

WORKING P A P E R

Merit Pay for Florida Teachers

Design and Implementation Issues

RICHARD BUDDIN, DANIEL F. MCCAFFREY,
SHEILA NATARAJ KIRBY, AND NAILING XIA

WR-508-FEA

August 2007

Prepared for the Florida Education Association

This product is part of the RAND Education working paper series. RAND working papers are intended to share researchers' latest findings and to solicit additional peer review. This paper has been peer reviewed but not edited. Unless otherwise indicated, working papers can be quoted and cited without permission of the author, provided the source is clearly referred to as a working paper. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors. **RAND®** is a registered trademark.

Abstract

Florida has enacted a plan to reward teachers based on their classroom performance, as measured on standardized student achievement tests and principal evaluations. This merit pay initiative is designed to provide a financial incentive for teachers to improve student outcomes, to encourage the retention of proficient teachers, and to attract high-skilled individuals to the teaching profession. The design and implementation of merit pay faces several key challenges. First, student outcomes are difficult to define and measure. Second, the contributions of individual teachers to student outcomes are difficult to disentangle from student background and prior achievement. The analysis shows serious deficiencies in several measures of teacher performance. Policy makers should be wary of adapting any measure without careful analysis of its properties and a plan to monitor how it is performing. The key issue is whether the incentive and sorting effects of an admittedly imperfect merit pay system can improve the quality of the teacher workforce.

Keywords: merit pay, pay for performance, teacher quality, student achievement, teacher pay incentives, teacher compensation

ACKNOWLEDGMENTS

The authors are grateful to Marshall Ogletree of the Florida Education Association for his support and encouragement of this research. We are indebted to officials from the Pinellas, Sarasota, and Sumter School Districts for providing the student achievement data used in our analysis. We are also grateful to Vi-Nhuan Le and Laura Hamilton of RAND for their comments and suggestions on an earlier draft of this paper.

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 and Kelley, 1996; Odden, Kelley, Heneman and 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—a system that has come under frequent attack. Most recently, the National Center on Education and the Economy (2007), in its report on the skills of the American workforce, called for an overhaul of the education and training system and singled out the teacher compensation system as badly in need of reform, bluntly describing it as “designed to reward time in service, rather than to attract the best and brightest of our college students and reward the best of our teachers.” A report by the Center for Teaching Quality (2007) by a panel of teacher experts declared:

Like the dusty blackboards still found in some school classrooms, the single-salary schedule has served its purposes and outlived its usefulness...

We agree that a carefully-crafted performance-pay system has huge potential to transform the teaching profession in ways that can help all students learn more. We do not shy away from the principle that teachers who perform at high levels and spread their expertise deserve extra compensation for their performance and accomplishments.

The last 30 years have seen a proliferation of merit pay programs as districts have sought to link teacher pay and performance, but these programs have generally not been successfully implemented. Many plans in the 1980's and 1990's struggled with problems in developing teacher evaluations that were valid, reliable, and fair. Ballou (2001) contends that the difficulties in implementing merit pay are exaggerated and that the main impediment to merit pay in public schools is the opposition of teacher unions. However, several studies (Murnane and Cohen, 1986; Hatry, Greiner, and Ashford, 1994; Hassel, 2002) point to the inherent complexity of designing systems that are valid and reliable in terms of being able to consistently identify and reward the most effective teachers: how best to measure “performance” and over what time period, how to link pay with performance levels; the size of the rewards, and how to eliminate preferential treatment from the performance appraisal systems. Springer (2007) argues that some of these evaluation problems have been diminished by increased availability of longitudinal student-level test score data and by sophisticated value-added statistical methodologies. We show below that the evaluation problems remain a significant problem for merit pay implementation.

More recently, districts have experimented with school-level rather than individual-level bonuses (Clotfelter and Ladd, 1996). Regardless of the type of pay-for-performance program, there have been few rigorous evaluations of how teacher- or school-based performance incentives serve to improve student achievement in the U.S. The little that is known, largely from abroad, paints an inconsistent picture of the effectiveness of pay-for-performance programs on student achievement (Burgess, Croxson, Gregg, and Propper, 2001; Lavy, 2002, 2004; Glewwe, Ilias and Kremer, 2003; Muralidharan and Sundararaman, 2006). In many of these studies, researchers have not successfully controlled for the influence of potentially confounding reform strands, nor explored how performance pay programs fit into systemic reform strategies. The current environment, with its emphasis on annual assessments linked to content and performance standards and accountability policies, is one that makes the issue of measuring performance at least theoretically more tractable (Smith and O'Day, 1991). In addition, the literature (Milkovich and Wigdor, 1991; Locke et al., 1981) points to some guiding principles for designing successful pay-for-performance systems. Such systems must ensure that the bonus system truly incentivizes teacher behavior, that teacher cooperation and collaboration is not reduced, that lead teachers are not encouraged to engage in dysfunctional behaviors, that there is a strong link between schooling outcomes and teacher rewards, and that gaming and other "unintended consequences" documented in some high-stakes testing and accountability environments are minimized.

Teacher Merit Pay in Florida

Florida has been at the forefront of proposals to implement merit pay for teachers over the past several years. Since 2001, Florida school districts have been "required" to pay bonuses for outstanding teaching as part of the Performance Pay Statute, although the plan was not fully implemented until 2006, when the Florida legislature passed the Special Teachers are Rewarded (STAR) plan, an individually-based merit plan that ranked teachers based on the performance of students on the Florida Comprehensive Assessment Test (FCAT). The key component of this plan was a value table, which assigned a "value" to students moving across performance categories. Because of opposition from teachers and school districts, the STAR plan was modified and a Merit Award Program (MAP) was adopted that allowed awards to both individuals and instructional teams and was not linked to the use of a value table. Because Florida is the only state to implement a state-wide pay-for-performance plan,¹ it offers a unique opportunity to examine both theoretical and practical issues related to design, implementation,

¹ The Quality Compensation for Teachers program in Minnesota is offered to districts statewide, but currently only 35 school districts and 14 charter schools are participating.

and eventual impacts of such a system on teacher performance, recruitment, retention, and distribution across different types of schools and districts.

Purpose and Scope of this Project

The Florida Education Association (FEA) asked the RAND Corporation to undertake a short-term assessment of Florida's STAR Plan. With the repeal of STAR and the implementation of MAP, the project refocused on broader issues of merit pay for teachers and how a system might be implemented at the district level. The project focused on four major research questions:

1. What are the theoretical underpinnings of pay-for-performance systems? What prior evidence exists about the effectiveness of such systems?
2. How has Florida's pay-for-performance plans been designed and implemented?
3. How is teacher performance being measured? Using FCAT data from selected districts, what can we say about the stability of teacher rankings?
4. In general, what are the challenges to implementing a pay-for-performance system like STAR or MAP?

To address these issues, we undertook a broad-based literature review of both the theoretical and empirical literature on teacher pay-for-performance and the challenges of measuring teacher performance; examined documents related to Florida's STAR plan and its recent metamorphosis into a different merit pay plan; and analyzed student achievement data from Florida to examine issues of measurement of teacher effects and the stability of teacher rankings.

This paper is organized into several sections. The next section examines the rationale for incentive pay and reviews the empirical literature on the impact of teacher incentives on teacher behaviors and student achievement. Section 3 presents an overview of Florida's STAR plan and the newer MAP plan. Section 4 presents findings from our analysis of the FCAT data and their implications for the design and implementation of Florida's pay-for-performance plan. Section 5 outlines the complex theoretical and practical issues that need to be addressed in terms of designing and implementing teacher pay-for-performance plans. A final section presents our conclusions and policy recommendations.

We should note that the paper is limited in scope—it is focused on measuring the performance of individual teachers using test scores and not other measures such as work samples or portfolios. We also do not address the larger questions that would need to be addressed in designing effective pay-for-performance systems such as optimal size of rewards, school versus

individual rewards, etc. Time and budget constraints precluded us from adopting a broader perspective.

2. RATIONALE FOR MERIT PAY AND REVIEW OF THE LITERATURE

This section has three subsections. The first discusses the theoretical rationale for linking pay to performance; the second surveys the empirical literature on how these compensation schemes have worked; and the third examines what is known about two design issues that are of particular relevance to Florida—individual versus team-based pay and using principal evaluations to measure teacher performance. Both of these are elements of Florida’s newly implemented MAP.

Rationale for Merit Pay

The concept of merit pay for teachers has been around for many years. The rationale behind such practice is straightforward—teachers, like all workers, are expected to respond to incentives inherent in their compensation structures. The current pay structure for teachers in the U.S. is input-based—teachers are paid on the basis of their skills, which are measured by education and certification and teaching experience (Lazear 1986, 2000; Barro and Beaulieu, 2003). The premise is that these input measures are ultimately linked to desired outcomes (i.e., more student learning or skills). On the other hand, merit pay programs are essentially output-based payment schemes that tie financial rewards to some direct metric of student performance.

Merit pay is “results oriented” in the sense that compensation focuses on the production of specific student outcomes. The challenge for designing a merit pay system for teachers is in defining an appropriate composite of student learning (output) and in measuring teacher performance in producing learning. Student test scores presumably reflect a portion of learning, but even well designed achievement tests miss some skills entirely and imperfectly measure others. In addition, teacher performance is difficult to disentangle from the mix of students assigned to a teacher, and teaching resources available in each classroom.

If there were consensus on what student outcomes mattered and measurement issues were resolved, merit pay could improve the quality of the teaching workforce in two ways. First, merit pay provides teachers a direct financial incentive to increase effort by improving the specified student outcomes. Teachers are encouraged to align their teaching objectives with the educational system for their students (Lazear, 2003; Lashway, 2001; Eberts, Hollenbeck, and Stone, 2002; Ballou, 2001).

The second mechanism by which merit pay can improve the teacher workforce is the sorting effect on teachers. Some teachers or potential teachers might have better inherent ability to

improve student outcomes. Perhaps these individuals are better able to motivate, to communicate, or to “make learning fun.” A merit pay system raises the relative pay of these higher ability teachers relative to lower ability teachers. Teachers sort themselves depending on their ability to improve student outcomes—higher-ability individuals are attracted to teaching and higher-quality teachers are encouraged to remain in the teaching profession (Lazear, 2003).² A traditional teacher pay schedule does not differentiate between high- and low-ability teachers in terms of student outcomes, so there is no relative advantage for high-ability teachers to enter or remain in the teaching profession.

However, merit pay may also present several disadvantages compared to a traditional pay schedule. First, some tasks inherently involve team production, so individual contributions are difficult to disentangle. A compensation system could reward team output, but this would create incentives for individuals on the team to “free ride” on the efforts of others. Second, individual rewards for quantity produced will encourage undue emphasis on quantity alone in some circumstances. For example, if teachers received bonuses for the number of students reaching a reading proficiency level, then they would have little incentive to focus on student above the proficiency level. Similarly, teachers might simply “teach the test” at the expense of promoting long-term learning. Third, most employees like predictable earnings and dislike large fluctuations in monthly income. This suggests that merit pay for teachers should comprise a portion of teacher pay, but not on the bulk of teacher compensation.

While theory predicts the effectiveness of merit pay programs in increasing student performance, there is little empirical evidence on this issue. One possible reason is the inherent nature of the education process, which may confound the potential efficacy of performance-based compensation systems. As many researchers have argued, education is a complex system with multiple stakeholders, disparate and poorly observable goals, multi-task jobs, and team production (Dixit, 2002; Eberts, Hollenbeck, and Stone, 2002). Even though teachers, as economic actors, may respond to incentives, “there may be several wedges between performance measures and the actions of teachers that tend to mitigate against individual level, incentive-based compensation schemes” (Eberts, Hollenbeck, and Stone, 2002: 917).

Empirical Findings on Teacher Incentive Programs

Arguments for merit pay programs rest on the following three presumptions:

² Merit pay might also be structured to improve the sorting of teachers within a school district. For example, suppose that improvements in student outcomes were particularly important for low-achieving students. If output measures assign large weights to these outcomes, then teachers would have incentives to take assignments in schools with at-risk students.

1. Financial rewards can motivate teachers to change their behaviors.
2. Teachers can influence student learning.
3. Performance-based incentive programs for teachers can be effective in improving student outcomes.

Evidence regarding the validity of each of these assumptions is provided below.

Teachers' Response to Financial Incentives

The limited research on teachers' response to merit pay programs suggests that financial incentives can induce changes in teacher behaviors. Glewwe, Ilias, and Kremer (2003) examined teacher behaviors in response to a school-level bonus program that tied teacher financial rewards to student test scores in primary schools (grades 4 to 8) in Kenya. In comparison with teachers in the control schools, teachers in the incentive schools increased their efforts to raise student test scores by conducting more test preparation sessions, and the incentive schools experienced significant increases in test scores in the short run.

Lavy (2004) evaluated the effects of cash bonuses for teachers based on their students' performance on high-school matriculation exams in Israel. The results suggest that teacher incentives have a significant effect on raising student achievement and that "the improvements appear to come from changes in teaching methods, after-school teaching, and increased responsiveness to students' needs" (Lavy, 2004: 3).³

In contrast, literature on the effect of salary levels on teachers' decisions to enter and remain in teaching is much more extensive. Many studies focus on teacher recruitment and leaving decisions in relation to their pay levels. Hanushek, Kain and Rivkin (1999) examined whether school districts with higher pay for teachers attracted better recruits, as measured by teacher test scores. They did not find a significant relationship after controlling for district fixed effects. However, using data from New York City, Jacobson (1995) found that the starting wage affects teacher recruitment and that the relative wage rate affects teacher retention rates. Similarly, Murnane and Olsen (1989, 1990) reported a positive impact of wages in alternative jobs on teacher attrition in North Carolina and Michigan.

³ Teacher emphasis on test-taking techniques might improve student scores without a corresponding increase in student learning. If so, then test score gains would be a misleading indication of the efficacy of merit pay.

Teacher Effects on Student Achievement

Teacher inputs appear to be one of the most important school factors in influencing student achievement. Hanushek, Kain and Rivkin (1998) reported that teacher effects account for at least 7.5 percent of the total variation in student attainment. In their evaluation of value-added literature on teachers' contributions to student learning, McCaffrey et al. (2003) analyzed and simulated the methods used in the existing literature. They concluded that teachers do affect student outcomes, though the size of teacher effects was unclear from the literature at that time.

Multiple studies have found evidence of wide variation in teacher impacts on student achievement even after adjusting for student characteristics such as baseline test scores, race and ethnicity, family income, gender, etc. (Koedel and Betts, 2007; Gordon, Kane and Staiger 2006; Rivkin, Hanushek and Kain 2005; Nye, Konstantopoulos and Hedges 2004; Rockoff 2004; Aaronson, Barrow and Sander 2003). Hanushek (1992) and Hanushek and Rivkin (2003) reported that, compared with teachers at the bottom of the quality distribution, high-quality teachers could get an entire year's worth of additional learning out of their students.

However, evidence on which teacher characteristics matter for student attainment is less clear-cut and studies that find large variance among teachers often find little evidence that specific teacher characteristics, like experience or education level, explain much of that variance (Koedel and Betts, 2007; Rivkin, Hanushek and Kain, 2005). In general, teacher experience is often found to have positive effects on student achievement. Hanushek, Kain and Rivkin (1998) and Hanushek et al. (2005) found that teacher experience appeared to matter in raising student test scores, but only in the first one or two years of teaching. Kreuger (1999) reported a positive but small effect of teacher experience. A review of 90 studies found that the majority of the studies reported positive impact of teacher experience, although only half of them were statistically significant (Hanushek, 1997; Hanushek and Rivkin, 2003).

