TRADEOFFS IN THE USE OF VALUE-ADDED ESTIMATES OF TEACHER EFFECTIVENESS BY SCHOOL DISTRICTS

by

Andrew David Baxter

A dissertation submitted to the faculty of
the University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Public Policy

Charlotte

2011

Approved by:

____
Dr. Jennifer L. Troyer

____
Dr. Robert K. Godwin

____
Dr. Suzanne M. Leland

____
Dr. Louis H. Amato

ABSTRACT

ANDREW DAVID BAXTER. Tradeoffs in the use of value-added estimates of teacher effectiveness by school districts.
(Under the direction of DR. JENNIFER L. TROYER)

A new capacity to track the inputs and outcomes of individual students' education production function has spurred a growing number of school districts to attempt to measure the productivity of their teachers in terms of student outcomes. The use of these value-added measures of teacher effectiveness is at the center of current education reform. This study links the technical work of academic researchers with the implementation and policy considerations school districts are likely to face in incorporating value-added measures in their teacher evaluations. First, I assess the choices the district must make in specifying one or more models. Then, I evaluate three potential threats to the validity of the inferences from value-added data: the influence of prior inputs in a student's education production function, ceiling effects in the test instrument, and the sorting of students to teachers. I end by considering to what extent value-added measures could be useful to districts in monitoring the distribution of effective teachers to its students and personnel decisions such as retention and compensation.

ACKNOWLEDGEMENTS

I have benefited from the wisdom and counsel of many good folks over the course of this study. I would like to thank the members of my dissertation committee—Jennifer Troyer, Ken Godwin, Suzy Leland, Ted Amato and David Swindell—for their helpful comments and guidance. I also learned a great deal through conversations with Jon Fullerton, Sarah Cohodes, Eric Taylor, Tom Kane and Doug Staiger of the Center for Education Policy Research at Harvard. I am appreciative, too, for conversations and help from colleagues at Charlotte-Mecklenburg Schools, especially Jason Schoeneberger, Samantha Salvador, Dana Lockwood, Tom Tomberlin and Cara Jokell. I would like to thank my wife, Susan Gover, for her patience and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

MOTIVATION

Spurred by a new wave of education reform marked by an emphasis on effective teachers as the primary lever through which school districts can raise the achievement of their students, school districts are increasingly looking at ways to measure the effectiveness of their teachers. Previous waves of education reform focused on student assignment (Angrist & Lang, 2004; Godwin, Leland, Baxter, & Southworth, 2006; Guryan, 2004; Reber, 2007); school choice (Godwin & Kemerer, 2002), the distribution of resources (Hanushek, 1986, 1996); and curricular reforms and management structures (Ladd, Hansen, & National Research Council (U.S.). Committee on Education Finance, 1999). Yet many of those reforms incurred significant costs for school districts while failing to achieve for sustained benefits for students (Hanushek, 2003, 2006; Hanushek & Rivkin, 1997).

Before the advent of standardized testing, and its mandated use prompted by the No Child Left Behind Act (NCLB) ("NCLB," 2001), districts did not have outcomes of student achievement by which they could measure teacher effectiveness across teachers and schools. Instead, the districts had to rely upon what they took to be signals of teacher effectiveness such as credentials or the type of qualifications easily listed on a resume. District recruiters sought out experienced teachers with advanced degrees and national board certification. The institutionalization of standardized testing led to an observable outcome by which districts can measure one dimension of student learning—achievement that can be measured by changes in standardized test scores. Districts are discovering what academic researchers have been noting, although not necessarily unanimously, for the last 10-15 years—credentials, or more broadly, qualifications, are not effective predictors of post-hire performance (Rockoff, Jacob, Kane, & Staiger, 2008). Returns to experience diminish greatly after a teacher's first 4-5 years of teaching. Advanced degrees do not seem to predict effectiveness with students (Kane, Rockoff, & Staiger, 2006). Neither do special certifications such as National Board Certification (Cantrell, Fullerton, Kane, & Staiger, 2008). In fact, one often cited study of North Carolina teachers suggests that these types of qualifications explain only 3% of the variation in effectiveness among teachers (Goldhaber & Brewer, 1997).

