distribution depend on how statistically complex the performance measures are or how transparent the calculations are? Will the responses of teachers depend on how the distribution of bonuses varies across teachers of different types of students? Will teachers be less likely to make changes in behavior in response to a system that consistently tends to reward teachers of minority students with less money? Because we clearly understand the potential payout of a bonus system based on each alternative performance measure and bonus decision rule, we can identify the features that could influence performance and elicit teachers' likely responses to those features. This information can then become the basis for designing future systems.

APPENDIX – STATISTICAL DETAILS

This appendix provides additional statistical details on the z-score transformation and the performance estimators.

Rank-based z-score transformation

For the base year, 2003-04 then rank based z-score equals the raw scale score transformed to the percentile of the empirical distribution function and then transformed to the corresponding quantile of the standard normal distribution by subject and grade level. For a given subject and grade-level let y denote raw score and $r(y)$ denote the rank of that score among all other scores for that subject and grade-level, then the z-score transform of y is

$$z = \Phi^{-1}(r(y)/(n+1)), \tag{A.1}$$

where Φ^{-1} is the inverse of the standard normal cumulative distribution function and n is the number of students with test scores for the given subject and grade-level. The resulting z-scores have mean zero and standard deviation one. For years other than the baseline we allow the mean and standard deviation of the z-scores to drift. We assume that each raw score is a monotonic nonlinear transformation (rank preserving) of a normally distributed variable. We assume that the nonlinear transformation is constant across year but that the distribution of underlying normal variables can change across year. Given that all the data is assumed normally distributed, we can assume that if the y in year 1 corresponds to the pth percentile then its value in the underlying normal distribution is z_p where z_p is the corresponding quantile of the standard normal distribution. For a score of y_2 at the pth percentile of the distribution of scores in year 2, any year other than the baseline, has corresponding normal value of $z_p\sigma_2 + \mu_2$, where μ_2 and σ_2 are the mean and standard deviation for the underlying normal distribution for the year. This value corresponds to $z_{p'}$ of the p' quantile of the standard. Because we assume the nonlinear function relating the normal variates to the scale scores is constant, the scale score y_2 should correspond to the p' percentile of the baseline population. Hence using the function A.1 on y_2 will yield $z_{p'}$. Because we know the score's percentile of the year 2 distribution we can solve forσ_2 and μ_2 if we repeat the procedure for multiple scores from year.

Regression Residuals

The regression models used for the regression residuals included separate intercepts and slope parameters for prior achievement by grade-level to allow for differences in the overall level of scores by grade-level and for the possibility that the relationship between current and prior achievement might differ by grade-levels.

To obtain the shrinkage estimator we used the regression residuals from fitting the linear regression models as outcome variables and fit a one-way random effects ANOVA model. The Best Linear Unbiased Predictors from this model yield shrunken estimates. This procedure ignores the sampling error in the estimated regression coefficients but with large numbers of students this should have a minimal effect on the estimates. Standard error estimates for both the shrunk and unshrunk estimators ignore the sampling error in estimating the model slope and intercept.

ANCOVA

The regression model includes separate intercept and slope parameters by grade-level. To enforce the sum to zero constraint, we subtract the intercept and the slope times the prior mathematics achievement score from each student's 2006-07 achievement score to obtain a residual. We averaged these to obtain the unshrunk performance measure and used these as outcomes in a random effects one-away ANOVA model to obtain the shrunken estimates. Standard error estimates for both the shrunk and unshrunk estimators ignore the sampling error in estimating the model slope and intercept.

Lookup tables

To estimate shrunk estimates we used the deviations for the expected scores as outcomes in a one-way random effects ANOVA. Standard error estimates for both the shrunk and unshrunk estimators ignore the sampling error in estimating expected scores.

Fixed effects

To create shrunken estimates we estimated the variability among true teacher effects using simple method of moments or meta-analytic methods. We estimated the variability among teachers with 10 or more students and subtracted from this the average of the squared standard errors for teachers' raw estimates to obtain the desired variance component, v^2. We restricted to teachers of 10 or more students to improve the stability of our estimates. The shrinkage factor for each teacher equals $v^2/(v^2 + SE^2)$, where SE denotes the standard error of the raw estimate for the teacher.

Table 1. Classification of Performance Measures[a]

Method	Scale of Student Achievement Measure	Statistical Adjustments
Deviations from Expectation Estimated with a Single Score	Raw scores	With shrinkage
ANCOVA	z-scores	Without shrinkage
Regression Residuals		
Lookup-Tables		
Gain Score		
Average gain scores		
Deviations from Expectation Estimated with Multiple Scores		
Multivariate ANCOVA		
Multivariate Mixed Effects Models		
Variable persistence model		
Layered (complete persistence) model		
Fixed Effects		
Student fixed effects for achievement scores		

[a] Performance measures were estimated using complete cross of method, scale of student achievement measure, and statistical adjustment for all methods except the two multivariate mixed effects model methods which were applied only to z-scores and only with shrinkage and lookup-tables which used only raw scores.