Teacher training and educational background are sometimes found to be associated with student attainment (Burgess et al., 2001; Hedges, Laine, and Greenwald, 1994; Dewey, Husted, and Kenny, 2000). Based on a natural experiment among schools in Jerusalem, Angrist and Lavy (2001) reported positive results from additional in-service pedagogical training. However, having a master's degree is not significantly associated with student test scores (Hanushek, 1997; Hanushek and Rivkin, 2003). Finally, literature provides mixed evidence on the effects of certification on student achievement (Burgess et al., 2001; Hanushek, 1997; Hanushek and Rivkin, 2003). Goldhaber and Brewer (1997) find that math students have higher test scores

when they are taught by teachers with degrees in mathematics relative to students whose teachers have degrees outside of math.⁴

Effects of Teacher Incentives on Student Outcomes

Empirical evidence on the efficacy of merit pay schemes is rather limited. Most literature on merit pay looks at the institutional experiences, which are often found to be short-lived and negative (Eberts, Hollenbeck, and Stone, 2002; Murnane and Cohen, 1986). The effectiveness of merit pay programs in improving student achievement in the U.S. has not been well researched.

The existing scant evidence is largely from abroad. Findings from the eight studies summarized below generally support the notion that performance-based teacher incentives, both group (school-level) and individual (teacher-level) merit pay programs, are associated with at least short-term increases in student test scores.

Among the few studies on the effectiveness of merit pay in raising student achievement, most examined group (school-level) incentive schemes, that is, an incentive system that rewards teachers based on the average performance of students in the school. For example, Lavy (2002) found that a school-level incentive program in Israel increased average test scores and reduced student dropout rates in the participating schools.

Using a randomized evaluation of a teacher incentive program for primary schools in India, Muralidharan and Sundararaman (2006) looked at both group (school-level) and individual (teacher-level) incentive programs. The program provided teachers with bonus awards equivalent to 3 percent of annual salary based on student gains in math and language skills. The study found that test scores rose by about 0.15 standard deviations in incentive schools compared with control schools. Moreover, students in incentive schools also performed better on subjects other than math and language skills that were not used to determine teacher rewards, “suggesting positive spillover effects from incentive to non-incentive subjects” (2006: 2).

Glewwe, Ilias and Kremer (2003) reported that a group incentive program in Kenya motivated teachers to raise test scores in the short term. The bonus awards were from 20 to 40 percent of annual salary. No long-term achievement gains were found among students, however. The authors argue that the short-lived achievement gains are an indication that the program encouraged teachers to focus on test-taking skills instead of promoting learning.

⁴ Teachers who are certified to teach particular subjects may not necessarily have a degree in that subject (their degree might be in education or mathematics education rather than mathematics, for example), although at the middle school and high school levels, most states now require a disciplinary degree.

By comparing schools in Dallas where school-level incentives were instituted with those in five other Texas cities without incentive programs, Clotfelter and Ladd (1996) and Ladd (1999) found that student pass rates in exams increased in Dallas relative to other cities.

Individual (teacher-level) incentive schemes—the model used in Florida STAR and MAP plans—reward individual teachers based on the average performance of the students they have taught. Lavy (2004) evaluated an Israeli rank-order tournament that rewarded teachers for improving their students' test scores on high-school matriculation exams in English, Hebrew and mathematics. The results suggest that performance incentives have a significant effect on increasing students' passing rates of these exams as well as some minor spillover effects on untargeted subjects that were not part of the incentive program. As mentioned earlier, Muralidharan and Sundararaman (2006) found significant positive effects of individual teacher incentives on student test scores in primary schools in India. Combining data from National Education Longitudinal Survey with their own survey, Figlio and Kenny (2006) found a significant association between merit pay for teachers and higher student test scores in the U.S.⁵

While the above studies looked at merit pay programs that reward teachers based on the test scores of their students, Eberts, Hollenbeck and Stone (2002) examined teacher-level incentive payments for student retention in a high school in the U.S. By comparing means across two schools, they found that merit pay was associated with a significant fall in dropout rates but had no effect on grade point averages.

Three studies have compared incentive schemes with programs investing in other school resources and found incentive programs to be more cost-effective in raising student achievement than other school investments. Lavy (2002) compared Israeli schools participating in a school-level incentive program with those receiving additional conventional resources such as extra teaching time and on-the-job teacher training. The results suggest that incentive intervention is much more cost-effective in improving student performance than conventional school resource programs. A cost-benefit comparison of an Israeli teacher-level incentive program with other relevant interventions by Lavy (2004) suggests that financial incentives for individual teachers are “more efficient than a program that targeted instruction time to weak students and as efficient as paying students monetary bonuses to improve their performance” (2004: 4). Muralidharan and Sundararaman (2006) found that both group and individual

⁵ Figlio and Kinney (2006) use cross-sectional data and find that student achievement is higher in schools with teacher performance incentives. The authors caution that this finding may mean that better schools are more likely to adopt teacher performance incentives rather than teacher incentives promoting stronger student achievement.

teacher incentive programs appeared to be more cost effective than the two alternative interventions, that is, an additional teacher aide and cash block grants to schools.

Evidence Regarding Two Design Issues

Group Versus Individual Incentives

One of the principal objections to STAR was the fact that it was individually-based and did not allow for team rewards, something that MAP allows. The issue of whether group (school-level) or individual (teacher-level) incentives are more appropriate and effective in raising student achievement has long been debated. Proponents of group incentives argue that teaching requires team-based cooperation. As Eberts, Hollenbeck and Stone (2002) described, “many elementary and middle schools are organized into teams of teachers” and “even in departmentalized secondary schools student performance on standardized tests depend on learning in several courses taught by different teachers” (2002: 916). Some researchers argue that the inherent nature of the education process is inconsistent with an individual-based compensation system that may result in unhealthy competition or even sabotage behaviors among teachers (Burgess et al., 2001; Glewwe, Ilias, and Kremer, 2003; Lavy, 2004).

Supporters of individual incentive schemes point out that most classroom settings involve only one teacher. They argue that school-level test scores are imperfect measures (Figlio and Kenny, 2006; Kane and Staiger, 2002) and that group-based plans may not provide sufficient incentives due to the free-rider problem especially when the group gets large (Lazear, 2003; Lavy, 2004). Two studies, both examining this issue in non-U.S. education settings, compared the effectiveness between school-level and individual teacher incentives. Lavy (2004) reported that financial incentives for individual teachers were more efficient than teachers’ group incentives based on a cost-benefit comparison of two Israeli incentive programs implemented at different times with different eligibility criteria. Muralidharan and Sundararaman (2006) conducted an experimental evaluation in India and found no significant difference in the effectiveness of school-level versus individual teacher incentives.

Principals’ Assessments of Teachers

Both STAR and MAP assign considerable weight to principal evaluations of teachers: 50 percent and 40 percent respectively in the overall score. The belief is that subjective evaluations might provide a more comprehensive or holistic measure of teacher contributions than is possible from student test scores or other objective measures of performance alone. Teacher efforts at discipline, mentoring, citizenship and other factors may be important to the overall learning environment at a school. Given the difficulty of measuring the full range of school outputs, principal evaluations are likely to play an important role in improving teaching.

While subjective evaluations are common in private sector firms, these evaluations have a number of limitations. First, evaluations are not likely to be consistent across evaluators, and the accuracy of evaluations is difficult to assess. Second, evaluations are frequently influenced by subordinate personality or demographic similarity to the supervisor (Jacob and Lefgren, 2005). Real or imagined favoritism in subjective evaluations may promote worker animosity. Third, subordinates are encouraged to curry favor with supervisors in ways that bear little relationship to firm objectives (Prendergast, 1999). Finally, supervisors are reluctant to differentiate between good and bad performance (Prendergast, 1999), since differentiation may agitate employees without encouraging improvements in performance. Compression of scores or rankings towards the upper end of the distribution is likely to occur when evaluations are used as part of a pay setting process.

Jacob and Lefgren (2005) examined whether principals' evaluations of teachers in their schools were related to student achievement. As part of their study, they surveyed school principals and asked them to rate teachers. These ratings were used as explanatory variables in a longitudinal model of student achievement, where students were linked with individual classroom teachers. The results showed that principal ratings were a significant predictor of student achievement and that the ratings were a stronger predictor of classroom success than were teacher characteristics like experience and educational preparation that are typically used in compensation tables. The principal rankings did not perform as well as overall teacher value-added measures, however.

The study found considerable compression of teacher ratings—the average teacher scored 8.1 on a 10-point scale where 8 is the top of the “very good” category and 9 is “exceptional.” The Lake Wobegon effect for principals means they are prone to rank nearly all teachers as well above average. This extreme compression occurs on a survey where teacher evaluations are not part of any formal district assessment of teaching proficiency.⁶ The pattern of compression is likely to be even greater with high-stakes principal evaluations that have consequences for teacher compensation (Asch, 2005; Prendergast, 1999).

With this as background, we now turn to an examination of Florida's pay-for-performance plans. The next section provides a brief overview of the elements of STAR and MAP and

⁶ The survey instrument included anchors to minimize compression of the scale. The exceptional category was defined as “the teacher is among the best I have ever seen in this area (e.g., in the top 1% of teachers).” With this explicit proviso, over 10 percent of principals rated all of their teachers as exceptional.

Section 4 uses FCAT data to model the results of using these plans to rank teachers and the persistence of these rankings over time.

3. FLORIDA'S TEACHER PAY-FOR-PERFORMANCE PLANS

As mentioned earlier, the 2006 Florida legislature appropriated \$147.5 million within the Florida Education Finance Program for the STAR performance pay plan. Early in 2007, the legislature revised the performance pay plan, replacing STAR with MAP. MAP builds on the STAR plans, so the features of STAR are important for understanding the environment for performance pay in Florida. Consequently, in this section of the report we first describe the STAR program and then provide detail on MAP.

For the STAR program, Section 1012.22 of the Florida Statutes required districts to adopt a salary schedule that linked some part of instructional employees' salary to performance and to adopt a pay-for-performance policy for both school administrators and instructional personnel. Rewards of at least 5 percent of base pay were to be provided to a minimum of 25 percent of instructional personnel.

In addition, the regulations required the State Board of Education to approve a district's STAR plan before the district could receive STAR funds. A technical assistance paper (Florida Department of Education, 2006a) outlines the key components of a STAR plan as well as those components that the plan *could not* include:

A district plan (1) must be fair and equitable, neither penalizing nor rewarding teachers for where their students begin the school year academically; (2) must allow for all instructional personnel to be eligible for an award; (3) must weight improved student achievement as at least 50 percent of the total evaluation score; (4) must be understandable and transparent; (5) must be designed to reward those teachers whose students show the greatest learning gains (i.e., the plan must reward high performing teachers in low performing schools, and not reward low performing teachers in high performing schools); (6) must be designed to reward individuals and not groups; and (7) must utilize a district-approved evaluation system with four to six levels of performance.

There are multiple options for measuring student learning gains. Districts may use the FCAT, NRT [norm-referenced test], AP [Advanced Placement], IB [International Baccalaureate], AICE [Advanced International Certificate of Education], or district end-of-course tests. Other options may be used as well, such as skill measures that assess baseline skills at the beginning of a course and growth at the end of a course (e.g., in PE, Art, Music). Regardless of what type of measure is used, there must be a learning gains component.

The Department (Florida Department of Education, 2006b) outlined a basic plan model and sample teacher evaluation to assist districts in constructing their STAR plans, but was careful to suggest that this was simply a model and that program requirements could be met in a variety of ways. The centerpiece of this plan was a value table, described as a “valid and reliable way to measure improved student achievement and give credit to teachers” (p. 4), and further as:

- “Fair and equitable
- Easy to calculate
- Transparent
- Flexible with regard to subject and grade levels
- Ensures a focus on specific educational goals and values” (p. 5).

Value Tables

Value tables are a recently developed approach to weight student achievement gains between adjacent years and produce a measure of school or teacher performance (Hill et al., 2005). Unlike most other performance measures, which are based on changes in scale scores, value tables use changes in discrete proficiency levels to measure teacher performance. The approach has been used primarily to assess school-level performance, but it was also advocated by the Florida Department of Education (FLDOE) as part of the STAR system. The idea is to translate value-added concepts into a simple metric that is easy to implement and easy to understand for educators.

Table 3.1 shows an example of a value table that was suggested by FLDOE as part of STAR.⁷ The table assigns points for each student in a teacher’s classroom based on the students FCAT achievement level in the prior year as compared with their achievement level in the current period. For example, if a student with minimal success at grade-level material (Level 1) improves and become partly successful at grade-level material (Level 3), then the teacher would receive a score of 300 for that student. Scores would be tallied across all students taught by each teacher in the year. The average score across students would be used to rank each teacher across grade and subject to determine whether teachers were performing well or poorly compared with their peers.

⁷ The table is available at http://www.fldoe.org/news/2006/2006_04_05/ValueTable.pdf (accessed on June 21, 2007).

Table 3.1. Possible Value Table for Assessing Teacher Performance

Year 1 Achievement Level Success at Grade-Level Content	Year 2 Achievement Level				
	1	2	3	4	5
1—Minimal success	0	190	300	415	500
2—Limited success	0	75	175	210	250
3—Partly successful	0	0	120	155	175
4—Mostly successful with challenging	0	0	0	130	180
5—Successful with most challenging	0	0	0	70	140

Value tables have several characteristics that distinguish them from other approaches to assessing student achievement progress. First, progress is only measured over discrete levels from one year to the next. Teachers are not given credit for gains or losses within an achievement level. Second, gains and losses are weighted according to a somewhat arbitrary point scheme chosen by policy makers to value behaviors they support. Generally, the point allocations are likely to reflect the perceived difficulty of moving from one achievement level to another one, but no objective criteria are used to assess whether this difficulty scale is appropriate.

A teacher would be assigned points for every student and an average ‘score’ calculated for the teacher by dividing her total score by the number of students. If the teacher taught more than one class, the teacher’s value score would be the average score for all students taught by the teacher. The teacher’s value score would be computed in each grade and subject area. Teachers would then be ranked in order and cut-off scores established for the top 25%, top 20%, and so on and teachers would be assigned points on an evaluation scale based on where they were located in the distribution.

Subjective Evaluations of Teachers

STAR mandated that at least 50 percent of teacher performance would be tied to student test score gains, and the remainder would be based on evaluations of teachers by principals or other school-based administrators. These evaluations would include assessments of each teacher’s ability to maintain appropriate discipline, knowledge of subject matter, ability to plan and deliver instruction, ability to evaluate instructional needs, ability to establish and maintain a positive relationship with students’ families, and other professional competencies or responsibilities as determined by the local school boards.

These process-type measures of teacher performance were traditionally collected in Florida, but the STAR plan linked the measures more directly to ranking teachers and to bonus awards.

Backlash to STAR

The STAR program faced considerable opposition from districts, teachers, and teacher unions. The concerns focused on a number of issues. First, the legislation was seen as a substantial intervention by the state and FLDOE in teacher assessments that had traditionally been the domain of local districts. FLDOE was seen as prescribing how districts should monitor teacher performance, without accommodating teacher or district concerns. Second, many saw the heavy emphasis on standardized student achievement gains as misplaced. Teachers were concerned that the performance ranks based on FCAT were unreliable and that evaluation methods in non-core subjects were even more deficient. Third, STAR limits on the share of teachers receiving awards were viewed as artificial. Fourth, there were concerns about the failure to recognize the contributions of teams rather than simply individuals.

Given these concerns, several unions and districts were unable or unwilling to craft and ratify plans that satisfied STAR criteria as set forth by FLDOE. Without approved performance plans, districts faced the potential loss of STAR funds in the spring of 2007. The prospect of districts foregoing STAR funding created further political pressure on the program, since the legislation called for undisbursed STAR funds to be reallocated to other districts with approved STAR performance plans. This controversy led to the passage of the MAP program to replace STAR.

From STAR to MAP

MAP was designed to address concerns about the STAR program and provided substantial flexibility to local districts and teacher representatives to design a merit pay system for teachers.⁸ Table 3.2 describes several important features of both programs. MAP expanded the coverage of school personnel to include school-based administrators as well as instructional personnel. The new program allowed awards for group performance as well as individual performance, but group size was limited to something smaller than an entire school.