The lack of pre-hire predictors of teacher effectiveness has been exacerbated by the lack of post-hire predictors available to school districts. Most districts rely upon classroom observations of teachers by principals. Typically a principal or her surrogate will observe the teacher 3-4 times over the course of a year in a single class period. The principal completes a rubric, and at the end of the year, the teacher is placed in a fairly broad category—satisfactory versus unsatisfactory or an equivalently crude categorization. These observations contain a nontrivial amount of noise. They are subject to a lack of inter-rater reliability, potentially biased by the principals' preferences for some teachers over others, and they are spasmodic.

Beyond these sources of noise is a larger one that lurks in the background—a notion of professionalism that simultaneously holds that teaching is a profession and that everyone can do it. There is a well-documented egalitarian strain within education that is largely unwilling to make distinctions among teachers about their performance (see, for example, Millman & Darling-Hammond, 1990). As a result, approximately 98% of the nation’s teachers are rated as satisfactory or better (Weisberg, Sexton, Mulhern, & Keeling, 2009).

So districts face a situation in which they have little ability to know before the hire how effective a teacher will be, and after the hire, they are receiving reports that 98% of their teachers are satisfactory despite stark evidence that many children are not learning. In states with collective bargaining, once teachers are hired, it is very difficult to fire them. Whether or not the district negotiates with a teacher union, the district often must abide by the tenure regulations that govern teacher employment. For example, in North Carolina, teachers are granted “career status” after their fourth year of teaching. This career status all but guarantees continued employment for the rest of the teacher’s career. In NC, approximately 90% of eligible teachers pass through their probation period into career status (Goldhaber & Hansen, 2010). If the teacher is subpar, then the net present value of the decision to grant career status is extremely high in absolute value terms both in terms of the costs in compensation and pensions as well as social welfare in terms of losses to student achievement (Staiger & Rockoff, 2010). A bottom quartile teacher who is granted tenure would likely teach 26 additional years before retiring with 30 years of service. Assuming an elementary teacher instructs 20 students a year, the district is subjecting 520 students to subpar teaching, assuming, as the evidence suggests, that the teachers are not likely to improve significantly after their first 4-5 years. The economic costs to the students in foregone wages can be quite large. A recent simulation study by Hanushek (2010) estimates that a teacher one standard deviation above the mean teacher would generate $400,000 in additional combined income for a classroom of twenty students. Another well-publicized study that reanalyzed the earnings and postsecondary outcomes of students from the Project STAR study came to similar conclusions (Chetty et al., 2010). Students bear the costs of a district's inability to identify the effectiveness of its teachers.

So do teachers. Without any strong signal of effectiveness, teachers are compensated on schedules laden with incentives to strengthen their qualifications—staying in teaching to gain more years of experience, enrolling in advanced degree programs, obtaining special certifications. These are the only ways in which teachers can improve their compensation and stay in teaching. Yet even with these potential gains, an entrant into the teacher workforce will likely face a highly regulated pay scale in which no matter how effective they are, they are not likely to reach their earning peak for 30 years (North Carolina Department of Public Instruction 2010). Candidates who are more likely to be effective will also be more likely to select out of a profession in which their effectiveness will not be recognized. This selection effect is likely to be large (Lazear, 2003). We have evidence that this rigid salary schedule, coupled with expanding opportunities in other labor markets, has spurred an exodus of many potentially effective teachers from the field (Corcoran, Evans, & Schwab, 2004).

The institutionalization of standardized tests coupled with the creation of longitudinal data systems that can now trace a student's academic trajectory from pre-kindergarten through graduate school has created the potential to measure teacher performance across schools and districts. A literature largely dominated by economists has evolved over the last two decades to attempt to identify teacher effectiveness from changes in students' standardized test scores from one year to the next. Using a production function framework of student achievement, economists seek to parse the teacher's contribution to these changes from the observable and unobservable characteristics of the student, the student's classroom or peers, and the student's school.