Table 2a. Summary Statistics for Performance Measures (Indicators of Signal)

Method	Raw or Shrunk	Scale	Cross Year Correlation	Correlation with LMT	R^2 for Prior year Teacher Effect(no covariates)	R^2 for Prior year Teacher Effect (controlling for prior scores)	R^2 for Prior year Teacher Effect (controlling for prior scores and schools)	Proportion of variance between schools
Regression Residuals	raw	raw score	0.17	0.24	0.11	0.02	0.01	0.04
Regression Residuals	shrunk	raw score	0.43	0.27	0.13	0.02	0.01	0.08
Regression Residuals	raw	z-score	0.32	0.29	0.11	0.02	0.01	0.07
Regression Residuals	shrunk	z-score	0.51	0.30	0.12	0.02	0.02	0.11
ANCOVA	raw	raw score	0.32	0.30	0.19	0.02	0.01	0.06
ANCOVA	shrunk	raw score	0.52	0.33	0.21	0.02	0.01	0.11
ANCOVA	raw	z-score	0.45	0.33	0.18	0.02	0.02	0.09
ANCOVA	shrunk	z-score	0.58	0.34	0.19	0.02	0.02	0.14
Lookup tables	raw	raw score	0.25	0.24	0.10	0.02	0.01	0.04
Lookup tables	shrunk	raw score	0.46	0.28	0.12	0.02	0.01	0.07
Gain Score	raw	raw score	0.04	0.08	0.03	0.02	0.01	0.02
Gain Score	shrunk	raw score	0.33	0.12	0.04	0.02	0.01	0.05
Gain Score	raw	z-score	0.16	0.07	0.01	0.02	0.01	0.04
Gain Score	shrunk	z-score	0.39	0.07	0.01	0.02	0.01	0.06
Multivariate ANCOVA	raw	raw score	0.39	0.21	0.13	0.03	0.02	0.05
Multivariate ANCOVA	shrunk	raw score	0.57	0.26	0.15	0.03	0.02	0.09
Multivariate ANCOVA	raw	z-score	0.36	0.21	0.09	0.03	0.02	0.07
Multivariate ANCOVA	shrunk	z-score	0.54	0.25	0.10	0.03	0.02	0.09
Mixed Models (Variable Persistence)	shrunk	z-score	0.51	0.32	0.13	0.03	0.02	0.11
Mixed Models (Layered Model)	shrunk	z-score	0.51	0.31	0.09	0.03	0.02	0.07
Fixed Effects	raw	raw score	0.28	0.18	0.10	0.03	0.02	0.05
Fixed Effects	shrunk	raw score	0.51	0.24	0.12	0.03	0.02	0.09
Fixed Effects	raw	z-score	0.36	0.11	0.02	0.03	0.02	0.08
Fixed Effects	shrunk	z-score	0.54	0.17	0.02	0.03	0.02	0.10

Table 2b. Summary Statistics for Performance Measures (Indicators of Bias)

						Systematic Error Index	
Method	Raw or Shrunk	Scale	Correlation with Average Mean Prior Z-score	Standardized Regression Coefficient for Percent Minority	Standardized Regression Coefficient for LMT	With average mean prior z-score	With percent minority
Regression Residuals	raw	raw score	0.39	-0.29	0.18	0.17	-0.05
Regression Residuals	shrunk	raw score	0.39	-0.26	0.21	0.18	-0.05
Regression Residuals	raw	z-score	0.33	-0.26	0.22	0.14	-0.05
Regression Residuals	shrunk	z-score	0.31	-0.24	0.23	0.12	-0.03
ANCOVA	raw	raw score	0.49	-0.35	0.27	0.33	-0.16
ANCOVA	shrunk	raw score	0.51	-0.31	0.29	0.31	-0.13
ANCOVA	raw	z-score	0.47	-0.34	0.26	0.29	-0.13
ANCOVA	shrunk	z-score	0.46	-0.31	0.28	0.27	-0.11
Lookup tables	raw	raw score	0.32	-0.19	0.19	0.10	0.00
Lookup tables	shrunk	raw score	0.29	-0.14	0.22	0.09	0.01
Gain Score	raw	raw score	0.10	-0.06	0.02	-0.17	0.13
Gain Score	shrunk	raw score	0.03	-0.05	0.06	-0.15	0.14
Gain Score	raw	z-score	-0.07	0.04	0.03	-0.28	0.18
Gain Score	shrunk	z-score	-0.12	0.04	0.04	-0.27	0.17
Multivariate ANCOVA	raw	raw score	0.42	-0.28	0.16	0.26	-0.04
Multivariate ANCOVA	shrunk	raw score	0.45	-0.28	0.22	0.27	-0.04
Multivariate ANCOVA	raw	z-score	0.25	-0.20	0.15	0.08	0.03
Multivariate ANCOVA	shrunk	z-score	0.21	-0.19	0.19	0.06	0.04
Mixed Models (Variable Persistence)	shrunk	z-score	0.14	-0.17	0.27	0.03	0.02
Mixed Models (Layered Model)	shrunk	z-score	0.09	-0.09	0.27	0.00	0.05
Fixed Effects	raw	raw score	0.21	-0.16	0.12	0.02	0.03
Fixed Effects	shrunk	raw score	0.21	-0.17	0.20	0.04	0.03
Fixed Effects	raw	z-score	-0.11	0.03	0.06	-0.31	0.22
Fixed Effects	shrunk	z-score	-0.12	0.01	0.13	-0.25	0.18