An important difference between STAR and MAP is the inclusion of learning proficiency measures in MAP as well as student improvement or gains. Under MAP, districts have the authority to award teachers based on end-of-course exams on student proficiency without considering the prior achievement of students assigned to each class. In contrast, STAR

⁸ Districts had several choices for the bridge year 2006-2007, depending on whether their STAR plan has been approved—they could choose to keep their STAR plans as approved or complete one to get approval, amend the plan to a MAP or develop a MAP.

required districts to compare how much students improved in a teacher's class, conditional on some performance level in the previous year.

The distinction between student proficiency and student gains are likely to have important implications for how teacher performance is measured. If a teacher is assigned students with high previous academic success, then those students are much more likely to achieve end-of-course proficiency than would lower-proficiency students assigned to a different teacher. While districts may still focus on gains under MAP, simple proficiency measures are likely to provide a distorted picture of what an individual teacher contributes to the learning of students assigned in to teacher's classroom.

The emphasis on standardized testing is larger under MAP than under STAR (60 percent of overall score as compared with 50 percent), but the districts have substantially more flexibility in how to use standardized testing under the new legislation. The various components of principal evaluations remain unchanged between the two programs.

The weight or share attributed to standardized test results in a teacher performance measure may be a misleading indication of how heavily these results factor in overall teacher rankings. As discussed above, there is often considerable compression in subjective evaluations especially when these evaluations are part of a merit pay system. In contrast, the construction of teacher test score ranks is guaranteed to produce a range of outcomes. A weighted sum of the subjective evaluations and test score rankings is likely to be highly driven by the test score rankings.

The MAP program removes the ceiling on the percentage of teachers receiving bonuses, but this revision is a bit misleading. The legislature left the budget allocation for the merit pay system at the same level for MAP as for STAR. This budget was predicated on awards to 25 percent of instructional personnel under STAR, so there are no corresponding budget dollars to advance the program beyond the 25 percent of personnel covered in the STAR program. In addition, MAP adds administrators to the pool of eligible recipients, so the same budget is spread over more eligible recipients. Smaller bonuses would allow for more awards to a greater number of individuals, but the new legislation keeps the 5 percent floor on the size of bonus payments.

An interesting provision of the new law is the tying of bonus payments to average district teacher salary instead of individual base salary. Teacher salaries vary considerably by teacher experience and education level. The linkage to average salary instead of individual salary means that bonus amounts will be scaled back for highly paid teachers and scaled up for lowly paid teachers. The implicit signal from the new legislative language is to provide relatively greater incentives for younger teachers to perform well as compared with older teachers.

Table 3.2. Comparison of STAR and MAP Features

Program Feature	Special Teachers Are Rewarded (STAR)	Merit Award Program (MAP)
Eligibility	Instructional personnel	Instructional personnel and school-based administrators
Basis for bonus awards	Individual performance	Individual or team performance
Types of performance	At least 50% based on improved student performance Remainder based on principal/supervisor evaluations	At least 60% based on learning gains, or proficiency or both Remainder based on principal/supervisor evaluations
Scope of awards	At least 25% of teachers receive a bonus of 5% or more of individual base salary	Top teachers and school-based administrators receive bonus of 5 to 10% of the district's average teacher salary
FLDOE oversight	FLDOE has substantial authority in recommending plan features, reviewing district plans, and recommending district revisions	FLDOE has limited discretion to assess whether district plans meet general guidelines
Size of Program	About \$150 million	About \$150 million
Allocation of undistributed funds	If some districts do not meet STAR requirements, funds reallocated to STAR eligible districts	Reallocation provision is dropped and undisbursed funds revert back to the state

FLDOE has less authority over district plans under MAP than under STAR. FLDOE will provide technical assistance and advice under the new legislation, but the major decisions will be made by the local districts and teacher representatives. MAP only provides broad outlines of what a performance plan should entail, so plans may vary considerably from district to district.

MAP makes program participation voluntary, but it does not disburse unallocated state funds to program participants. Unlike the STAR program, MAP participants will not receive extra bonus dollars if other districts choose not to participate in the program.

Because MAP is relatively new and has yet to be implemented, there is little evidence on exactly how districts will structure performance plans under the law. In the next section, we examine various teacher performance measures (including value tables) and assess possible limitations in using these measures in a merit pay plan.

4. USING STUDENT ACHIEVEMENT TESTS TO ASSESS TEACHER QUALITY

Student achievement tests are a powerful, albeit imperfect, metric for assessing student learning and academic progress. In Florida, the FCAT is aligned to state curriculum standards and provides an objective basis for comparing how students in the state are performing in different subjects. These comparisons provide a basis for assessing the efficacy of educational reforms and determining how alternative configurations of school resources affect learning. However, converting test scores to desirable measures of teacher performance is not straightforward because student backgrounds and preparation differs across classrooms and because many teachers teach relatively few students in each year. The performance measure must account for differences in the student populations among classrooms and provide precise estimates that distinguish among teachers.

This section examines the practical issues related to measuring teacher performance from FCAT data for a district in Florida. The section discusses six questions about how teacher performance measures behave.

1. How does student achievement vary across students, teachers, and schools?
2. How much variability is there among teachers for each estimated performance measure?
3. How similar are teacher rankings from alternative performance measures?
4. Are estimated performance measures stable across time?
5. How much do sampling errors affect teacher performance measures?
6. Do performance measures reward teachers for the students they teach?

The answers to these types of questions provide insights into how student achievement tests might be used a part of a merit pay system.

A cautionary note is that this analysis is conducted on student achievement data that were gathered when teacher compensation did not include merit pay directly tied to student test scores. A premise of merit pay is that this incentive improves performance through improved effort or perhaps through innovative teaching methods. If this premise is true, then the patterns of student achievement in existing data might provide an incomplete indication of the properties of performance measures in a merit pay environment. The analysis does provide insights into the tools available for assessing teacher quality and potential problems in isolating how teachers contribute to student achievement.

The Data

The student achievement analysis focuses on data from Pinellas County Schools in Pinellas County, Florida.⁹ Pinellas is the 23rd largest district in the United States and the 4th largest school district in Florida. The database consists of FCAT scores and student characteristics in reading and math for grades three through ten for the 2003-2004, 2004-2005, and 2005-2006 school years. Each student record is linked with a specific reading/language arts and math teacher. Most elementary students (grades three through five) are taught in self-contained classrooms and have the same teacher for reading and math. In contrast, most secondary students have different reading and math teachers. The file includes student achievement records for about 89,000 students, 3100 reading/language arts teachers, and 2700 math teachers.¹⁰

The Pinellas data is longitudinally linked, so student achievement progress is tracked over time as the student moves across grades, teachers, and schools. Within a single year, students are nested within teachers' classrooms and schools, but across years, students have several teachers and sometimes different schools. The year-to-year movement of students across teachers is useful to isolate how particular teachers affect the learning trajectory and achievement gains of students.

Performance Measures

We consider three methods that span a range of simple models that might be used to estimate teacher performance. The first is the value table score (as discussed above), the second is a gain score model of student achievement, and the third is a random effects test score model. Each method attempts to isolate the contributions of a current year teacher to student achievement, while controlling for the types of students assigned to each teacher.

⁹ We have also analyzed data from Sarasota and Sumter school districts in Florida. The results from those districts are consistent with those reported for Pinellas. The Sarasota data only contained two years of test score information, so we were unable to perform the full range of comparisons as reported for Pinellas. Sumter is a small district, so the teacher comparisons were only based on a small number of teachers.

¹⁰ The Pinellas data include grade, school year, race/ethnicity, gender, and free/reduced lunch status for students. The data do not include demographic characteristics, educational attainment, or experience of teachers assigned to classrooms. This analysis is focused on identifying overall quality differences across teachers—these overall differences are generally the issue in assigning merit pay bonus based on student achievement outcomes. Of course, a portion of these teacher effects may be related to teacher education or experience, and these two teacher attributes are generally rewarded in existing teacher compensation systems.

As described above, the value table approach assigns points to a teacher based on improvements in FCAT levels. The comparison assumes that the prior year score captures student preparation and that the current year score indicates student subject matter knowledge at the completion of a teacher's term. The change in level is assumed to provide information about the teacher's input into learning and the weighting rewards inputs that are considered most desirable. The final scores for teachers are assumed to provide fair indication of teacher performance. The scores are clearly not exact measures of teachers' effects on student outcomes since the weighting of changes in performance is set according to policy values rather than empirical studies of student achievement.

The gain score is another simple and transparent method for estimating teacher performance (Rivkin, Hanushek, and Kain, 2005; Harris and Sass, 2006). The approach takes the difference in students' developmental scale scores for adjacent grades, and the average gain among students assigned to each teacher is a measure of teacher performance for the year. Gain scores are further adjusted to control for student characteristics and other factors that might distort the contribution of teachers to student achievement. The model is

$$G_{ijkt} = x_{ikjt} \beta^G + \delta_j + \varepsilon_{ijkt} \quad (1)$$

where G_{ijkt} is the test score gain in reading or math for the i^{th} student taught by the j^{th} teacher in the k^{th} school in year t . Measured characteristics (the vector x) can include student background (race/ethnicity, gender, and eligibility for free/reduced lunch), as well as indicator variables for grade, year, and school. The parameter vector β^G shows the contributions of these measured factors to achievement gains. The vector δ shows average teacher effects on student test score outcomes. The gain score approach is an improvement on a simple measures of student status like average scale score and percent proficient, because the gain score approach accounts for the ability level of incoming students prior to the start of the school year.

In this report, we consider only a simple gain score approach that uses the average of the classroom gain scores without further adjustment as the estimated teacher effect. In terms of the model Equation 1, the approach used here is equivalent to excluding the student background variables x from the model.

Random effects models are an example from the class of models for the joint distribution of the repeated test score for each student and are a special case of a hierarchical linear model (Raudenbush and Bryk, 2002). Other examples from this class that have been applied to estimating teacher effects include fixed effects model and mixed models such as the layered model of the Tennessee Value-Add Assessment System (Sanders, Saxton, and Horn, 1997) or the variable persistence model (McCaffrey et al., 2004, Lockwood et al., 2007).

In our random effects model, student achievement is based on a four-way error component where the four components consist of a student-specific effect, a teacher-specific effect, a school-specific effect, and a year-specific effect (Abowd, Creecy, and Kramarz, 2002; Abowd, Kramarz, and Margolis, 1999; Andrews, Schank, and Upward, 2004; Andrews, Schank, and Upward, 2005). The year-specific effect is easily incorporated into time-varying factors affecting achievement. The dependent variable is the student test score, T_{ijt} , observed for student i with teacher j , in school k at time t . Separate test specifications are estimated in reading and math.¹¹ The formal model is

$$T_{ijkt} = x_{ikjt} \beta^R + \theta_i + \psi_j + \gamma_k + \varepsilon_{ijkt} \quad (2)$$

where x_{ikjt} is a vector of measured student, teacher, school, and time (year dummy variables) characteristics¹²; β^R is a parameter vector linking these factors to test scores in the random effects model. The student, teacher, and school effects are represented by θ_i , ψ_j and γ_k . The last component of the model is a random error term, ε_{ijkt} , that is orthogonal to all other effects in the model. The main goal here is to separate the effects of teachers from the background of students that they are assigned and from schools where they are stationed.

In our model, the student, teacher, and school effects in Equation 2 are random variables with mean zero and unspecified variance which is estimated from the data. The statistical model estimates these components using heterogeneity in the data at each level of aggregation. One key assumption of the model is that, after implicitly controlling for students' scores in other years, the heterogeneity in scores at the teacher level is attributable to teacher effects rather than unequal distribution of students among classrooms. A second key assumption of this approach is that the composite error term is uncorrelated with measured characteristics in the model. If either assumption is violated, then the estimated teacher effects from this model are biased even for large samples of students.

We use a relatively simple random effects model for this study. Other models are more complex, with additional parameters for the persistence of teacher effects on their students' subsequent test scores or variables to account for peer effects and other factors (Lockwood et al., 2007; McCaffrey et al., 2004). Alternatively, some researchers have used fixed effects specifications that include indicator variables for individual students, teachers and schools

¹¹ The test scores in reading and math are standardized to have mean zero and standard deviation of one. The standardization was done by grade in the random effects analysis.

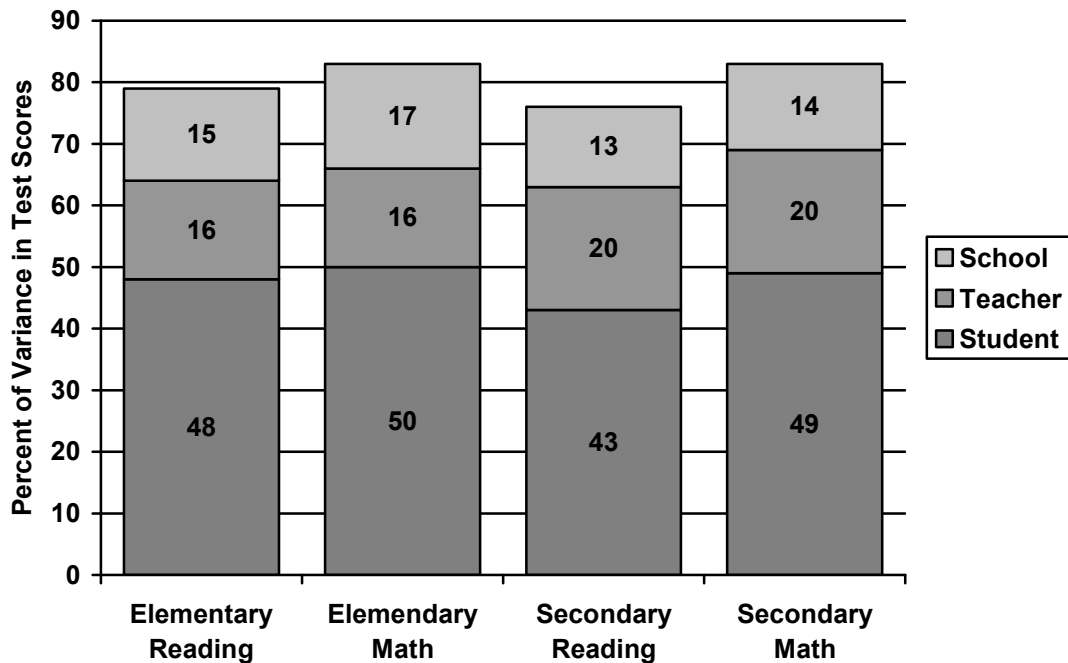
¹² The dataset has limited information on student characteristics. The model includes measures of student race/ethnicity (Black, Hispanic, or other), gender, and eligibility for free/reduced school lunch.

(Rivkin, Hanushek, and Kain, 2005; Clotfelter, Ladd, and Vigdor, 2006). Other models used fixed effects with gain scores (Rivkin, Hanushek, and Kain, 2005; Clotfelter, Ladd, and Vigdor, 2006; Harris and Sass, 2006; Koedel and Betts, 2006). Although estimated effects can be sensitive to model specification, the random effects model we present demonstrates some important issues when considering measures of teacher performance. Moreover, a fixed effects specification provided very similar results to our random effects estimates.

How Does Student Achievement Vary Across Students, Teachers, and Schools?

Figure 4.1 presents a decomposition of the variability in student test score into the variability within classrooms, among classrooms within schools, and between schools. As is almost always found in achievement data for students in the United States, a very large proportion of the variance is within classroom, with smaller proportions between classrooms within a school and between schools. This distribution of the variability in scores has implications for the estimation of teacher performance. The within classroom variability serves as a source of sampling error in estimated effects that can lead to instability in the estimated effects. In particular, as described in Lockwood, Louis, and McCaffrey (2002), the ratio of within classroom variance to the variance in teacher effects determines the precision of ranking and our ability to distinguish among teacher performance with statistical confidence.