Much of the current literature is marked by statistical arguments over identification strategies, especially relating to endogeneity concerns. The debate is highly technical. Often missing are the implications of one alternative versus another for a school district that wants to add a value-added measure of teacher effectiveness to its evaluation of teacher performance. Academicians often spend large parts of their papers explaining how they arrived at their sample. Typically they are working with large administrative data sets that cover a state or a large urban district. They have large numbers of student observations. They are not interested in identifying a particular teacher's effect so much as they are identifying the variation in teacher effectiveness as a whole.

Key Questions Addressed By the Study

This study will serve as a bridge between the academy and district office. It addresses three key questions faced by school districts considering the use of the value-added measures of teacher effectiveness. In this study, I evaluate some of the technical considerations that districts must consider in choosing one or more value-added models for human capital decisions such as recruitment, retention and compensation. The district's concerns in choosing a value-added model certainly include wanting to get the econometrics right. Yet they also include implementation concerns that are not as urgent for the econometrician.

The three lines of inquiry are as follows:

1. What are the benefits and costs of various value-added models in terms of the identification and specification of teacher effects?

2. How serious are two often-cited threats to the validity of value-added estimates—ceiling effects in the test instrument and the sorting of teachers to students? What can be done to mitigate the risks they pose?
3. Are value-added estimates suitable for use in (a) considerations of a district's equitable allocation of its resources across students, and (b) high-stakes personnel decisions?

To answer the first question, I replicate a number of the models that are current in the literature, models that are designed to account for a particular concern, (e.g., measurement error). I estimate the models and then the correlations of their teacher effects. Where they are different, I try to explore why. In assessing the benefits and costs, I evaluate the alternatives on their statistical properties (e.g., comparisons of model fit) as well as the constraints they pose on the district (e.g., requiring three years of student data rather than two). I arrive at a preferred model that maximizes the number of teachers a district can evaluate in a way that minimizes misidentification of an individual teacher's effects.

To address the second question, I test the preferred model against two threats to its validity—ceiling effects in the test instrument and the sorting of students to teachers. Ceiling effects could bias the estimates of teachers with high proportions of students who scored at the high end of the distribution in the prior year. The sorting of students to teachers could also bias the estimates of teachers even with robust controls for sorting on observables. I test for the presence of such sorting and evaluate the magnitude and probability of its impact on estimates from the preferred model.

In answering the third question, I use the effect estimates from the preferred model to examine two possible policy uses of the estimates. First, I use the estimates to explore the distribution of high and low value-added teachers to students both across and within schools. Second, I evaluate the usefulness of the estimates for high-stakes decisions such as tenure or performance-based compensation.

Section 2 reviews prior research. Section 3 details the data. Section 4 discusses the methods. Section 5 reviews the results. Section 6 concludes.

LITERATURE REVIEW

Over the last twenty-five years, a robust literature on the specification and use of value-added models has reshaped education policy. The emergence of the value-added measure in policy discussions has been fueled by the creation of large longitudinal administrative data sets that allow researchers to utilize panel data techniques to analyze the effect of various educational policies. Extensive data sets in Texas (Rivkin, Hanushek, & Kain, 2005), Florida (Harris & Sass, 2006), Chicago (Jacob, Lefgren, & Sims, 2008), North Carolina (Clotfelter, Ladd, & Vigdor, 2006; Goldhaber, 2007) and New York City (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2008b) have been used repeatedly by groups of researchers first to test various specifications and then to assess potential uses of value-added estimates.

The availability of the data has enticed economists who have both the econometric tools to exploit the large longitudinal data sets and a framework—the production function—for understanding the effect of educational inputs on student achievement. The work of (Hanushek, 1979; Lazear, 2001; Todd & Wolpin, 2003, 2004) posed education as a technology in which current and past inputs from the student, family and school, including the student's unobserved endowment influence the student's cumulative acquisition of knowledge. This work, particularly that of Todd and Wolpin, illuminated the assumptions under which causal inferences about the relative effect of different inputs could be made.