Table 2c. Summary Statistics for Performance Measures (Indicators of Noise)

Method	Raw or Shrunk	Scale of Test	Percent of teachers with small classes in extreme deciles[a]	Percent of teachers with large classes in extreme deciles[b]
Regression Residuals	raw	scale score	49.28	10.23
Regression Residuals	shrunk	scale score	26.44	17.62
Regression Residuals	raw	z-score	38.76	20.51
Regression Residuals	shrunk	z-score	22.44	20.63
ANCOVA	raw	scale score	44.71	13.02
ANCOVA	shrunk	scale score	29.81	16.67
ANCOVA	raw	z-score	42.03	13.69
ANCOVA	shrunk	z-score	23.56	18.76
Lookup tables	raw	scale score	45.89	11.2
Lookup tables	shrunk	scale score	25.24	17.67
Gain Score	raw	scale score	52.15	10.2
Gain Score	shrunk	scale score	27.88	18.59
Gain Score	raw	z-score	43.69	12.58
Gain Score	shrunk	z-score	19.32	21.65
Multivariate ANCOVA	raw	scale score	43.27	10.98
Multivariate ANCOVA	shrunk	scale score	26.44	15.69
Multivariate ANCOVA	raw	z-score	39.23	13.31
Multivariate ANCOVA	shrunk	z-score	20.87	20.3
Mixed Models (Variable Persistence)	shrunk	z-score	13.46	22.37
Mixed Models (Layered Model)	shrunk	z-score	12.44	22.2
Fixed Effects	raw	scale score	46.41	12.19
Fixed Effects	shrunk	scale score	20.67	20.81
Fixed Effects	raw	z-score	46.41	13.38
Fixed Effects	shrunk	z-score	13.59	23.46

[a] Small classes have less than 10 students

[b] Large classes have 20 or more students

Table 3. Summary of Teachers Whose Performance is Classified as Significantly Better by Four Selected Performance Measures

	ANCOVA	Gains	Mixed Models	Fixed Effects
	Percent Classified Significantly Better than Average			
All Teachers	29%	20%	19%	19%
Teachers by Class Size				
Less than 10 students	7	14	4	11
10 to 19 students	12	17	13	15
20 or more students	44	23	27	23
Teachers by Average Student Prior Achievement				
Less than one std deviation below mean	6	18	7	15
Between one standard deviation below and the mean	23	21	18	20
Between the mean and one standard deviation above	42	19	24	21
More than one standard deviation above the mean	70	23	40	20
Teachers by Course Taught				
Only general education courses[a]	36	21	23	22
Only special education courses	5	17	5	12
Only advanced courses[b]	88	13	63	13
	Kappa Statistic			
Cross Year Agreement	0.55	0.24	0.53	0.50

[a] Excludes teachers who taught special education, algebra or other advanced mathematics courses

[b] Includes teachers who taught only algebra or other advanced mathematics courses

Table 4. Pairwise Comparison of Four Performance Measures on the Basis of Teachers Classified as Significantly Better than Average

Methods	**Pearson Correlation Coefficient**		**Cross Tabulation of Classification as Significantly Above Average by Two Methods** Method A: Fixed Effects			**Classified as above average by Method A** **Classified as above average by Method B**	**Yes** **No**	**No** **Yes**
A. Fixed Effects **B. Multivariate Mixed Model**	0.86	Method B: Mixed Models	Significantly Above Average	Other	Total	Average Percent Minority Student	***0.70***	***0.46***
		Significantly. Above Average	135 14%	45 5%	180 19%	Average Prior Achievement	***-0.77***	***0.72***
		Other	45 5%	729 76%	774 81%	Percent Special Education Teachers	***0.40***	***0.02***
		Total	180 19%	774 81%		Average LMT Score	-0.44	0.43

Methods	**Pearson Correlation Coefficient**		**Cross Tabulation of Classification as Significantly Above Average by Two Methods** Method A: Fixed Effects			**Classified as above average by Method A** **Classified as above average by Method B**	**Yes** **No**	**No** **Yes**
A. Fixed Effects **B. Gain Scores**	0.59	Method B: Gain Scores	Significantly Above Average	Other	Total	Average Percent Minority Student	0.67	0.65
		Significantly. Above Average	122 13%	58 6%	170 18%	Average Prior Achievement	-0.12	-0.27
		Other	58 6%	721 76%	779 82%	Percent Special Education Teachers	***0.12***	***0.33***
		Total	180 19%	769 81%		Average LMT Score	0.17	0.16