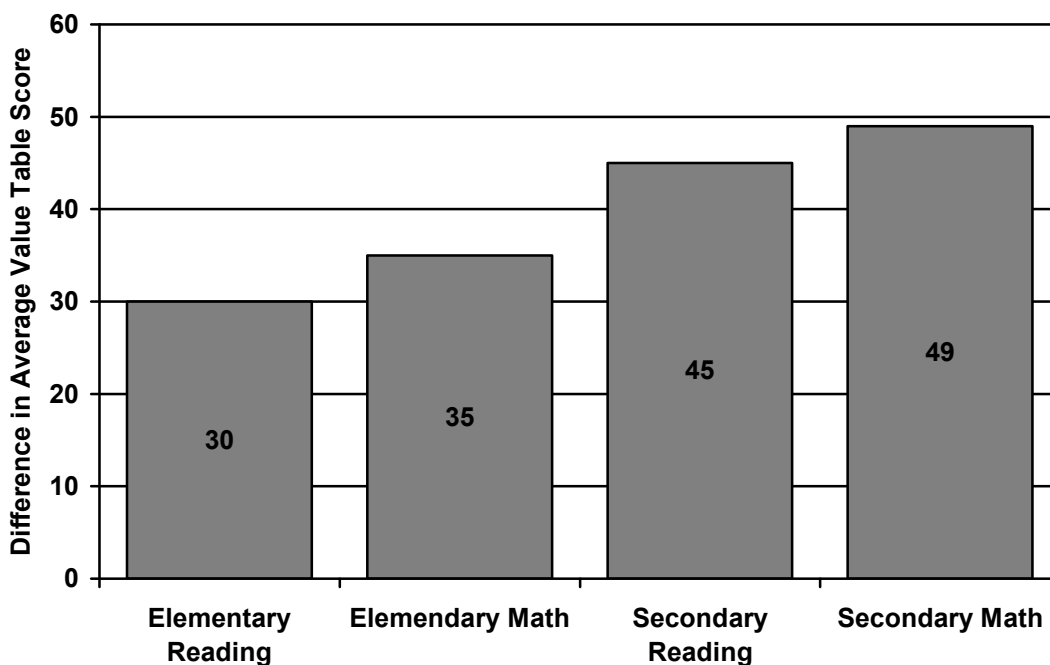
Figure 4.1. Decomposition of Test Score Variance Across Students, Teachers and Schools.



How Much Variability is There Among Estimated Performance Measures?

Figure 4.2 shows how value table scores differ over the range of teachers in Pinellas' schools. The scores are based on the value table illustrated in Table 3.1. The results show that higher ranking teachers have substantially better outcomes than lower ranking teachers. For example, the average value table score for an elementary math teacher is about 35 points. Teachers at the 25th percentile of value table scores score 77 points, meaning that 25 percent of teachers have value table scores of less than 77. At the 75th percentile, the score is 112. The interquartile difference (i.e., the difference between teachers at the 25th and 75th percentiles) ranges from 30 points in elementary reading to 49 points in secondary math.

Figure 4.2. Difference in Value Table Scores for Teachers at the 25th Percentile and 75th Percentile of the Distribution



The value table metric is difficult to interpret since scores depend in nonlinear ways both on where students start and on how much they improve. Suppose we compare two elementary math teachers with all level 3 (partially successful) students. One of the teachers is at the 25th percentile because she maintains 64 percent of students at level 3 while 36 percent decline to level 2. This classroom performance would produce a value table score of 77 ($0 \times 0.36 + 120 \times 0.64$). At the 75th percentile, the other teacher maintains 93 percent of students at level 3 and only 7 percent decline to level 2, so the teacher earns a score of 112. This differential effect of the two

teachers is consistent with a gap of 35 points for elementary math. Another way to calibrate this difference is to compare it to the variability among students. In elementary school math, the standard deviation in value scores for individual students is about 30 points. Thus the interquartile range is large relative to the student variability and the spread among classrooms is quite substantial.

The gain score results also show considerable variation in performance among teachers. For elementary teachers, the interquartile range in both reading and math scores is over 200 points. This difference is about the size of a standard deviation in student gains. Secondary teachers have lower overall gain scores, and this corresponds with a smaller interquartile range of scores in both reading and math. This range of gains is about 40 percent of a standard deviation in the corresponding reading and math gain scores. The gain score results show large differences in gain scores especially for elementary school teachers.

Figure 4.3. Gain Score Differences for Teachers at the 25th Percentile and 75th Percentile of the Distribution

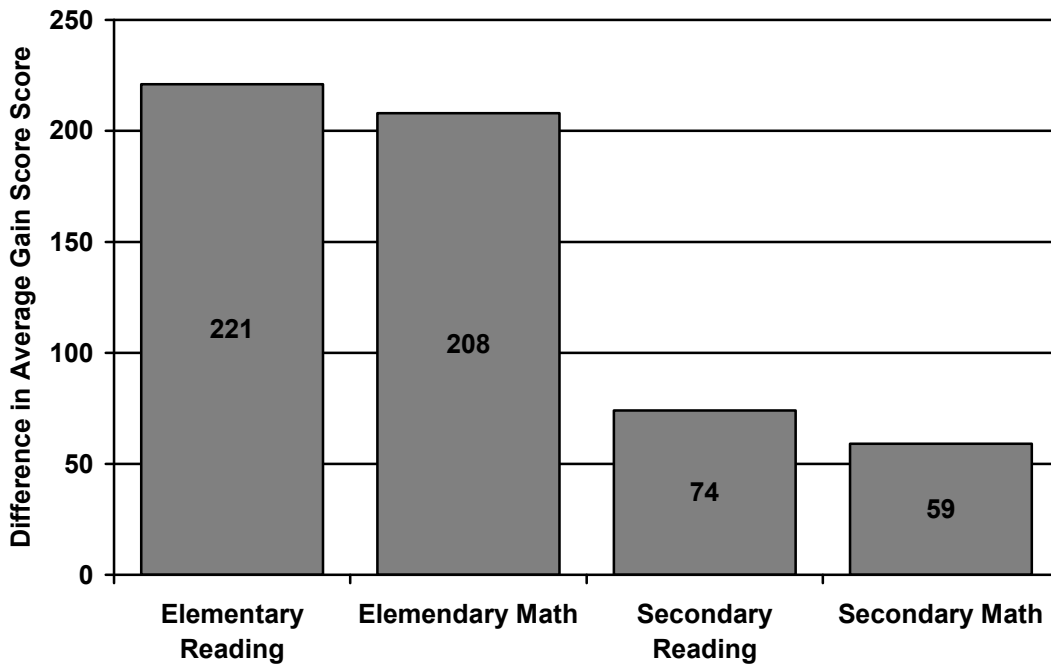
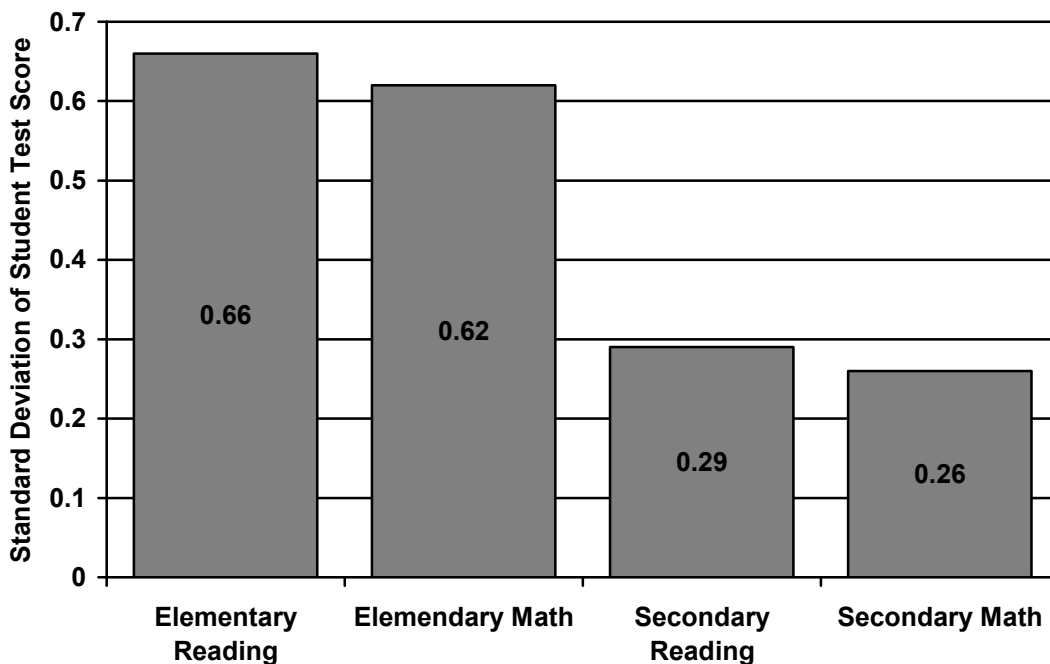


Figure 4.4 shows differences in the random-effects performance measures for teachers at the 25th and 75th percentile. The estimates are scaled by the standard deviation in student test scores. In elementary school, the interquartile range for teacher quality represents over 0.6 standard deviations in student test score performance. The interquartile range for secondary students is

only half as large, but the 0.3 difference across teachers is still substantial. For example, the test score gap between minority and white students is typically in the range of 0.8 to 1 standard deviation units. Thus, the differences among teachers cover a large portion of that gap.

Although all three methods suggest that there is considerable variability among teachers, the value table method differs from the other two methods in that the variability among secondary teachers is slightly greater than the variability among elementary teachers from value tables, but the variability among elementary teachers is substantially greater than the variability among secondary teachers for the other methods. One of the possible sources of this difference is that, as described below, value tables fail to remove differences among student background variables, and these are strongly differentiated across secondary teachers due to tracking and student course selection.

Figure 4.4. Difference in Random-Effects Estimates for Teachers at the 25th Percentile and 75th Percentile of the Distribution

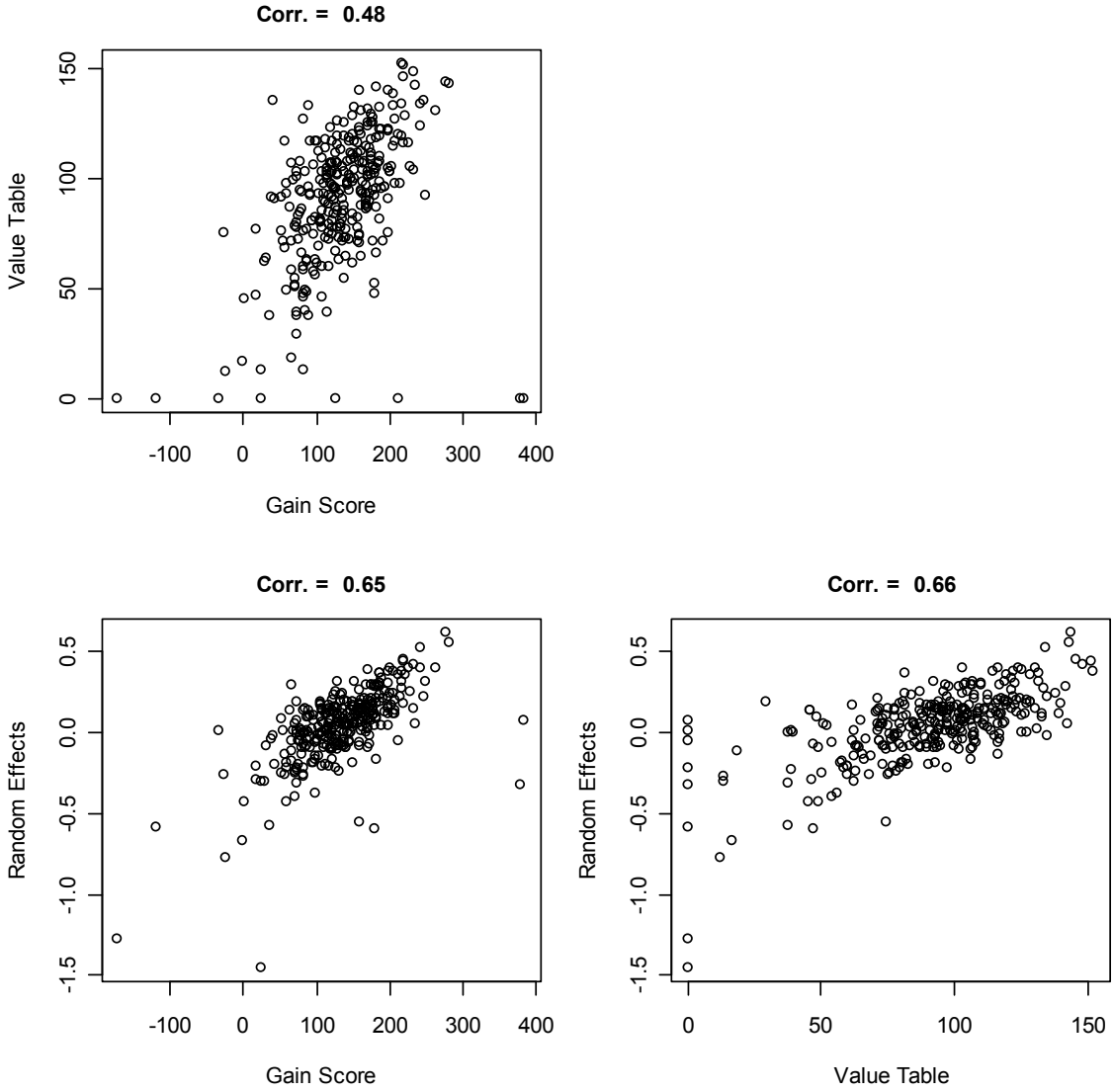


The results in Figures 4.2, 4.3, and 4.4 are consistent with those of Rivkin et al. (2006), Harris and Sass (2006), and Koedel and Betts (2007). These studies also find large overall effects of teacher differences on student achievement.

How Similar are Teacher Rankings from Alternative Performance Measures?

Although each performance measure finds variability among teachers, the measures do not identify the same teachers at the top and bottom of the distribution. As shown in Figure 4.5, correlation between estimates from pairs of methods range from 0.48 to 0.66 for fifth grade mathematics teachers in 2006. As shown in the figures, there are a few outlying points in each comparison that might lower the correlation coefficients, but there is at best only modest agreement among the methods. For example, the STAR program proposed teacher bonuses for teachers who scored in the top 25 percent of their grade cohort. Although MAP is less specific about the bonus requirements, this cutoff is still likely to receive considerable attention. However, this classification is somewhat unstable across the different measures. Only 13% percent of 5th grade mathematics teachers would be classified as in the top quartile by both value tables and gain scores. Similarly only 16% of teachers would be classified as in the top quartile by both gain scores and random effects and only 14% would be so classified by both value tables and random effects. Thus, only about half to two-thirds of teachers classified as in the top-quartile by one measure would be so ranked by one of the alternatives. Administrators and policy makers must be aware that the choice of performance measure will have implications for the teachers who are awarded bonuses in any given year. Another concern for merit pay is the resiliency of the teacher performance in subsequent years. In the next subsection, we turn to this issue of the stability of estimated teacher effects over time.

Figure 4.5. Scatterplots of Alternative Performance Measures for 5th Grade Teachers in 2006.



Are Estimated Performance Measures Stable Across Time?

In this subsection we consider if the performance measures for a teacher from different years provide consistent information about the teacher’s performance relative to the overall performance of teachers. Differences across time in a teacher’s performance measure result from differences in the teacher’s true performance across time, measurement errors in the test, and from the sampling errors in each year’s estimate. Regardless of its source, the concern is that instability in performance measures across time might make it challenging for teachers and administrators to respond to the measures and might reduce the credibility of the measures and thus limit their value as an incentive mechanism. Hence, when considering these measures, stability of the measures across time is of particular concern.

Value Table Score

Value table rankings for individual teachers are variable from year-to-year. Table 4.1 shows that the secondary teachers with high value table scores in one year are likely to have high scores in other years as well. The correlation is weaker for elementary school teachers, where the year-to-year correlation is much smaller between 2005 and 2006 than for other years.