Although the production function included but was not limited to human capital inputs, the role of teachers quickly became the center of the research program. The work of Eric Hanushek and his collaborators (1986, 1996, 2003, 2005) sparked many economists to shift the research agenda away from measuring the effects of inputs such as programs or curricula or the allocation of financial resources. Decades of increased investment in schools in terms of public expenditures spent on reducing class sizes or new curricula seem to have done little to increase student achievement (Hanushek, 2003). Although not undisputed (Card & Krueger, 1996; Greenwald, Hedges, & Laine, 1996; Krueger, 2003), this line of research has changed fundamentally the policy questions asked by districts.

At the same time, the new administrative data sets have allowed researchers to uncover the effect of teachers on student achievement, at least the type of student achievement that can be measured by a standardized test. In a recent review of the literature, Hanushek and Rivkin (2010b) summarize the consistent finding across studies of significant differences among teachers in their effect on student achievement. These studies, from different researches using different data sets, consistently show that that a one standard deviation difference in teacher effectiveness results in changes in student achievement of 0.11-0.36 student-level standard deviations. To put the magnitude of these effects into perspective, the Tennessee STAR experiment yielded effect sizes of 0.2 student level standard deviations for decreasing class sizes from 22 to 15 students (Krueger, 2003).

Table 2 summarizes the canonical studies that formed the first wave of value-added research. These value-added studies were structured similarly. They compared different specifications of the models to test their sensitivity to changes in the specification. The work centered on assumptions about the decay of prior inputs, the form of teacher effects (random or fixed), and ways to handle the unobserved endowment of the student. Many of these papers were not testing hypotheses about policies, but rather specifying different possibilities for the actual estimate and then gauging the impact of those different specifications on the results.

Although this literature did not result in a consensus on the proper specification of value-added models, there have been few papers since these that focus on the overall specification of the models. A consensus seems to have emerged that controlling for student, classroom, or school characteristics is less a technical issue and more of a policy issue (for an exception, see Ballou, Sanders, & Wright, 2004). Similarly, the decision about whether to model school heterogeneity through school fixed effects is also cast as a policy decision concerning how the estimates will be used (e.g., for cross-school comparisons). A consensus has emerged that the teacher estimates must be shrunken of their sampling variation through either one of several variations of empirical Bayes estimators or by using the correlations of the adjacent year teacher effects (I will discuss this technique in the Methods section).

The one specification issue where there seems to be ongoing disagreement is how to deal with measurement error in the test. An oft-cited recent paper argues for the use of dynamic panel data estimators to deal adequately with the potential of measurement error in the pretest to render the teacher effect estimates inconsistent (Andrabi, Das, Khwaja, & Zajonc, 2009). Similarly Boyd et al. (2008a) argue for an alternative approach to dealing with the measurement error that exploits the covariance structure of the tests.

As considerations about the fine points of the specifications have subsided, a newer literature is examining threats to validity of the teacher effect estimates that would pertain across most specifications. The most controversial of these critiques is that introduced by Jesse Rothstein in two recent papers (Rothstein, 2009, 2010). In these papers, Rothstein uses a North Carolina administrative dataset to test the strict exogeneity assumptions of the value-added approaches. He conducts a falsification test such as that suggested by Todd and Wolpin (2003) and finds that the distribution of effects of the teachers at time t+1 on students at time t are almost as large as those of the teacher at time t, leading him to conclude that sorting of teachers to students renders estimates of the teacher effects biased and inconsistent. His critique has gained great traction in the policy community (for example, see Baker et al., 2010).

Among economists, Rothstein's critique generated a vigorous response. Koedel and Betts (2009a) replicate his results and show how adding additional years of data to in calculating the teacher's effect mitigates the bias completely. Others such as Hanushek (2009) have shown that the impact of the teacher at t+1 on the student at time t is actually a function of tracking of students to teachers on the basis of test scores.