Methods	**Pearson Correlation Coefficient**		**Cross Tabulation of Classification as Significantly Above Average by Two Methods** Method A: Fixed Effects			**Classified as above average by Method A** **Classified as above average by Method B**	**Yes** **No**	**No** **Yes**
		Method B: ANCOVA	Significantly Above Average	Other	Total	Average Percent Minority Student	***0.72***	***0.50***
A. Fixed Effects	0.51	Significantly. Above Average	132 13%	129 14%	261 27%	Average Prior Achievement	***-0.74***	***0.56***
B. ANCOVA		Other	48 5.%	640 67%	688 73%	Percent Special Education Teachers	***0.33***	***0.02***
		Total	180 19%	769 81%		Average LMT Score	***-0.30***	***0.34***

Methods	**Pearson Correlation Coefficient**		**Cross Tabulation of Classification as Significantly Above Average by Two Methods** Method A: Mixed Models			**Classified as above average by Method A** **Classified as above average by Method B**	**Yes** **No**	**No** **Yes**
		Method B: Gain Scores	Significantly Above Average	Other	Total	Average Percent Minority Student	***0.53***	***0.69***
A. Multivariate Mixed Models	0.66	Significantly. Above Average	120 13%	50 5%	170 18%	Average Prior Achievement	***0.52***	***-0.81***
B. Gain Scores		Other	59 6%	722 76%	781 82%	Percent Special Education Teachers	***0.00***	***0.52***
		Total	179 19%	772 81%		Average LMT Score	***0.42***	***-0.41***

Methods	**Pearson Correlation Coefficient**		**Cross Tabulation of Classification as Significantly Above Average by Two Methods** Method A: Mixed Models			**Classified as above average by Method A** **Classified as above average by Method B**	**Yes** **No**	**No** **Yes**
		Method B: ANCOVA	Significantly Above Average	Other	Total	Average Percent Minority Student	***0.71***	***0.53***
A. Multivariate Mixed Models	0.74	Significantly. Above Average	165 17%	96 10%	261 27%	Average Prior Achievement	***-0.49***	***0.36***
B. ANCOVA		Other	14 1.%	676 71%	690 73%	Percent Special Education Teachers	0.21	0.06
		Total	180 19%	769 81%		Average LMT Score	0.31	0.20

Methods	**Pearson Correlation Coefficient**		**Cross Tabulation of Classification as Significantly Above Average by Two Methods** Method A: ANCOVA			**Classified as above average by Method A** **Classified as above average by Method B**	**Yes** **No**	**No** **Yes**
		Method B: Gain Scores	Significantly Above Average	Other	Total	Average Percent Minority Student	***0.52***	***0.75***
A. ANCOVA	0.87	Significantly. Above Average	129 14%	41 4%	170 18%	Average Prior Achievement	***0.55***	***-0.98***
B. Gain Scores		Other	132 14%	649 68%	781 82%	Percent Special Education Teachers	***0.00***	***0.56***
		Total	261 27%	690 73%		Average LMT Score	***0.35***	***-0.49***

Figure 1. Expected Bonus Award for Eight Alternative Decision Rules for Different Values of the True Teacher Performance and Standard Errors of Performance (Assuming no Bias).