Table 4.1. Correlations Between Estimated Teacher Value Table Scores Across Years

a) Elementary Reading				c) Secondary Reading			
	2004	2005	2006		2004	2005	2006
2004	1			2004	1		
2005	0.81	1		2005	0.73	1	
2006	0.54	0.19	1	2006	0.61	0.69	1

b) Elementary Math				d) Secondary Math			
	2004	2005	2006		2004	2005	2006
2004	1			2004	1		
2005	0.81	1		2005	0.80	1	
2006	0.70	0.51	1	2006	0.71	0.70	1

Table 4.2 shows the stability quartile rankings of teachers based on value tables across year. Only 29 percent of elementary reading teachers in the top quartile in 2005 repeat this teaching performance in the next year. About 19 percent of the “best” reading teachers in 2005 are in the worst reading quartile in 2006. The top quartile of elementary math teachers has more stability, with 48 percent remaining in the top quartile for the next year.

Table 4.2. Quartiles for Teacher Value Table Scores in Successive School Years

a) Elementary Reading					c) Secondary Reading				
Quartile in 2006					Quartile in 2006				
Quartile in 2005	1	2	3	4	Quartile in 2005	1	2	3	4
1 (Worst)	35	22	22	22	1 (Worst)	60	23	10	6
2	21	32	28	19	2	19	42	28	11
3	15	26	22	37	3	11	22	42	25
4 (Best)	19	22	30	29	4 (Best)	0	6	24	69
b) Elementary Math					d) Secondary Math				
Quartile in 2006					Quartile in 2006				
Quartile in 2005	1	2	3	4	Quartile in 2005	1	2	3	4
1 (Worst)	48	26	16	10	1 (Worst)	57	27	11	5
2	26	36	22	16	2	19	40	33	8
3	16	20	32	31	3	7	22	45	26
4 (Best)	6	13	33	48	4 (Best)	2	10	25	64

In contrast with elementary teachers, value table scores of secondary teachers are fairly stable from year-to-year. Table 4.2 shows that about two-thirds of reading and math teachers scoring in the top quartile in 2005 are also in the top group in 2006. Similarly, about 60 percent of teachers in the bottom quartile remain there in the second year.

Gain Scores

The gain scores are more stable over time than the value table scores. Table 4.3 shows correlations across year are higher than 0.7 for most categories. The correlations for secondary reading are lower, but they are still over 0.6.

As expected from the correlations, the gain score results in Table 4.4 show that teachers are likely to remain in the same quartile ranking from year-to-year. The probability of a teacher remaining in the top quartile for a second consecutive year ranges from 66 percent for secondary reading to 79 percent in elementary reading.

Table 4.3. Correlations Between Estimated Teacher Gain Scores Across Years

a) Elementary Reading				c) Secondary Reading			
	2004	2005	2006		2004	2005	2006
2004	1			2004	1		
2005	0.74	1		2005	0.63	1	
2006	0.72	0.92	1	2006	0.67	0.66	1

b) Elementary Math				d) Secondary Math			
	2004	2005	2006		2004	2005	2006
2004	1			2004	1		
2005	0.75	1		2005	0.71	1	
2006	0.78	0.93	1	2006	0.82	0.80	1

Table 4.4. Quartiles for Teacher Gain Scores in Successive School Years

a) Elementary Reading					c) Secondary Reading				
	Quartile in 2006					Quartile in 2006			
Quartile in 2005	1	2	3	4	Quartile in 2005	1	2	3	4
1 (Worst)	64	21	13	1	1 (Worst)	57	25	11	7
2	23	40	28	8	2	24	41	24	11
3	10	33	40	17	3	10	28	41	20
4 (Best)	3	5	13	79	4 (Best)	6	8	21	66

b) Elementary Math					d) Secondary Math				
	Quartile in 2006					Quartile in 2006			
Quartile in 2005	1	2	3	4	Quartile in 2005	1	2	3	4
1 (Worst)	80	17	1	2	1 (Worst)	69	22	5	3
2	13	79	4	4	2	21	56	19	5
3	0	11	61	29	3	4	26	58	11
4 (Best)	1	1	29	69	4 (Best)	7	2	16	76

Random Effects Estimates

The random effects results in Table 4.5 show that estimated teacher performance measures vary somewhat more from year-to-year for the random effects estimator than the other methods. In elementary reading, the table shows that the correlation between teacher effects in 2004 and 2005 is 0.5 as compared with 0.6 between 2005 and 2006. The correlations vary somewhat between years and subjects at the elementary and secondary levels. This evidence suggests that a teacher identified as good in one year is more likely to be rated highly in another year, but the tendency is not overwhelming.

Table 4.5. Correlations Between Estimated Random Teacher Effects Across Years

a) Elementary Reading				c) Secondary Reading			
	2004	2005	2006		2004	2005	2006
2004	1			2004	1		
2005	0.52	1		2005	0.70	1	
2006	0.44	0.61	1	2006	0.54	0.64	1

b) Elementary Math				d) Secondary Math			
	2004	2005	2006		2004	2005	2006
2004	1			2004	1		
2005	0.46	1		2005	0.65	1	
2006	0.49	0.71	1	2006	0.52	0.59	1

Again, although the top 25 percent cutoff for bonus is not a hard and fast rule, it is useful to assess whether the same teachers would likely earn bonuses each year using this rule, or whether the bonus recipients would vary considerably from year-to-year.

Table 4.6 examines the stability of teacher effect across quartiles from one year to the next. The results show that 42 percent of elementary reading teachers in 2005 are in the top reading quartile in 2006. The persistence is somewhat stronger in math where 51 percent of the elementary teachers remain in the top quartile in the next year. The results for secondary teachers are stronger with over 60 percent of teachers in the top quartile remaining there in the next year.

Table 4.6. Random Teacher Effect Quartiles in Successive School Years

a) Elementary Reading					c) Secondary Reading				
Quartile in 2006					Quartile in 2006				
Quartile in 2005	1	2	3	4	Quartile in 2005	1	2	3	4
1 (Worst)	49	18	14	19	1 (Worst)	53	26	14	7
2	24	30	24	22	2	26	38	24	12
3	18	30	31	21	3	14	27	36	23
4 (Best)	8	20	30	42	4 (Best)	4	10	19	66
b) Elementary Math					d) Secondary Math				
Quartile in 2006					Quartile in 2006				
Quartile in 2005	1	2	3	4	Quartile in 2005	1	2	3	4
1 (Worst)	57	25	8	10	1 (Worst)	51	19	21	9
2	21	35	26	18	2	21	36	35	9
3	9	27	33	31	3	14	22	29	35
4 (Best)	5	12	32	51	4 (Best)	4	14	20	62

The results for other teacher effects quartiles are mixed. In each group, the worst quartile (like the best) is rather stable. The trend in the intermediate quartiles is inconsistent across groups. About 30 percent of teachers remain in the second and third quartile in next year, but teachers are about equally likely to improve to a higher quartile as to decline to a lower one.

As noted above, sampling error and measurement error both contribute to the variability in estimated teacher performance. As discussed in the next subsections, the methods have differing levels of sampling error and this can contribute to the difference in their stability of their estimates across time. Also, the methods might not adequately control for student background variables. Student backgrounds are relatively stable across years and increase the variability among teachers within a year. Hence methods that do not adequately control for background variables are more stable across time.

The results here suggest difficulties in implementing merit pay for teachers based on FCAT student achievement scores. The models show moderately high stability of measured teacher performance over time for some measure and groups, but the stability is only modest in some years and for some groups. Financial awards to teachers based on these models would likely result in substantial differences in the so-called “best” teachers from year-to-year. Without

metrics to clearly define teacher performance over time, the effectiveness of the bonuses in encouraging teacher performance could be diminished

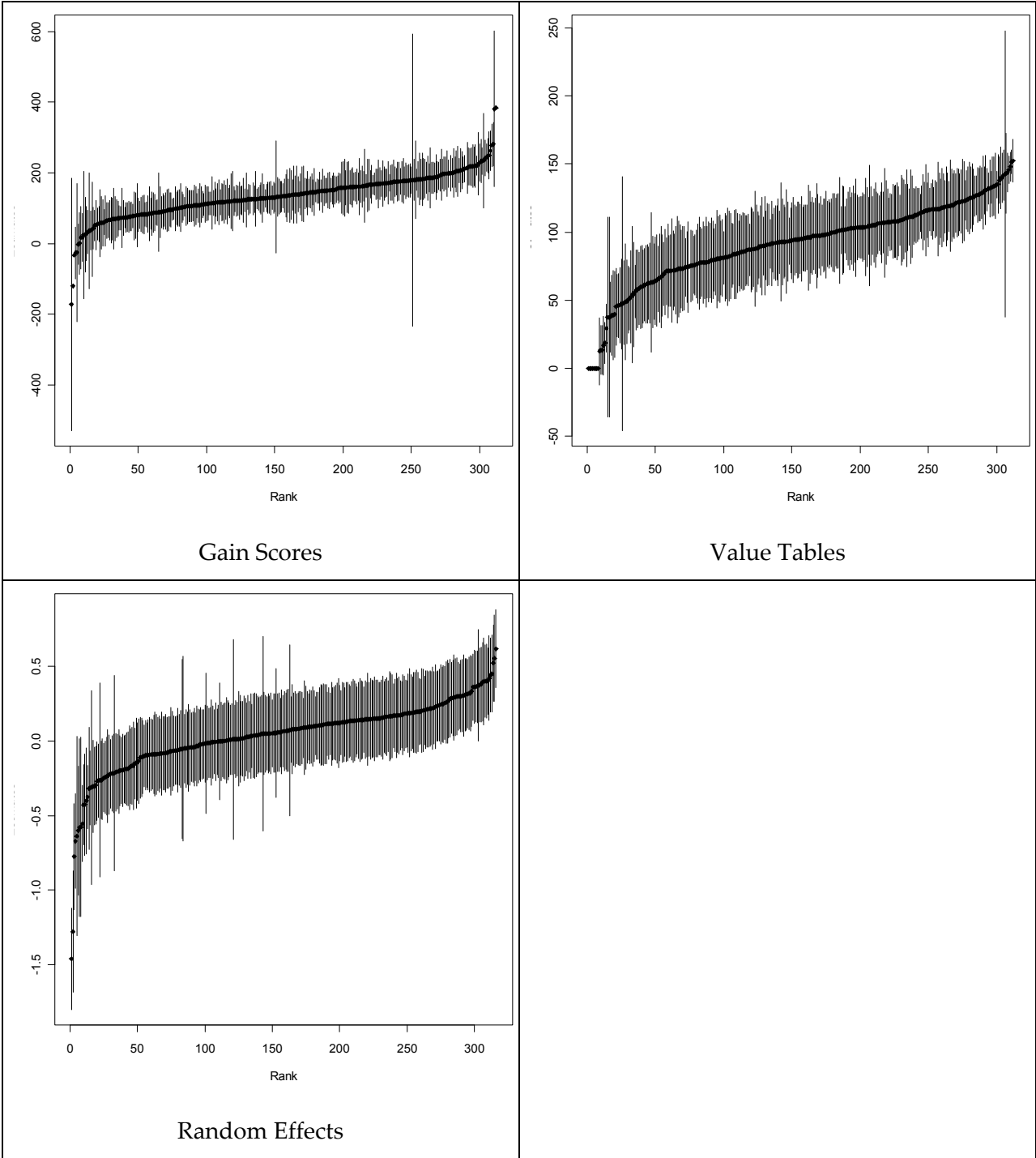
How Much Do Sampling Errors Affect Teacher Performance Measures?

One of the sources of the year-to-year differences in teacher effects is the sampling error due to the idiosyncratic characteristics of the students in a teacher's class each year. Because many teachers teach a moderately small number of students, this variability in students' outcomes can result in substantial variability in estimated annual performance measures. Figure 4.7 shows the uncertainty due to sampling error in estimated performance measures for 5th grade mathematics teachers in 2006. In each panel, the teachers are sorted by their estimated performance measure from lowest to highest. The estimates are plotted as a solid dot. The 95% confidence interval for the estimates (the interval that is likely to contain the "true" value for 95% of potential samples) is shown by the vertical lines. The top left panel presents gain score estimates, the top right panel presents value table estimates, and the bottom left panel presents the random effects estimates.

In all cases the sampling errors are large. One gauge of the magnitude of the sampling errors is that 64% of the estimated effects for 5th grade math teachers from the random effects model were greater than zero but only 11% of the 95% confidence intervals were entirely greater than zero. Hence for over 50% of the teachers, the sampling errors are sufficiently large that we cannot say with confidence that the true effect is positive even though the estimate is greater than zero.

As shown in the figure, the gain scores can have very large sampling errors for some teachers resulting in wide confidence intervals. This is because individual gain scores can be highly variable and because the requirement that the students have prior year test scores can result in very small samples for some teachers. Because random effects estimators can use data from students with incomplete test score data and "shrink" estimates with from small classes toward zero, they tend to have narrower and more uniform confidence intervals than gain scores. In this example, the value-tables also appear to generally have slightly narrower confidence intervals than either of the other methods. This will not always be the case. In particular, with more years of data on students (i.e., secondary students) the random effects estimates would be expected to have smaller sampling errors.

Figure 4.7. Estimate Teacher Performance Measures for 5th Grade Teachers in 2006 with 95% Confidence Intervals



For all three performance measures, sampling errors constitute a large fraction of the variability among the estimators across teachers. For example, the average variability due to sampling error compared to the overall variance in the estimated teacher performance measures are 0.31 for gains, 0.24 for value tables, and 0.38 for random effects. The result is that with each method, differentiating among teachers with a single year's worth of data will be difficult.

The values for our evaluation seem to suggest that value tables might provide the most efficient estimates because the sampling error is small relative to overall variability. However, this comparison of ratios is somewhat misleading for value tables because some teachers have value table values of zero and that probably underestimates the true sampling error in their estimates. Also, the variability in the performance measures can be a function of the unadjusted heterogeneity among classrooms in student background variables. That is, the teacher performance measures can vary among teachers, in part, because of differences among the students they teach. Even though the measures are meant to remove these influences from the estimates of teacher performance, none do so perfectly, and the variability among teachers for each of the estimators is due in part to confounding and the variability in the background characteristics of the students in their classes. Importantly, the greater the variability due to uncorrected confounding, the smaller the relative size of the sampling error. Because we do not fully know what proportion of the variability in the various performance measures is from bias, we need to be careful when comparing the ratios of sampling error to overall variability.

Do Performance Measures Reward Teachers for the Students They Teach?

As just noted, the measures do not necessarily remove the influences of student background from estimates of the teachers' performances and the relative sizes of the biases of the different measures have not been fully studied. In general, random effects models can remove much of the bias under the right circumstances and assuming a large number of tests (Lockwood and McCaffrey, 2007). Lockwood and McCaffrey (2007) report from simulation studies that possibly as few as five tests can reduce the bias to very small values, and other empirical research suggests that as few as three tests can be sufficient to greatly mitigate the bias (Sanders, 2006). Gain scores can also remove bias, but under relatively restrictive conditions (Lockwood and McCaffrey, 2007). Properties of value table measures have not been studied.

With empirical data, we cannot determine the bias since we do not know the teachers' true performance levels. However, strong correlations between performance measures and the classroom average of prior achievement is a signal that the performance measures has not fully controlled for student background variables and that teachers of historically poor-performing students will not have the same chances for rewards as teachers of historically high-performing

students. We find that for 5th grade mathematics teachers, the gain score measures correlate -0.18 with classroom average prior achievement in 2006 and -0.45 in 2005. For value tables, the correlation with classroom average prior achievement is 0.62 in 2006 and 0.50 in 2007. For the random effects measures the correlation is 0.32 and 0.31 for 2006 and 2005 respectively.

For the value tables the correlation with classroom average prior achievement is very high, and teachers teaching classes with higher incoming achievement will be more likely to be rewarded by a bonus program. It is possible that this high correlation reflects true differences in teacher performances that are correlated with student backgrounds; however, the other measures do not show similar relationships with prior achievement. Thus, it is likely that this is a source of systematic error that is in conflict with the goals of creating a fair and equitable system and could limit the value of this performance measure.

Alternative scorings for growth in the value table are possible and they might result in measures that are not so strongly related to the students' prior achievement; these might be explored. Also, the other properties of value tables (transparency and explicit valuation of specific outcomes) might make them useful even if the correlation with prior achievement exists. However, more extensive investigations of the properties of the value table measures are clearly warranted.