A recent experimental study by Kane and Staiger (2008) conducted an experiment in which pairs of teachers in Los Angeles were randomly assigned to classrooms within their schools. Kane and Staiger used the strength of the relationship of the difference in the value-added of the teachers pre- and post-random assignment as a measure of the robustness of the specification against the threat of sorting to students. They found that pre-experimental models that included peer effects but excluded student fixed effects yielded results that were not statistically different from the experimental differences in the teacher's value-added. Thus, in their relatively small study, they were able to conclude that sorting of students was not biasing the estimates of teacher effects. It is unclear whether the findings would generalize to other districts.

A second line of critiques returns to the assumptions of value-added models that are rooted in production function framework (Ishii & Rivkin, 2009; Reardon & Raudenbush, 2009). The concerns include the difficulty value-added models have in accounting for the joint production of achievement by teachers within the same school; the assumption that all prior inputs decay geometrically at the same rate; and more generally what they hold to be violations of the assumptions of strict exogeneity.

A third line of inquiry surrounds the assumption that economists estimating value-added models are ignoring properties of the test (Koretz, 2002; McCaffrey, Lockwood, Koretz, & Hamilton, 2003). They argue that academic researchers are treating the test as a prima facie indicator of student achievement when in fact the properties of the can render results between students incomparable. For example, tests may not be scaled on an interval basis so that movement along one part of the distribution could be easier than another. This line of argument can be summarized by concerns of the psychometricians that their tests are being used for a purpose—evaluating teachers—for which they were not intended.

Beyond these arguments about the specification of the value-added models, researchers are engaged in a heated debate about the proper use of the value-added estimates. This debate is largely being fought through competing reports written by academics for policy think tanks. These reports are passed along in the policy circles as evidence that researchers cannot agree on the proper use of the estimates. Perhaps the most prominent one was published by the Economic Policy Institute (Baker, et al., 2010).

It critiques the emphasis on value-added measures for their imprecision, inability to account for summer loss, sorting of students to teachers, incentives to narrow curriculum to tested subjects, and inability to control for factors outside the teacher's control. A panel from the National Academy of Science reached similar conclusions (Chudowsky, Koenig, Braun, National Research Council (U.S.). Center for Education., & National Academy of Education., 2010).

The Brookings Institution issues a rejoinder paper (Glazerman, Loeb, Goldhaber, Raudenbush, & Whitehurst, 2010). The counter argument in this and other similar policy briefs is to accentuate the counterfactual in which current teacher evaluation systems do not generally distinguish among teachers. Although it is not a perfect measure, it does add information for the district policy maker that is not available from the customary proxies of teacher effectiveness such as national board certification, advanced degrees, or years of experience.

Although some of the literature around uses of value-added estimates has focused on the placement of teachers (Clotfelter, Ladd, & Vigdor, 2005; Clotfelter, et al., 2006; Hanushek & Rivkin, 2010a), more has focused on the use of value-added estimates to inform high-stakes personnel decisions. In the context of national cutbacks in the teacher workforce, several studies simulate how districts might employ value-added results in hiring and firing decisions (Goldhaber & Hansen, 2010; Staiger & Rockoff, 2010; Yeh & Ritter, 2009). Another line of research involves the use of value-added estimates to inform performance-based compensation systems (Roland G. Fryer, 2011; Lazear, 2003; Neal, 2011; Podgursky & Springer, 2007; Springer et al., 2010)

There are far fewer studies that illuminate the policy decisions districts face in adopting one or more value-added models. A recent overview for policymakers (Harris, 2011) should help districts assess the strengths and weakness of value-added measures. The greatest contribution to districts has come from the work of McCaffrey et al. of the RAND Corporation (McCaffrey, Han, & Lockwood, 2010; McCaffrey, et al., 2003; Steele, Hamilton, & Stecher, 2010).