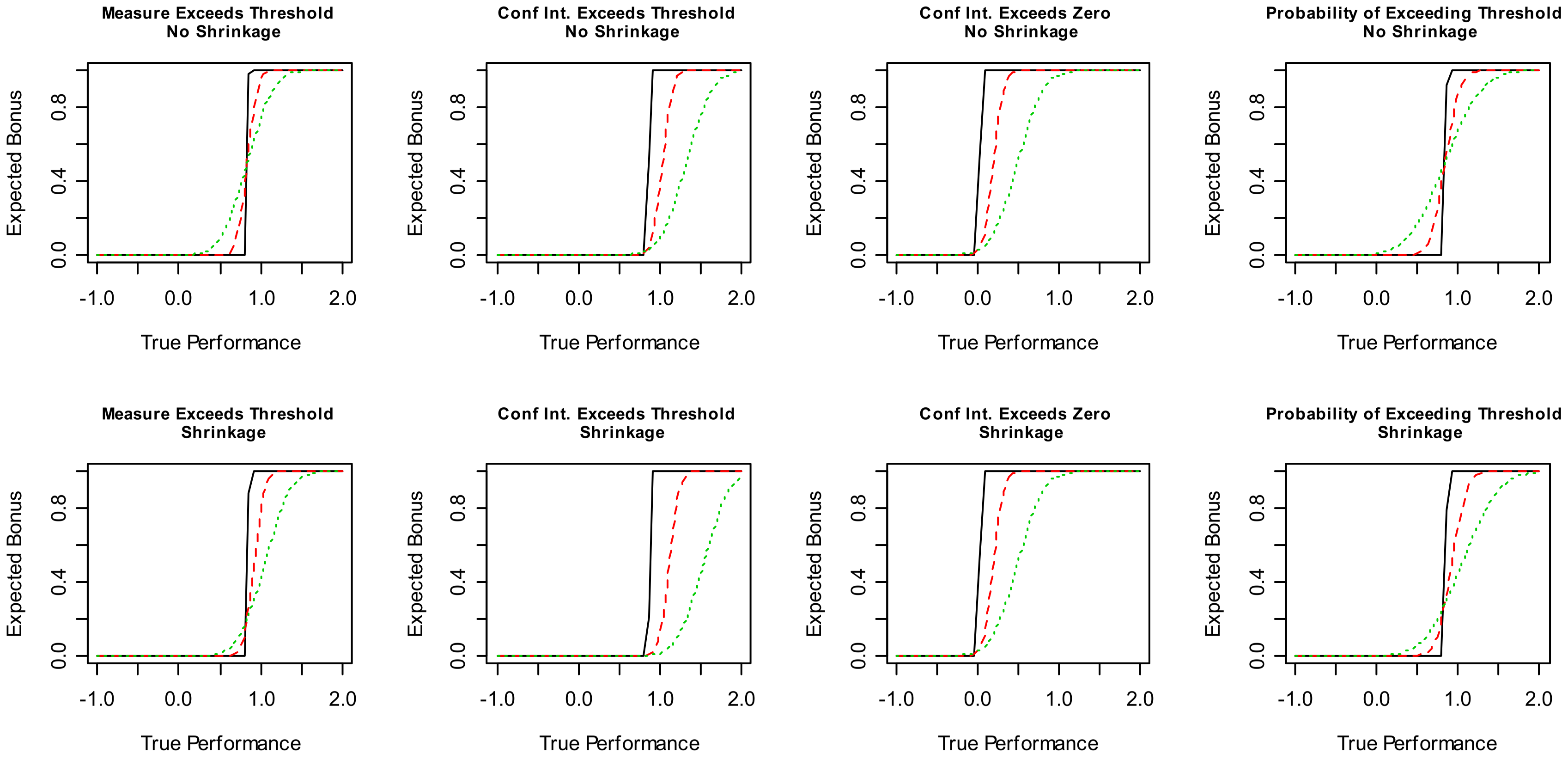
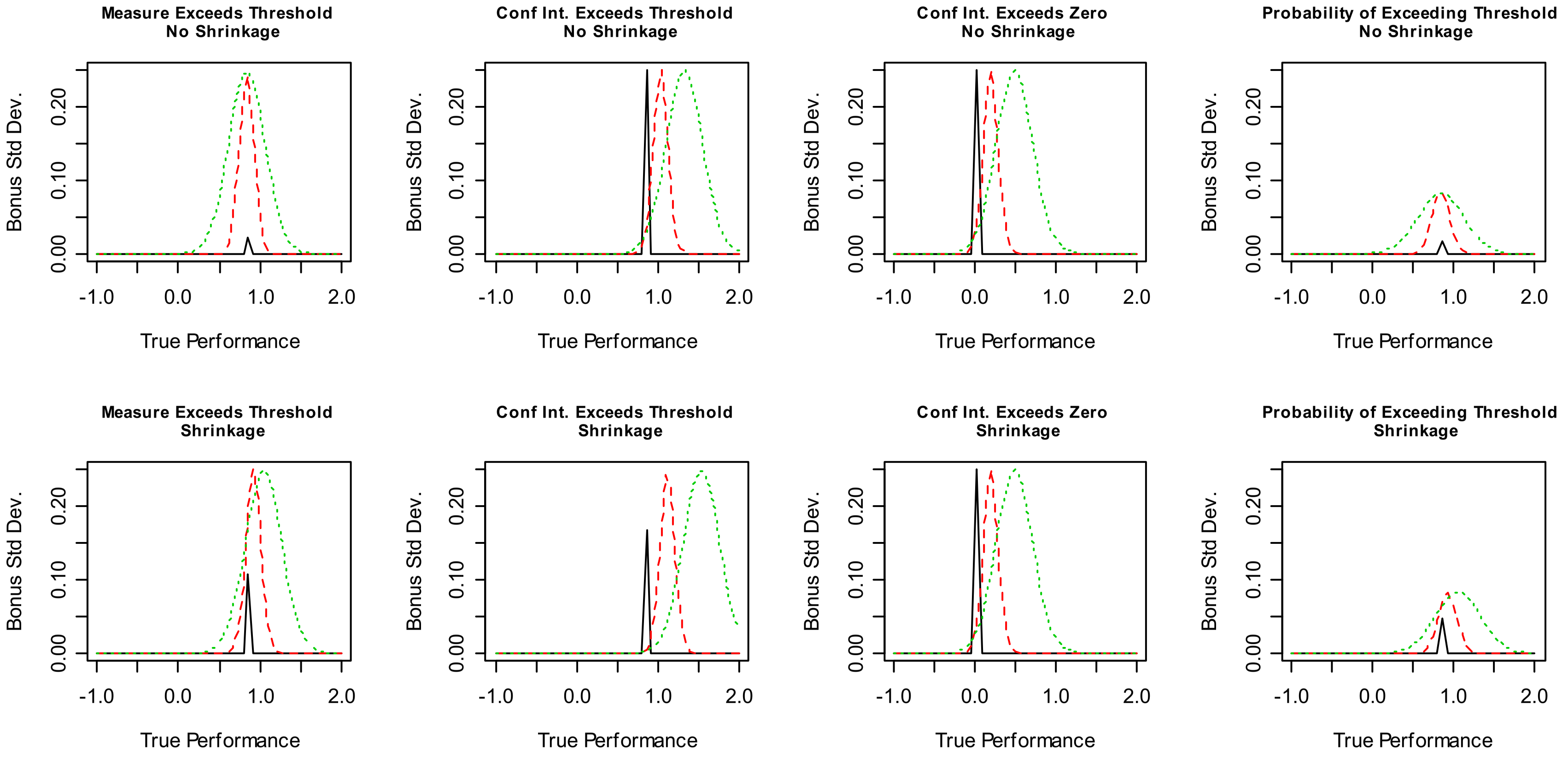