We also estimated these correlations for grade 8 mathematics teachers and find that the values tended to be even more extreme. The gain score measures correlate -0.42 with classroom average prior achievement in 2006 and -0.60 in 2005. For value tables the correlation with classroom average prior achievement is 0.75 in 2006 and 0.83 in 2005. For the random effects measures the correlation is 0.68 and 0.78 for 2006 and 2005 respectively. The large values likely result from large heterogeneity in the grouping in secondary mathematics in order to accommodate different curriculums such as pre-algebra, algebra, or more advanced classes.

Summary

FLDOE had listed desired attributes of its pay-for-performance system, which included fairness, ease of calculation, transparency, flexibility with regard to subject and grade levels, and a focus on specific educational goals and values. Our analysis demonstrates that meeting these goals in practice will be challenging for any measure.

It is clear that value tables and the simple average gain models meet the requirements of ease of calculation and transparency. The random effects method does not these requirements. The value table can also be tuned to focus on specific educational goal and values.

However, value tables clearly fail to provide fair and equitable measures of performance because as construed in the STAR plan the measures will less likely reward teachers of historically poor-performing students and more likely reward teachers of historically high-performing students. The other methods are less likely to be correlated with the student characteristics.

The large sampling errors and instability of the measures across years when based on a single year's worth of data could result in rewards being distributed to teachers who are not truly highest performing and lack of continuity in award across time. Incorrectly awarding teacher bonus can hardly be considered fair. Moreover, inconsistency across years is likely to diminish the face validity of the measures among teachers and limit their ability to incentivize behavior.

Because the measures are at best moderately correlated with each other, the choice of metric is likely to reward different teachers. Different measures will have different strengths and weaknesses, and these need to be considered carefully when choosing among the measures. Average gains are simple and transparent, but are sensitive to scaling of scores, cannot use incomplete data, have large sampling errors, and likely to be correlated with student background variables.¹³ Value tables are similar. They are transparent, easy to implement, and can be tuned to value certain outcomes. However, they cannot use data from students without prior year scores and are strongly correlated with student background variables. The random effects methods are not transparent, able to use incomplete data and with more years of score likely to have smaller sampling errors and be less correlated with student background variables than the alternative methods. However, as our example shows, with two years of prior scores these methods produced estimated teacher effects with similar statistical properties to the other simpler alternatives, even though the alternative methods all identify different teachers as high performing. Thus, every method will be imperfect; the designers of the performance pay system should carefully evaluate the alternative and consider the likely consequences of alternative methods for their teachers and their data.

5. OTHER TECHNICAL CONCERNS FOR MERIT PAY SYSTEMS

The previous section presented an empirical investigation of the properties of three alternate performance measures and clearly demonstrated the complexities and tradeoffs associated with measuring performance using one of these methods. This section examines three additional issues that arise when using student achievement data to measure teacher performance as part of a pay-for-performance program. First we discuss factors that make performance measures

¹³ Incomplete and missing data are discussed further in Section 5.

imperfect measures of teacher effectiveness at improving student achievement and the possible implications this has for performance-related pay systems. Next we discuss some concerns for implementing the systems, and finally we discuss the potential for inflation of the performance measures without equivalent gains in student achievement when measured more broadly. These issues and concerns are well-documented in the literature, but what is less well understood is that the impact of these factors differs across different performance measures and as such, attention needs to be paid to them when designing and selecting performance-based systems.

Performance Measures are not Measures of Teacher Effectiveness

One of the most pressing challenges of incentive programs is the measurement of the output (Dixit 2002; Eberts, Hollenbeck and Stone 2002; Burgess et al. 2001). The goal of teacher performance-related pay is to compensate teachers who are better performing according to the established criteria, typically high levels of student achievement. However, various sources of error mean that performance measures are limited measures of teachers' effectiveness at improving student achievement. We discuss these below.

Sampling Errors

As described in the previous section, a very large source of error in estimated performance measures is sampling error that results from idiosyncratic attributes of the students in a teacher's class in any given year. Because sampling errors are larger for teachers with small classes than teachers with large classes, sampling errors can result in differential treatment of teachers with different class sizes. For example, assuming that the true distribution of teacher effectiveness is equal for teachers with small and large classes, rewarding teachers on the basis of rankings or exceeding a predetermined threshold will tend to favor teachers of small classes over teachers of large classes. On the other hand, decision rules for allocating rewards that account for the estimates of performance and their errors, such as rules based on statistical significance, would tend to favor teachers of large classes over teachers of small classes who are equally high performing.

As shown in the previous section, different methods for estimating performance measures are more or less efficient at reducing sampling errors. The implications of sampling errors will depend on the decision rules, and careful evaluation of the performance of alternative measures in the context of a system's rewards program is advisable prior to choosing any one method.

The best way to reduce sampling error is to include information from more students. This can be accomplished by pooling estimates across years, for example, using a three- year average of performance measures rather than a measure from a single year. Different number of years

might be used for teachers with large and small classes to try and make the sample errors more comparable across these groups.

Confounding Effects of Uncontrolled Student Background Variables

Also as discussed in the previous section, the performance measures might confound teacher effectiveness with differences in the background characteristics of the students they teach. For example, the value table measures were highly correlated with students' prior achievement and would appear to favor teachers teaching classes with high levels of prior achievement and /or students from higher income families.

Empirical evidence and theoretical results suggests that complex methods such as random effects models that control for multiple prior year test scores can remove the potential biasing effects of individual student background variables (Lockwood and McCaffrey, 2007; Sanders, 2006; Ballou, Sanders, and Wright, 2004). Thus, these methods might be preferable to simpler methods such as average gain scores or value tables on the basis of removing potential confounding. However, other considerations such as transparency might favor the simpler methods.

However, even complex statistical methods have their own limitations. Theoretical results and simulation studies show that when student populations are stratified so that students with different potential for achievement growth do not share common teachers, even the best available methods can provide biased estimates of teachers' effects (Lockwood and McCaffrey, 2007; McCaffrey et al., 2004).¹⁴

If teachers teaching in different strata or disjoint subsets of the population are to be compared in a performance system, it might be appropriate to monitor the awards to see if they differ across the subsets of the population and if the differences correspond to different characteristics of the population. For example, is a disproportionate share of the rewards going to teachers not in the Title 1 schools? If so, can additional data be collected to determine if this corresponds to true differences in performance? For example, teachers could be tracked over time to determine if teachers who change assignments between the groups demonstrate similar performance in both groups or change performance in a direction that might suggest errors. Although such data

¹⁴ McCaffrey et al (2005) provide an example where teachers in whole-school Title 1 schools tend to have lower performance measures than other teachers in a large urban school district and very few students cross between these two groups of schools. In such a situation it is impossible to know for certain if the teachers in the Title 1 schools are poorer performing or if the performance measures are biased.

would not provide perfect evidence about the quality of measures, they would be an important component of any evaluation.

Alternatively, the system might limit comparisons to teachers teaching similar students. Some states, Ohio for example, classify schools on the basis of students they serve. Teachers might be compared within these classifications. This might remove the potential for bias, but it might create additional problems for the system, such as unequal treatment of students.

Moreover, differences in student populations do not occur only between schools but they can also occur within schools. In many schools, especially secondary schools, students are tracked by their prior performance and might not share classes with students of differing prior achievement. In addition, students in different classes may be exposed to different materials and have access to differential resources. Also, the alignment between the test and curriculum may vary across courses, which affects the extent to which teachers' efforts will lead to improved test scores. These differences will influence the estimates of teacher effects (Braun, 2005).

Models using multiple prior test scores to estimate teacher performance can again mitigate some of the potential biases from tracking, but the potential for bias will continue to exist. Monitoring the allocation of bonuses, tracking teacher performance, and using alternative measures are three approaches to dealing with such potential bias.

Issues Arising from the Use of Achievement Tests as an Outcome

Performance measures based on student achievement data are meant to measure teachers' contributions to student achievement as opposed to other attributes of students or other aspects of teacher performance. As Gratz (2005) put it, one assumption for linking student achievement to teacher incentives is that "student achievement can be assessed with sufficient rigor, breadth, validity, and reliability that it can be used to make decisions about teacher pay" (2005: 574). However, measuring student achievement and growth consistently and precisely is at best difficult, if not infeasible. In this section, we discuss scaling and errors in the achievement testing and their implications for performance measures.

Scaling of the Test. Students' responses to test items are combined into scores and scaled in ways that are designed to allow interpretation of the scores. For example, one goal of scaling is that differences in score points anywhere along the scale imply equal differences in achievement. Longitudinal data from multiple school grades requires methods to compare scales from different grade-level testing. One method for facilitating such comparisons is to place all scores on a single developmental scale that spans multiple grades.

Psychometricians have raised concerns about whether such scales meet the assumptions of the statistical models and inferences people make about growth (Schafer, 2006, Rigney and Martineau, 2005, Schafer and Twing, 2005, McCaffrey et al., 2003). For example, there is concern that the intervals between score points might not have the same interpretation at all points of the scales. A difference of 100 points at grade 4 might not measure the same difference in achievement as a 100 point difference at grade 5 because they might be based on tests with different content (Schafer, 2006). Moreover, there is no single way to scale tests across grades, and different scales can change the nature of the test scores and interpretations about growth (McCaffrey et al, 2003).

The statistical methods used in estimating teacher performance make assumptions about the scaling of the tests. Many methods (e.g., gain scores, fixed effects, growth modeling, the “layered” model of Sanders, Saxton, and Horn (1997)) are appropriate only under the assumption of a single vertical or developmental scale for scores across years of a student’s schooling. Random effects models and regression type adjustments are less dependent on such assumptions (Lockwood and McCaffrey, 2007).

A related issue is multidimensionality of tests. A mathematics or reading test does not measure only one skill or construct—each test measures multiple dimensions of achievement at each grade. The mix of dimensions measured can change across grades (Schafer, 2006; Rigney and Martineau, 2005). If statistical methods ignore the changing mix of measured constructs, then the resulting estimates of teacher performance may contain errors that favor some teachers over others (Martineau, 2006).

Lockwood et al. (2007) found estimated teacher effects to be highly sensitive to the test used for measuring achievement. This study estimated teacher effects for middle school math teachers separately using scores from two different subsets of the same test (mathematics procedures and mathematics problem-solving). They found that the two measures of a teacher’s performance were only weakly correlated. In simulations that assigned differential weights to the two scores, they found that teachers who were significantly above or below average with one set of weighted scores were rated no different from average with other sets of weights.

Errors in the test. No test measures without error the achievement it is designed to measure. If students were tested with an alternative test form of the same content or on a different occasion they would have different scores. This type of error, commonly referred to as measurement error, contributes to the variability in student scores and in the estimates of teacher performance.

Measurement error makes a single prior year test score a weak proxy for student background variables, and performance measures that rely on a single test score to control for such factors (simple linear regression, for example) may be of limited value because they can be highly correlated with measures such as socio-economic status of the students (Sanders, 2006). The use of multiple test scores to control for student backgrounds (as is done in random effects models) can reduce this impact of measurement error.

As shown above, estimated teacher performance measures are likely to be sensitive to unique features of the tests used for assessing student achievement. Methods like random effects models might be least sensitive to measurement error and the potential shortcomings of developmental scales. However, all methods will be sensitive to the content of the tests. As a result, it is important to undertake a careful evaluation of the test and its relevance to the curriculum and content being taught by all teachers and to monitor the curriculum and course content of the classes taught by teachers rewarded for their performance. Such monitoring could determine the biases in the system and whether these biases align with the goals of policymakers.

Disentangling Effects of Earlier Teachers and Schools from Estimated Teacher Effects

One particularly difficult methodological issue is how to account for the contribution of prior teachers and schools to students' outcomes in the current year. This is particularly problematic when the interval between tests spans a large portion of two school years. Methods such as gain score adjustments or regression adjustments (analysis of covariance) will not be biased by the contributions of prior year teachers only under certain assumptions. For example, gain score methods will be free of the prior year teacher effects if these persist undiminished into all subsequent years of schooling, and covariate methods require schooling effects to decay at the same rate as student characteristics.

Some models allow for less stringent assumptions about the persistence of teacher effects (Lockwood et al., 2007; McCaffrey et al. 2004). The models can result in different inferences about teachers (Lockwood et al., 2007) than the simpler alternative models. However, these models are both difficult for people to understand and often difficult to implement, making them less desirable in a performance system. Given that little data exist to support a decision about the preferred estimation method, one approach that might be useful is for systems to choose one method but also estimate alternative measures. In this way, systems could determine the characteristics of teachers who are treated differently by the alternative measures and monitor the comparability of the alternative measures over time.

Effect of Timing of Test

Because tests span two school years, changes in achievement also include changes during summer recess. Several authors have demonstrated that changes in achievement over summer recess are related to student characteristics such as socioeconomic status and ethnicity (Alexander et al., 2001). However, a small simulation study conducted by McCaffrey et al. (2003) suggests the differences in summer gains across groups are sufficiently small to be unlikely to have significant impact on teacher performance measures. Adding both fall and spring testing could avoid problems related to the testing interval spanning multiple school years, but it could raise additional problems. First, additional testing is likely to be very unpopular. Furthermore, Linn (2000) suggests that fall-spring testing introduces other, perhaps larger, biases than the more common spring-to-spring testing. For example, large performance improvements by students between fall and spring could benefit teachers and this could create unwanted incentives for teachers to suppress fall scores.

Implementation Issues

Missing Data

Real-world longitudinal student achievement data inevitably contain incomplete student achievement records, both in terms of incomplete data for students who leave in mid-year or who switch classes, or missing data in terms of missing links between students and teachers. As McCaffrey et al (2003) point out, “The factors that contribute to missing links and missing test scores are common: students are mobile, with large proportions transferring among schools every year.” For example, McCaffrey et al (2005) showed that for a large urban district, only 20 percent of students who started in grade 1 had fully observed scores linked to teachers from grades 1 to 5.

Students can be missing test scores in the current year for several reasons. The most common reasons are that the student transfers out of the district (or state) prior to testing, is exempted from testing because of limited English proficiency or a disability, or in rare cases is habitually absent or misses testing for some other reason such as hospitalization.

For students who transfer out of the district, there is a question of whether or not these students provide valid information about their teachers’ performances. States have rules for including students in school performance measures such as Adequate Yearly Progress (AYP) and similar rules are needed when measuring teacher performance.

For those remaining students with missing scores, the options are limited. They can be excluded from the calculations, scores can be imputed for them based on other variables such as prior year scores and background variables, or the performance measure could be adjusted to account for the proportion of students tested. This last option might be considered unfair by

some teachers and constituents since teachers potentially have limited control over whether or not students transfer schools.

The issues are similar for students who miss testing because of absenteeism or other factors. Rules are required for which of these students should be considered eligible for contributing to the performance measure, and then methods accounting for these students will need to be developed.

Students who are missing scores because they are not eligible for testing are more problematic. Clearly teachers' performance with these students is important to consider; however, the school district has determined that the available tests are inappropriate for these students and cannot always provide an alternative assessment of achievement. The lack of an alternative achievement test will be particularly likely for district-developed tests or tests the district purchases. If alternative tests exist, such as a Spanish language form, then scores on these tests might be used. For example, value tables could be based on proficiency levels determined by the alternative test or a test given with accommodations for students with disabilities. Although there has been some research on the use of alternative achievement tests and testing conditions, the effect of combining scores from alternative achievement tests on the properties of the performance measures has not been studied.

Excluding students from the performance measures for any reason has the potential to create negative consequences for a performance-based pay system. The theory of incentives suggests that teachers might be motivated to shift resources away from students who will not provide test scores. There is no research on how teachers might behave toward such students, but there is clear evidence that schools have focused their energies on students near the cutoff of proficiency in efforts to meet the requirements of AYP (Hamilton et al., 2007; Booher-Jennings 2005).