McCaffrey, Han and Lockwood provide perhaps the most useful guidance. They walk districts through a number of decisions ranging from matching teachers and students to different estimators. They list a number of considerations for the business rules of matching teachers and students that a district should establish (e.g., the number of days a student must be enrolled with the teacher in order to be attributed to the teacher). Then, the authors test several models—regression residuals, ANCOVA, Look-Up Tables, Gain Scores, Multivariate ANCOVA, Mixed Models, and student fixed effects—for their sensitivity to bias due to sorting and their relative degree of precision. They find that the mixed models and student fixed effects generally produce more precise and less biased estimates of teacher effects. Finally, the authors explore the implication of the uncertainty in the estimates for decision rules about who would qualify for a bonus.

This study extends McCaffrey, Han and Lockwood to help districts align their decisions on value-added models to their policy goals. From issues of covariate selection to the form of the estimated teacher effects, it evaluates many of the statistical concerns from the academy by criteria—technical, practical, political—which are of significant concern to the district. It addresses directly to specific doubts often raised in districts concerning the validity of value-added models—ceiling effects and the sorting of teachers to students. Finally, it goes beyond the current literature in exploring how the value-added effects could be used as measures of equity within a school district.

DATA

This study employs data from an administrative data set supplied by Charlotte-Mecklenburg Schools (CMS) that includes cohorts of students in grades 5-8 during the period 2008-2010. It contains information on the students, their courses and teachers, and their schools.

STUDENTS

Data on the students include their demographic characteristics, test score history, enrollment patterns (e.g., mobility between schools), attendance and behavioral record. The demographic data on students include their gender, English proficiency, designation as academically gifted by the district, ethnicity, special education designation, and age relative to their peers. Unlike many data sets used in similar studies, the data do not include an indicator of the student's eligibility for free or reduced lunch, the usual proxy for a student's socio-economic background. Due to legal constraints, the district no longer allows researchers access to this information.

Student test scores are from the standardized end-of-grade tests administered by the NC Department of Public Instruction (NCDPI) 3-4 weeks prior to the conclusion of the school year. These are norm-referenced tests that measure student mastery of the NC Standard Course of Study for that grade and subject. NCDPI converts the students' raw scores to scale scores that are designed to be vertically linked across grades and to possess the property of interval scaling such that a one point increase in the scale score reflects the same amount of learning across the scale score distribution. Following general practice, I standardize the scale scores by grade and year so that the scores have mean of zero and unit variation. Students who fail the tests are allowed, and in some cases, required, to take a re-test often within the week of the first test. I use the score from the first administration.[1] There are also a number of accommodations given to students with special needs. These range from having extended time on the test to taking alternate forms of the test. In this study, I use only those students who take the regular administration of the test.

The data also contain information on the student's enrollment in school. They indicate whether a student is repeating a grade, is enrolled in a school for the first time and/or has changed schools within the academic year. Data on the student attendance include the number of absences as well as the number of days enrolled in school within the year. The data also include the percentage of time the student spent in out-of-school or in-school suspension during the year.

Courses

As with many urban districts, Charlotte-Mecklenburg Schools establishes the link between students and teachers through course registration data entered by the school through the scheduling interface of the student information system. There is a record for each student in each course period. The record contains the name of the course, an associated course code provided by the state, a course code specific to CMS, the course day and period, and an associated instructor. In NC, end-of-course standardized tests are required for any student who is enrolled in courses with a given state course code. For example, any student enrolled after the 20th day of the term in a course with a 2001 state course code must take the end-of-grade math test. Researchers seeking to establish the

[1] Using a subsequent attempt introduces the effect of taking the test a second time into the student's score. Another alternative—averaging the test and the retest—changes the underlying variance of the score.

primary teacher for the student in the tested subject often default to the teacher listed by this course code. This assumption may not always hold for a variety of reasons.

First, student mobility among classrooms threatens the validity of teacher-student matches. When these students move, their course records are overwritten by the student information system. So, for example, a researcher who pulls the course table at the end of the school year will not observe the students' course enrollments in their previous schools. Aside from being unable to apportion the student's instructional dosage among more than one teacher, the researcher's calculations of the composition of the student's peers can be invalid. The mobile student may have been enrolled in the first term with Teacher A in Class B at School C, but when computing the classroom averages for that class, the student will not contribute to those peer means.