Figure 2. Variance in Bonus Award for Eight Alternative Decision Rules for Different Values of the True Teacher Performance and Standard Errors of Performance (Assuming no Bias).



Measure Exceeds Threshold
No Shrinkage

Conf Int. Exceeds Threshold
No Shrinkage

Conf Int. Exceeds Zero
No Shrinkage

Probability of Exceeding Threshold
No Shrinkage

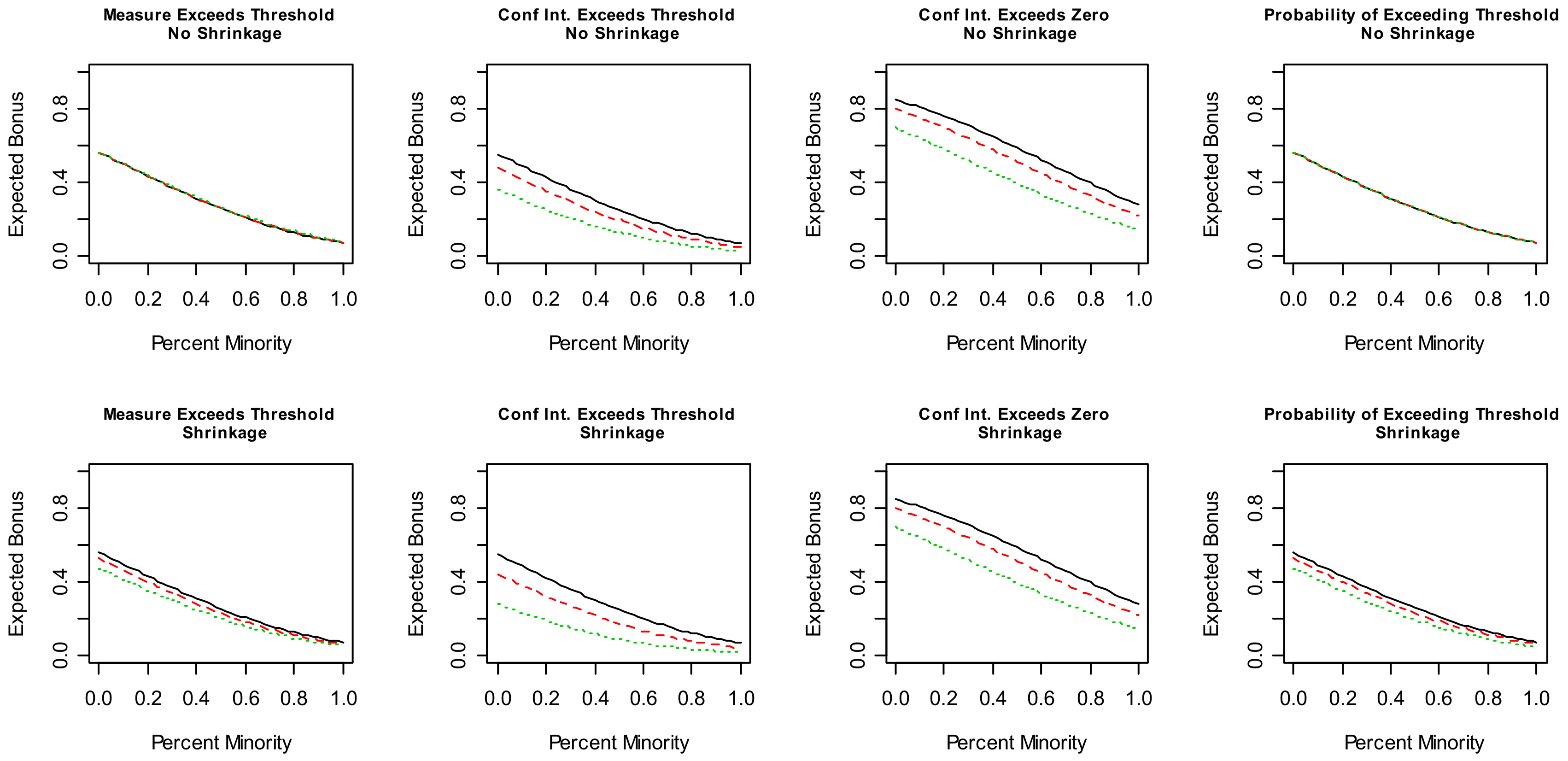
Measure Exceeds Threshold
Shrinkage

Conf Int. Exceeds Threshold
Shrinkage

Conf Int. Exceeds Zero
Shrinkage

Probability of Exceeding Threshold
Shrinkage

Figure 3. Average Expected Bonus by Percent Minority Students for Performance Measures with Strong Bias that Depends on Percent of Minority Students in the Teachers Class



REFERENCES

Aaronson, D., L. Barrow, and W. Sander (2003). Teachers and student achievement in the Chicago public high schools. Technical report, Federal Reserve Bank of Chicago.

Ballou, D., W. Sanders, and P. Wright (2004). Controlling for students background in value-added assessment of teachers. Journal of Educational and Behavioral Statistics 29(1), 37.66.

Betebenner, D. W. (2007) Estimation of Student Growth Percentiles for the Colorado Student Assessment Program. Dover, New Hampshire: National Center for the Improvement of Educational Assessment.

Buddin, R., McCaffrey, D.F., Kirby, S. N. and Xia, N. (2007). Merit Pay for Florida Teachers: Design and Implementation Issues. Working paper, WR-508-FEA. Santa Monica, CA: The RAND Corporation.

Carlin, B.P. and Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Committee for Economic Development (2004). *Investing in learning: School funding policies to foster high performance.* Washington, DC: Committee for Economic Development.

Fleiss, Joseph L. (1981). *Statistical methods for rates and proportions*. New York, NY: *John Wiley & Sons*

Gordon, R., T. Kane, and D. Staiger (2006). Identifying effective teachers using performance on the job. Technical report, The Brookings Institution. White Paper 2006-01.

Harris, D. and T. Sass (2006). Value-added models and the measurement of teacher quality. Unpublished manuscript.

Hill, H. C. (2007). Mathematical knowledge of middle school teachers: Implications for the No Child Left Behind policy initiative. *Educational Evaluation and Policy Analysis, 29*(2), 95-114.