Students with missing prior year scores are also candidates for being excluded from the performance measure calculations because these prior year scores are necessary to calculate measures like value tables and average gain scores. Students who miss testing tend to be lower-scoring students, and excluding these students from the testing could result in performance measures that favor teachers with more or less mobile students. Random-effects methods do not require every student to have a prior test score and could avoid some of the problems with the simpler methods. Furthermore, Lockwood et al (2004) and McCaffrey et al. (2005) show that the ordering of estimated teacher effects from complex methods like random effects models appears robust to missing data in some settings.

Definition of a Teacher's Students

If all students were taught by a single teacher in each subject, then that individual teacher would clearly be accountable for the student performance. However, this is often not the case. In many instances students have multiple teachers for the same subject. There are at least five reasons why a student can have multiple teachers for a tested subject: 1) the student is taking multiple courses in a subject; 2) the student switched courses midway through the school year; 3) the teacher for a course switched midway through a school year or two teachers share a course; 4) the student received special education or other supplemental services for this subject; and 5) achievement tests are not administered at the end of a school year, so instruction spans two teachers in consecutive school years. These issues are likely to affect all performance measures and will need to be addressed regardless of which statistical procedure is used.

To deal with these issues, what is needed is complete information on all students' courses and teachers, which is often hard to obtain because administrative data might have incomplete information on students' courses. Teachers are often the best recourse for correcting administrative data, but this can be time-intensive. Students taking multiple courses in the same subject are typically identified in administrative data. Sometimes the challenge is determining which courses provide instruction for a tested subject. For example, should a drama teacher be accountable for students' English/Language Arts test scores?

It is not clear what is the best approach for allocating teachers to students who receive instruction from multiple teachers. In Tennessee, teachers are assigned students for the fraction of instructional time they provide (Sander, Saxton and Horn, 1997). However, this weighting of teachers is predicated on the untested assumption that instructional time is equivalent to effort or share of the learning that takes place. Alternative weighting could be used and sensitivity of inferences about teachers to this assumption should be investigated.

Potential for Inflation of Performance Measures

Theoretical models such as the multi-task principal-agent model, developed by Holmstrom and Milgrom (1991), suggest that when multidimensional tasks, such as teaching, are measured by outcome measures that do not equally assess all aspects of performance, then performance monitoring systems will lead the "agent" (i.e., the teacher) to focus on that tasks measured by the performance metric and divert attention away from other tasks. Standardized tests, for instance, are likely to measure only some aspects of the teaching goals. Therefore, many analysts worry that linking teacher compensation to student test scores could divert teachers' efforts away from promoting curiosity and creative thinking towards basic skills tested on the exams (Hannaway 1992; Glewwe, Ilias and Kremer 2003; Kane and Staiger 2002). Also, teachers

might divert attention from subjects that are not tested and students whose performance is not considered or down-weighted. Research on the response of schools and teachers to the requirements of the federal No Child Left Behind Act shows that educators report focusing more attention on reading and mathematics than untested subjects and focusing their attentions on students just below proficiency (Hamilton et al., 2007; McCaffrey and Hamilton, 2007). Similarly, Glewwe, Ilias and Kremer (2003) found that teachers increased effort to raise short-run test scores in response to a teacher incentive program in Kenya, but that these gains were not retained after the end of the program.

Many researchers are concerned that other consequences of merit pay schemes can be more pernicious than teaching to the test. For example, schools and teachers may deliberately label lower-performing students as special education students or English language learners (ELLs) (who are normally exempt from standardized testing and/or allowed to take alternative assessments), retain them in grade, or at worst cause them to drop out in order to raise average scores on the exams (Glewwe, Ilias and Kremer 2003). Furthermore, alternative performance methods could create or exacerbate the potential for negative consequences. For instance, simple procedures like gain scores and value tables exclude students without prior test scores and this could motivate teachers to focus their attention away from these students.

A clear consequence of this type of narrowing of focus or “gaming” the testing system is that it could have negative consequences for student learning. It can result in score inflation, which can distort performances of teachers and limit the values of tests for assessing the achievement of students. Score inflation refers to increases in scores that do not reflect a commensurate increase in mastery of the domain. Such inflation can be caused by a wide variety of teacher behaviors, ranging from simple cheating to inappropriately narrowing focus on the content of the test at the expense of other material and/or teaching test-taking tricks. Several studies have shown that scores can become seriously inflated under high-stakes conditions (e.g., Jacob, 2002; Klein et al., 2000; Koretz and Barron, 1998; Koretz, Linn, Dunbar, and Shepard, 1991) and that score inflation can be much larger than meaningful gains in test scores. If teachers engage in different practices that promote score inflation this could result in differential inflation and distort rewards. If teachers see that behaviors that promote score inflation are rewarded, they could become distrustful of the bonus system, causing it to fail to provide the desired incentives.

Hamilton, McCaffrey and Koretz (2005) suggest that well-designed tests that broadly measure the domain of interest and use diverse items might reduce score inflation. Moreover, making all students’ scores count toward performance measures could help reduce the negative consequences of testing. Monitoring teachers’ responses to testing through surveys might also

allow for identifying limiting behaviors and modifications to the system to reduce those behaviors and the long-term negative impact from them.

6. SUMMARY AND CONCLUSIONS

Florida is at the forefront of a national debate on merit pay for teachers. In the past two years, the state legislature has passed reform legislation that would fund bonuses for teachers based on their classroom performance. Performance is measured in terms of student achievement on standardized tests as well as principal assessments of teachers.

The goal of merit pay is to target compensation towards specific student outcomes and use pay to leverage improvements in teacher quality. If compensation is tied specifically to successful outcomes, then teachers have a specific monetary incentive to improve those outcomes. Similarly, outcome-based pay should encourage the retention of proficient teachers and attract more high-skilled individuals to the teaching profession. The reform contrasts with traditional teacher compensation schedules that link pay with teacher inputs, primarily experience and educational achievement, without any direct linkages to classroom outcomes.

However, rewarding teachers for their productivity at increasing student learning may be challenging for several reasons. First, comprehensive learning outcomes for students are difficult to define and measure. The limitations in the achievement tests may make it difficult to accurately sort teachers on the basis of their students' achievement. Second, students are not randomly assigned to classes and classes tend to differ in terms of the students' preparation and prior achievement. Careful statistical controls will be necessary to disentangle the contributions of individual teachers from the prior background and achievements of students assigned to their classes; otherwise teacher performance measures may better reflect competencies of students assigned to specific teachers than the success of those teachers in improving student learning.

This study focuses on the usefulness of standard achievement tests in assessing teacher quality. Florida has mandated that 60 percent of teacher performance should be tied to student achievement test scores. Our analysis uses student-level data from a large school district in Florida to explore issues that arise when using statistical methods and student achievement test scores as measures of teacher performance. We consider three measures:

1. **Value tables**, which award points to students on the basis of their changes in proficiency levels from the prior year of testing; points are averaged across all students assigned to each teacher.

2. **Average gain scores**, which average the change in scale scores for all students assigned to a teacher.
3. **Random effects model**, which simultaneously controls for each student's scores at multiple time points to estimate improvements in student achievement in any given year and aggregate these estimates across all the students assigned to a teacher.

The three measures represent classes of models that might be used to assess teacher performance as part of a merit pay system.

Lessons from Comparing Alternative Measures

The analysis demonstrates important features of the statistical properties of the teacher performance estimates and highlights issues that require attention when designing merit pay systems. First, consistent with the literature, our study found that all three methods show substantial variation among teachers' estimated performances. Students assigned to teachers at the 25th percentile of each performance measure are likely have much smaller test score gains than are students assigned to teachers at the 75th percentile of the performance measure.

Estimated performance measures vary considerably from method to method, however, so bonus awards would be sensitive to which method is used to measure teacher performance. Only about half to two-thirds of teachers classified as in the top quartile by one method would be so ranked by one of the alternatives. These differences in simple metrics suggest that rewards for teacher are sensitive to the method, and systematic differences in the teachers who would receive awards might make one method preferable to another.

The estimated performance measures for each method are also somewhat unstable over time. For some methods and in some years, performance in one year is weakly correlated with performance in the subsequent year. This inter-temporal variability might generate confusion for teachers in how to respond to the measures. This confusion might generate mistrust in how bonuses are awarded and discourage teachers from modifying teaching practices or increasing effort to earn a merit bonus.

The performance measures have moderate to high correlations with the prior class achievement. This suggests that estimated performance for some measures is sensitive to the types of students assigned to teachers. Using the value table measure, for example, elementary school teachers are more likely to rank in the top quartile if they are assigned students with high prior achievement than if they are assigned students with lower achievement scores. Refinements of the value table or additional years of achievement data to use as controls might reduce the sensitivity of the performance measures to the mix of students assigned to teachers.

Apart from the issues identified in the empirical analysis, there are other technical concerns regarding the use of student achievement test scores to measure teacher performance. As we discussed above, performance measures are not true measures of teacher effectiveness because of sampling errors, confounding by student background variables; the scaling of the test, the timing of tests that typically span two school years and include changes during summer recess; missing data; accurate definition of a teacher's students; and the narrowed focus on measured outputs. The choice of performance measure can mitigate the problems associated with some of these issues. For other issues, however, the only recourse may be the recognition of the potential problems and the development of plans to monitor and respond to problems.

Policy Recommendations

Our results suggest serious challenges to using standardized test scores to measure teacher performance as part of a merit pay system. Ideally, a system would be transparent and easy to understand, since this would help policy makers and teachers to understand what exactly is needed to earn a bonus award. A system must also isolate the contributions of each teacher from the prior academic achievement of students assigned to each class. Policy makers should be wary of adapting performance measures without a thorough understanding of the properties of a particular measure. Errors in defining teacher merit will distort the incentive effects of bonuses and may distort the effort of teachers in promoting student learning.

Performance measures should be based on multiple years of data on students and teachers. Single-year measures of teacher performance are highly volatile given the relatively small numbers of students taught by a teacher in a given year and the idiosyncrasies of student assignments.

Districts should monitor the bonus awards and examine patterns of awards across teachers and schools. While high- or low-quality teachers might be disproportionately assigned to some grades, subjects, or schools, strong patterns in the data may also suggest that the performance measure is biased. For example, suppose that bonus incidence is higher for suburban schools than for central city schools. This finding might indicate that teaching performance is truly better at the suburban schools, but the finding might alternatively indicate that the measure of teacher performance is correlated with prior class achievement. Districts should be careful to adopt performance measures with good statistical properties, but they should also be proactive in searching out and resolving potential problems with the measure that is adopted.

Districts should also monitor the trends in student achievement scores overall and for various student groups (e.g., low- versus high-proficiency students). A key benchmark for the merit pay reform is whether overall achievement levels rise with the implementation of a merit pay

system. Disproportionate changes for some groups relative to others may indicate that the teacher reward system is placing undue emphasis in improvements for some segments of students.

Teacher performance should be based on growth in student achievement and not on the proficiency level of students in a teacher's class. Teachers who are assigned students with high prior achievement are much more likely to have high end-of-year proficiency scores than are teachers who are assigned students with low prior achievement. Simple proficiency measures will distort the contribution of an individual teacher to student learning and provide undue "merit" rewards for teachers who are assigned high proficiency students.

Our primary analysis has focused on measuring teacher performance using standardized test scores. However, many performance-related pay systems, including MAP, also include subjective evaluation of teacher performance in the award calculation. The literature on subjective evaluations suggests that these evaluations will be compressed, especially when the evaluations are part of a merit pay system. As a result, the evaluations may implicitly carry a small weight as compared with test-based performance measures and play little role in how bonuses are awarded. Districts should monitor evaluations to assess their effectiveness and role in the merit awards. If evaluations do uncover important differences among teachers, the ranking system should be adjusted to assure that these evaluations are weighted appropriately.

More research is needed on how classroom evaluations from principals and other observers correspond with performance measures based on student achievement results. Little is known about what specific classroom practices and strategies translate into test score outcomes. Similarly, teacher performance measures, like those used in a merit pay plan, might be related some teacher preparation courses or professional development programs. A better understanding of what factors contribute to better classroom success would help districts to hire and train better teachers and to encourage more effective classroom practices.

What are the prospects for merit pay? The traditional compensation schedule links teacher pay to educational background and experience—two factors that have weak to nonexistent relationships with classroom success. It is true that no merit pay system will meet all the challenges and develop a perfect measure of teacher performance. The key issue is whether the incentive and sorting effects of an admittedly imperfect merit pay system can improve the quality of the teacher workforce. We believe that piloting such systems and carefully monitoring their results is a valuable exercise.

REFERENCES

- Aaronson, D., L. Barrow, and W. Sander, *Teachers and Student Achievement in the Chicago Public High Schools*, Chicago, IL: Technical report, Federal Reserve Bank of Chicago, 2003.
- Abowd, J., R. Creecy, and F. Kramarz, *Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data*, Technical Paper 2002-06, U.S. Census Bureau. As of September 6, 2005: <http://lehd.dsd.census.gov/led/library/techpapers/tp-2002-06.pdf>
- Abowd, J., F. Kramarz, and D. Margolis, "High-Wage Workers and High-Wage Firms," *Econometrica*, Vol. 67, No. 2, 1999, pp. 251–333.
- Alexander, K., D. Entwisle, and L. Olson, "Schools, Achievement, and Inequality: A Seasonal Perspective," *Educational Evaluation and Policy Analysis*, Vol. 23, No. 2, 2001, pp.171-191.
- Andrews, M., T. Schank, and R. Upward, "Practical Estimation Methods for Employer-Employee Data," IAB Discussion Paper No. 3, 2004. As of September 6, 2005: <http://doku.iab.de/discussionpapers/2004/dp0304.pdf>
- Andrews, M., T. Schank, and R. Upward, "Practical Fixed Effects Estimation Methods for the Three-Way Error Components Model," 2005. As of September 6, 2005: <http://www.nottingham.ac.uk/economics/staff/details/ru/leed2.pdf>
- Angrist, Joshua, and Victor Lavy, "Does Teacher Training Affect Pupil Learning? Evidence from Matched Comparison in Jerusalem Public Schools," *Journal of Labor Economics*, Vol.19, No. 2, 2001, pp. 343-369.
- Asch, Beth J., "The Economic Complexities of Incentive Reform," in Robert Klitgaard and Paul C. Light, eds., *High-Performance Government: Structure, Leadership, and Incentives*, Santa Monica, CA.: RAND Corporation, 2005, pp. 309-342.
- Ballou, D., "Pay for Performance in Public and Private Schools," *Economics of Education Review*, Vol. 20, No.1, 2001, pp. 51-61.
- Ballou, D., "Value-Added Assessment: Lessons from Tennessee," in R. Lissetz ed., *Value Added Models in Education: Theory and Applications*, Maple Grove, MN: JAM Press, 2005, pp. 272-303.
- Ballou, D., W. Sanders, and P. Wright, "Controlling for Students' Background in Value-Added Assessment of Teachers," *Journal of Educational and Behavioral Statistics*, Vol. 29, No. 1, 2004, pp. 37-66.

- Barro, Jason, and N. Beaulieu, *Selection and Improvement: Physician Responses to Financial Incentives*, Cambridge, MA: National Bureau of Economic Research, Working Paper 10017, 2003.
- Booher-Jennings, J. "Below the Bubble: 'Educational Triage' and the Texas Accountability System," *American Educational Research Journal*, No. 42, Vol. 2, 2005, pp. 231-268.
- Braun, H., *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models*, Technical report, Princeton, NJ: Educational Testing Service, Policy Information Center, 2005.
- Burgess, S., B. Croxson, P. Gregg, and C. Propper, *The intricacies of the relationship between pay and performance of teachers: Do teachers respond to performance related pay schemes?* Bristol, UK: Centre for Market and Public Organization Working Paper, Series No. 01/35, University of Bristol, 2001.
- Center for Teaching Quality, *Performance Pay for Teachers: Designing a System that Students Deserve*, Hillsborough, NC: Center for Teaching Quality. As of July 23, 2007: http://www.teacherleaders.org/teachersolutions/TSexec_summary.pdf
- Clotfelter, C., and H Ladd, "Recognizing and Rewarding Success in Public Schools, " in Ladd, H., ed., *Holding Schools Accountable: Performance-Based Reform in Education*, Washington, DC: The Brookings Institution, 1996, pp. 23-63.
- Clotfelter, C., H. Ladd, and J. Vigdor, *How and Why Do Teacher Credentials Matter for Student Achievement?* Cambridge, MA: National Bureau of Economic Research, Working Paper 12828, 2006.
- Committee for Economic Development, *Investing in Learning: School Funding Policies to Foster High Performance*, Washington, DC: Committee for Economic Development, 2004.
- Dewey, J., T. Husted, and L. Kenny, "The Ineffectiveness of School Inputs: A Product of Misspecification?" *Economics of Education Review*, Vol, 19, No. 1, 2000, pp. 27-45.
- Dixit, Avinash, "Incentives and Organizations in the Public Sector: An Interpretative Review," *The Journal of Human Resources*, Vol. 37, No. 4, 2002, pp. 696-727.
- Eberts, Randall, K. Hollenbeck, and J. Stone, "Teacher Performance Incentives and Student Outcomes," *The Journal of Human Resources*, Vol. 37, No. 4, 2002, pp. 913-27.
- Florida Department of Education, *Star Technical Assistance Paper II*, August 2006. As of December 12, 2006 at http://info.fldoe.org/docushare/dsweb/Get/Document-3888/k12_06_115att.pdf).

- Florida Department of Education, *Using Value Tables to Determine Teacher Effectiveness in Florida*, 2006b. As of December 12, 2006 at <http://www.fldoe.org/PerformancePay/pdfs/ValueTable2005-06.pdf>.
- Figlio, David N., and L. Kenny, *Individual Teacher Incentives and Student Performance*, Cambridge, MA: National Bureau of Economic Research, Working Paper 12627, 2006.
- Glewwe, Paul, N. Ilias, and M. Kremer, *Teacher Incentives*, Cambridge, MA: National Bureau of Economic Research, Working Paper 9671, 2003.
- Goldhaber, D., and D. Brewer, "Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity," *Journal of Human Resources*, Vol. 32, No. 3, 1997, pp. 505-523.
- Gordon, Robert, T.J. Kane, and D.O. Staiger, *Identifying Effective Teachers Using Performance on the Job*, Cambridge, MA: The Brookings Institution, Discussion Paper 2006-01, 2006.
- Gratz, Donald B, "Lessons from Denver: The Pay for Performance Pilot," *Phi Delta Kappan*, Vol. 86, No. 8, 2005, pp. 569-81.
- Hamilton, L., D. McCaffrey, and D. Koretz, "Validating Achievement Gains in Cohort-to-Cohort and Individual Growth-Based Modeling Contexts," in Lissitz, R.L., ed., *Longitudinal and Value Added Models of Student Performance*, Maple Grove, MN: JAM Press, 2006.
- Hamilton, L., B. Stecher, J. Marsh, J. McCombs, A. Robyn, J. Russell, S. Naftel, and H. Barney, *Standards-Based Accountability under No Child Left Behind: Experiences of Teachers and Administrators in Three States*, Santa Monica, CA.: RAND Corporation, MG-589-NSF, 2007.
- Hannaway, Jane, "Higher Order Thinking, Job Design, and Incentives: An Analysis and Proposal," *American Education Research Journal*, Vol. 29, No. 1, 1992, pp. 3-21.
- Hanushek, Eric A., "The Trade-off between Child Quantity and Quality," *Journal of Political Economy*, Vol. 100, No. 1, 1992, pp. 84-117.
- Hanushek, Eric A., "Assessing the Effects of School Resources on Student Performance: An Update," *Educational Evaluation and Policy Analysis*, Vol. 19, No. 2, 1997, pp. 141-64.
- Hanushek, Eric A., John F. Kain, Daniel M. O'Brien, and Steven G. Rivkin, *The Market for Teacher Quality*. Cambridge, MA: National Bureau of Economic Research, Working Paper 11154, 2005.

- Hanushek, Eric A., John F. Kain, and Steven G. Rivkin, *Teachers, Schools, and Academic Achievement*, Cambridge, MA: National Bureau of Economic Research, Working Paper 6691, 1998.
- Hanushek, Eric A., John F. Kain, and Steven G. Rivkin, *Do Higher Salaries Buy Better Teachers?* Cambridge, MA: National Bureau of Economic Research, Working Paper 7082, 1999.
- Hanushek, Eric A., and S. G. Rivkin, *How to Improve the Supply of High Quality Teachers*, Paper prepared for the Brookings Papers on Education Policy, 2003.
- Harris, D. and T. Sass, *Value-Added Models and the Measurement of Teacher Quality*, Unpublished manuscript, 2006.
- Hassel, B., *Better Pay for Better Teaching: Making Teacher Compensation Pay off in the age of Accountability*, Washington, DC: Progressive Policy Institute, 2002.
- Hatry, H., J. Greiner, and B. Ashford, *Issues and Case Studies in Teacher Incentive Plans*, Washington, DC: Urban Institute Press, 1994.
- Hedges, L., R. Laine, and R. Greenwald, "Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential Inputs on Student Outcomes." *Educational Researcher*, Vol. 23, No.3, 1994, pp. 5-14.
- Heneman III, Herbert, Anthony Milanowski, and Steven Kimball, *Teacher Performance Pay: Synthesis of Plans, Research, and Guidelines for Practice*, CPRE Policy Briefs, RB-46. Philadelphia, PA: Consortium for Policy Research in Education, 2007.
- Hill, R., B. Gong, S. Marion, C. DePascal, J. Dunn, M. Simpson, "Using Value Tables to Explicitly Value Student Growth," Conference on Longitudinal Modeling of Student Achievement. Dover, NH: The Center for Assessment, 2005. As of July 24, 2007: http://www.nciea.org/publications/MARCES_RH07.pdf, 2007
- Holmstrom, Bengt, and Paul Milgrom, "Multi-Task Principal-Agent Analysis: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics and Organization*, Vol. 7, 1991, pp. 24-52.
- Jacob, B., *Accountability, Incentives and Behavior: Evidence from School Reform in Chicago*, Cambridge, MA: National Bureau of Economic Research, NBER Working Paper 8968, 2002.
- Jacob, B., and L. Lefgren, "Principals as Agents: Subjective Performance Measurement in Education," Working Paper Number RWP05-040, Cambridge, MA: Harvard University, 2005.

- Jacob, B., and L. Lefgren, "When Principals Rate Teachers," *Education Next*, Vol. 6, No. 2, 2006, pp. 58.
- Jacobson, S. L., "Monetary Incentives and the Reform of Teacher Compensation: A Persistent Organizational Dilemma," *International Journal of Education Reform*, Vol. 4, No. 1, 1995, pp. 29-35.
- Kane, Thomas J., and Douglas O. Staiger, "The Promise and Pitfalls of Using Imprecise School Accountability," *Journal of Economic Perspectives*, Vol. 16, No. 4, 2002, pp. 91-114.
- Klein, S. P., L. S. Hamilton, D. F. McCaffrey, and B. M. Stecher, *What do Test Scores in Texas Tell Us?* Santa Monica, CA: RAND, IP-202, 2002.
- Koedel, C., and J. R. Betts, *Re-Examining the Role of Teacher Quality in the Educational Production Function*, Unpublished Manuscript, 2006.
- Koretz, D., and S. I. Barron, *The Validity of Gains on the Kentucky Instructional Results Information System (KIRIS)*, Santa Monica, CA: RAND, MR-1014-EDU, 1998.
- Koretz, D., R. L. Linn, S. B. Dunbar, and L. A. Shepard, "The Effects of High-Stakes Testing: Preliminary Evidence About Generalization Across Tests," in R.L. Linn (chair), *The Effects of High Stakes Testing*, symposium presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April, 1991.
- Kreuger, Alan, "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, Vol. 114, No. 2, 1999, pp. 497-932.
- Ladd, Helen. "The Dallas School Accountability and Incentive Program: An Evaluation of Its Impacts on Student Outcomes," *Economics of Education Review*, Vol. 18, No. 1, 1999, pp. 1-16.
- Lashway, Larry, *Incentives for Accountability. ERIC Digest.*, Eugene, OR: ERIC Clearinghouse on Educational Management, 2001.
- Lavy, V, "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement" *Journal of Political Economy*, Vol. 110, No. 6, 2002, pp. 1286-1317.
- Lavy, V, *Performance Pay and Teachers' Effort, Productivity and Grading Effects*, Cambridge, MA: National Bureau of Economic Research, Working Paper 10622, 2004..

- Lazear, Edward P., "Salaries and Piece Rates." *Journal of Business*, Vol. 59, No. 3, 1986, pp. 405-31.
- Lazear, Edward P., "Performance Pay and Productivity," *American Economic Review*, Vol. 93, No. 5, 2000, pp. 1346-61.
- Lazear, Edward P., "Teacher Incentives," *Swedish Economic Policy Review*, Vol. 10, No. 2, 2003, pp. 179-214.
- Linn, R., "Assessments and Accountability," *Educational Researcher*, Vol. 29, No. 2, 2000, pp. 4-14.
- Locke, E., K. Shaw, L. Saari, and G. Latham, "Goal Setting and Task Performance: 1969-1980," *Psychological Bulletin*, Vol. 90, No. 1, 1981, pp. 125-152.
- Lockwood, J.R., D.F. McCaffrey, L.S. Hamilton, B. Stecher, V.-N Le, and F. Martinez, "The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures," *Journal of Educational Measurement*, Vol. 44, No. 1, 2007b, pp. 47-67.
- Lockwood, J.R., D.F. McCaffrey., L.T. Mariano, and C. M. Setodji Bayesian Methods for Scalable Multivariate Value-Added Assessment, *Journal of Educational and Behavioral Statistics*, Vol. 32, No. 2, 2007a, pp. 125-150.
- Lockwood, J.R., T. Louis, and D.F. McCaffrey, "Uncertainty in Rank Estimation: Implications for Value-Added Modeling Accountability Systems," *Journal of Educational and Behavioral Statistics*, Vol. 27, No. 3, 2002, pp. 255-270.
- Malanga, Steven, "Why Merit Pay will Improve Teaching," *City Journal*, Vol. 11, No. 3, 2001. As of July 24, 2007: http://www.city-journal.org/html/11_3_why_merit_pay.html
- Martineau, J., "Distorting Value Added: The Use of Longitudinal, Vertically Scaled Student Achievement Data for Growth-Based Value-Added Accountability," *Journal of Educational and Behavioral Statistics*, Vol. 31, No. 1, 2006, pp. 35-62.
- McCaffrey, D. F., and L. S. Hamilton, *Value-Added Assessment in Practice: Lessons from the Pennsylvania Value-Added Assessment System Pilot Project*, Santa Monica, CA: RAND Corporation, TR-506-CC, forthcoming 2007.
- McCaffrey, D. F., J. R. Lockwood, D. M. Koretz, and Laura S. Hamilton, *Evaluating Value Added Models for Teacher Accountability*, Santa Monica, CA: RAND, MG-158-EDU, 2003.

- McCaffrey, D.F., J. R. Lockwood, D. M. Koretz, T. Louis, and L.S. Hamilton, "Models for Value-Added Modeling of Teacher Effects," *Journal of Educational and Behavioral Statistics*, Vol 29, No. 1, 2004, pp. 67–101.
- McCaffrey, D. F., J. R. Lockwood, L. Mariano, and C. Setodji, "Challenges for Value-Added Assessment of Teacher Effects," in Lissitz, R., ed., *Value-Added Models in Education: Theory and Applications*, Maple Grove, MN: JAM Press, 2005.
- Mildovich, G., and A. Wigdor, *Pay for Performance: Evaluating Performance Appraisal and Merit Pay*, Washington, D.C.: National Research Council, 1991.
- Muralidharan, Karthik, and Venkatesh Sundararaman, *Teacher Incentive in Developing Countries: Experimental Evidence from India*. Unpublished manuscript, 2006.
- Murnane, Richard J., and David K. Cohen, "Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and a Few Survive," *Harvard Educational Review*, Vol. 56, No. 1, 1986, pp. 1-17.
- Murnane, R., and R. Olsen, "The Effects of Salaries and Opportunity Costs on Duration in Teaching: Evidence from Michigan," *Review of Economics and Statistics*, Vol. 71, No. 2, 1989, pp. 347-52.
- Murnane, R., and R. Olsen, "The Effects of Salaries and Opportunity Costs on Duration in Teaching: Evidence from North Carolina," *Journal of Human Resources*, Vol. 25, No. 1, 1990, pp. 106-24.
- National Center for Education Statistics, *Digest of Education Statistics, 2005*, Washington, DC: National Center for Education Statistics, Institute of Education Science, US Dept. of Education, 2005. As of December 21, 2006:
<http://nces.ed.gov/programs/digest/d05/index.asp>
- National Center on Education and the Economy, *Tough Choices or Tough Times: The Report of the New Commission on the Skills of the American Workforce*. San Francisco: Jossey-Bass, 2007.
- Nye, B., S. Konstantopoulos, and L. V. Hedges, "How Large are Teacher Effects?" *Educational Evaluation and Policy Analysis*, Vol. 26, No. 3, 2004, pp. 237-257.
- Odden, A., and C. Kelley, *Paying Teachers for What They Know and do: New and Smarter Compensation Strategies to Improve Schools*. Thousand Oaks, CA: Corwin Press, 1997.
- Odden, A., C. Kelley, H. Heneman, and A. Milanowski, *Enhancing Teacher Quality Through Knowledge and Skills-Based Pay*, Philadelphia, PA: Consortium for Policy Research in Education, 2001.

- Prendergast, C., "The Provision of Incentive in Firms," *Journal of Economic Literature*, Vol. 37, No.1, 1999, pp. 7-63.
- Raudenbush, S.W. and A.S. Bryk, *Hierarchical Linear Models: Applications and Data Analysis, Second Edition*, Thousand Oaks, CA: Sage Publications Inc., 2002..
- Rigney, S., and J. Martineau, "NCLB and Growth Models: In Conflict or in Concert?" in Lissitz, R.L., ed., *Longitudinal and Value Added Models of Student Performance*, Maple Grove, MN: JAM Press, 2006, pp. 47-81.
- Rivkin, Steven G., Eric A. Hanushek, and John F. Kain, "Teachers, Schools, and Academic Achievement," *Econometrica*, Vol. 73, No. 2, 2005, pp. 417-58.
- Rockoff, J., "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *AEA Papers and Proceedings*, Vol. 94, No. 2, 2004, pp.247-252.
- Sanders, W. L., "Comparisons Among Various Educational Assessment Value-Added Models," presented at The Power of Two--National Value-Added Conference, Columbus, Ohio, October 16, 2006. As of July 24, 2007:
www.sas.com/govedu/edu/services/vaconferencepaper.pdf
- Sanders, W., A. Saxton, and B. Horn, "The Tennessee Value-Added Assessment System: A Quantitative Outcomes-Based Approach to Educational Assessment," in Millman, J., ed., *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluational Measure?* Thousand Oaks, CA: Corwin Press, Inc., 1997, pp. 137-162.
- Schafer, W., "Growth Scales as an Alternative to Vertical Scales," *Practical Assessment, Research and Evaluation*, Vol. 11, No. 4, 2006, pp.1-6.
- Smith, M. S., and J.A. O'Day, "Systematic School Reform," In Fuhrman, S., and B. Malen, eds., *The politics of Curriculum and Testing*, New York: The Falmer Press, 1991, pp. 233-267.
- Southern Regional Education Board, *Teacher Salaries and State Priorities for Education Quality: A Vital Link. Educational Benchmarks 2000 Series*, Atlanta, Georgia, 2000.
- Springer, M., "Social Science Theory and Performance-Based Teacher Pay," Nashville, Tennessee: Vanderbilt University, 2007. As of July 9, 2007:
<http://www.ncctq.org/webcasts/payforteach/SpringerNCCTQ.pdf>
- Zingheim, P., and J. R. Shuster., "How are the New Pay Tools Being Deployed?" *Compensation and Benefits Review*, Vol. 27, No. 4, 1995, pp. 33-39.