Hill, H. C., & Ball, D. L. (2004). Learning mathematics for teaching: Results from California's mathematics professional development institutes. *Journal for Research in Mathematics Education, 35*(5), 330-351.

Hill, H. C., Ball, D. L., Blunk, M., Goffney, I. M., & Rowan, B. (2007). Validating the ecological assumption: The relationship of measure scores to classroom teaching and student learning. *Measurement: Interdisciplinary Research and Perspective, 5*(2/3), 107-118.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 317-406.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *The Elementary School Journal, 105*(1), 11-30.

Hassel, B. (2002). *Better pay for better teaching: Making teacher compensation pay off in the age of accountability.* Progressive Policy Institute, Washington, DC.

Humphrey, D. C., & Wechsler, M. E. (2007). *Insights into alternative certification: Initial findings from a national study*. Menlo Park, CA: Center for Education Policy, SRI International.

Humphrey, D. C., Wechsler, M. E., & Hough, H. J. (2008a). Characteristics of effective alternative teacher certification programs. *Teachers College Record, 110*(4), 1-47.

Humphrey, D. C., Wechsler, M. E., & Hough, H. J. (2008b). Insights into alternative certification: Initial findings from a national study. *Teachers College Record, 110*(4), 1-47.

Kane, T., J. Rockoff, and D. Staiger (2006). What does certification tell us about teacher effectiveness? Evidence from New York City. Unpublished Manuscript.

Lockwood, J.R., and McCaffrey, D.F. (2007). "Controlling for Individual Heterogeneity in Longitudinal Models, with Applications to Student Achievement." *Electronic Journal of Statistics*, 1(1), 223-252.

Lockwood, J.R., D.F. McCaffrey., L.T. Mariano, and C. M. Setodji. (2007). Bayesian methods for scalable multivariate value-added assessment. Journal of Educational and Behavioral Statistics.

Malanga, S. (2001). *Why merit pay will improve teaching.* City Journal, 11 (3).

Martineau, J. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for value-added accountability. *Journal of Educational and Behavioral Statistics* 31, 35–62.

McCaffrey, D., J. R. Lockwood, D. Koretz, T. Louis, and L. Hamilton (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics 29* (1), 67{101.

McCaffrey, D. F., J. R. Lockwood, D. M. Koretz, and Laura S. Hamilton. (2003). *Evaluating Value Added Models for Teacher Accountability.* MG-158-EDU. Santa Monica, CA: RAND.

McCaffrey, D.M., Lockwood. J.R., Mariano, L. and C. Setodji (2005). Challenges for value-added assessment of teacher effects. In R. Lissitz (Ed.) *Value Added Models in Education: Theory and Applications.* Maple Grove, MN: JAM Press. pp. 111–144.

Odden, A., & Kelley, C. (1996). *Paying teachers for what they know and do: New and smarter compensation strategies to improve schools.* Thousand Oaks, CA: Corwin Press.

Odden, A., Kelley, C., Heneman, H., & Milanowski, A. (2001). *Enhancing teacher quality through knowledge and skills-based pay*. Consortium for Policy Research in Education. Philadelphia, Pennsylvania.

Podgursky, M. & Springer, M.G. (2007). "Credentials Versus Performance: Review of the Teacher Performance Pay Research." *Peabody Journal of Education,* 82(4), 551-573.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. (2005). "Teachers, Schools and Academic Achievement." *Econometrica, 73*(2), 417-58.

Sanders, W., A. Saxton, and B. Horn (1997). The Tennessee Value-Added Assessment System: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading*

Teachers, Grading Schools: Is Student Achievement a Valid Evaluational Measure? Thousand Oaks, CA: Corwin Press, Inc., 137–162.

Schilling, S. G., Blunk, M., & Hill, H. C. (2007). Test validation and the MKT measures: Generalizations and conclusions. *Measurement: Interdisciplinary Research and Perspective, 5*(2/3), 118-128.

Schilling, S. G., & Hill, H. C. (2007). Assessing measures of mathematical knowledge for teaching: A validity argument approach. *Measurement: Interdisciplinary Research and Perspective, 5*(2/3), 70-80.

Snedecor, G.W., and Cochoran, W.G. (1980). *Statistical Methods, Seventh Edition.* Ames, IA: The Iowa State University Press.

Southern Regional Education Board (2000). *Teacher salaries and state priorities for education quality: A vital link*. Educational Benchmarks 2000 Series. Atlanta, Georgia.

Webster, W. and R. Mendro (1997). The Dallas value-added accountability system. In J. Millman (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure*? Thousand Oaks, CA: Corwin Press, Inc., pp. 81–99.

Faculty and Research Affiliates

Matthew G. Springer
Director
National Center on Performance Incentives

Assistant Professor of Public Policy
and Education
Vanderbilt University's Peabody College

Dale Ballou
Associate Professor of Public Policy
and Education
Vanderbilt University's Peabody College

Leonard Bradley
Lecturer in Education
Vanderbilt University's Peabody College

Timothy C. Caboni
Associate Dean for Professional Education
and External Relations
Associate Professor of the Practice in
Public Policy and Higher Education
Vanderbilt University's Peabody College

Mark Ehlert
Research Assistant Professor
University of Missouri – Columbia

Bonnie Ghosh-Dastidar
Statistician
The RAND Corporation

Timothy J. Gronberg
Professor of Economics
Texas A&M University

James W. Guthrie
Senior Fellow
George W. Bush Institute

Professor
Southern Methodist University

Laura Hamilton
Senior Behavioral Scientist
RAND Corporation

Janet S. Hansen
Vice President and Director of
Education Studies
Committee for Economic Development

Chris Hulleman
Assistant Professor
James Madison University

Brian A. Jacob
Walter H. Annenberg Professor of
Education Policy
Gerald R. Ford School of Public Policy
University of Michigan

Dennis W. Jansen
Professor of Economics
Texas A&M University

Cory Koedel
Assistant Professor of Economics
University of Missouri-Columbia

Vi-Nhuan Le
Behavioral Scientist
RAND Corporation

Jessica L. Lewis
Research Associate
National Center on Performance Incentives

J.R. Lockwood
Senior Statistician
RAND Corporation

Daniel F. McCaffrey
Senior Statistician
PNC Chair in Policy Analysis
RAND Corporation

Patrick J. McEwan
Associate Professor of Economics
Whitehead Associate Professor
of Critical Thought
Wellesley College

Shawn Ni
Professor of Economics and Adjunct
Professor of Statistics
University of Missouri-Columbia

Michael J. Podgursky
Professor of Economics
University of Missouri-Columbia

Brian M. Stecher
Senior Social Scientist
RAND Corporation

Lori L. Taylor
Associate Professor
Texas A&M University

NATIONAL CENTER ON
Performance Incentives

EXAMINING PERFORMANCE INCENTIVES
IN EDUCATION

National Center on Performance Incentives
Vanderbilt University Peabody College

Peabody #43
230 Appleton Place
Nashville, TN 37203

(615) 322-5538
www.performanceincentives.org

VANDERBILT
PEABODY COLLEGE