A similar overwriting happens at CMS when a student withdraws from school. The course records of students who withdraw from the district are purged from the course data, making it impossible to include them in *ex post* calculations of peer means. To the extent that these students are clustered in certain schools or among certain teachers, their omission from the data could bias estimates of the teacher effects, especially if their withdrawal was endogenous to the teacher's effect.

A second risk of invalid teacher-student matches stems from within-year teacher mobility. A teacher may leave the school and a new teacher takes over. If the school administrator does not update the database, the teacher of record will be the first teacher. Or suppose a teacher takes maternity leave. The student may have an interim teacher or the student's class may be instructed by a principal or other group of teachers who fill in for the teacher. This type of arrangement will likely never be recorded in the administrative dataset and will be unobservable to the district office.

A third threat to the validity of the teacher-student matches arises from non-traditional teacher arrangements. The scheduling software used by CMS presupposes static groupings of students with one or two teachers. Yet increasingly in its elementary schools, CMS principals and teachers are using more dynamic groupings of teachers and students. For example, the principals may have all 5th grade students instructed in math by one 5th grade teacher who is particularly adept at math. They may or may not record this change in the student information system, even if the system allows this type of input. Or principals might adopt a type of flexible grouping approach where students are grouped into small subsets of the class based on instructional needs and then regrouped every few weeks. Each iteration of the groups could have a different instructor.

A final difficulty in matching students and teachers arises from students having varying exposures to different teachers in the same subject. Some low-performing students in a class are pulled out by a resource teacher or specialist for intensive instruction in reading.   There is a similar problem with students having varying amounts of time with the same teacher. For example, a student could be enrolled in a math course that has a lab component taught either by the same teacher or by a different one. If it is taught by the same teacher, and this is not a standardized practice across the district, then the teacher has greater exposure to the student than the teacher’s peers. To the extent that the teacher’s effectiveness is linked to the time with the student, this could conflate time and effectiveness and make comparisons among teachers difficult. Note that this could also be a problem if retained students have the same teacher from the prior year. In some cases, researchers using large state administrative data sets cannot observe these complementary teachers. When they do, they often drop students with multiple instructors from the sample.

Table 3 summarizes the exposure of students to multiple dosages of teachers and subjects within a term in 2010. Of the 42,552 students in grades 5-8 in the sample for that year, 51% had only one math course in the year and 48% had two math-related courses. Most of the students with two math-related courses had the same teacher for both. Eleven percent of the sample had two math-related courses with two different teachers.

Notwithstanding the potential threats to the validity of the teacher-student matches, I calculated the classroom means by aggregating the individual student data to the classroom level while excluding each individual student from his/her classroom mean.

SCHOOLS

Similarly, I aggregated all individual data to the school level to create school-level means. In addition, I included the school's percentage of students who are eligible for free or reduced lunch. This measure could influence student achievement in two possibly opposite directions. The school's free and reduced lunch eligible population is an indicator of the needs of its students—needs that are correlated negatively with the student's academic achievement. At the same time, this percentage is a trigger for a school's receipt of Title I funds from the federal government. This funding is tied to smaller class sizes or access to technology such as smart boards that could conceivably be correlated with improved instruction.

Arriving at the Sample

This section outlines the steps I took to arrive at the estimation samples.   First, I selected all CMS students in grades 5-8 during the period 2008-2010 from the master student file (N=160,527). I began with grade 5 because I want to have two prior test scores available for each student.

Second, I merged these students with the test score file. I summarized the demographic characteristics of the 11.4% of students who are missing a test score on the regular administration of the test in math in the current year and report this in column 1 of Table 4 (N=18,285).   These students are more likely to be special education students who would not take the regular administration or students who scored much lower than their peers in prior years.

Third, I merged these students with the course file, again summarizing the characteristics of the students who are not found in the course file in column 2 of Table 4. There are 21,893 students who do not have course records (13,785 of these also had no current year test score). Of the 21,893 with no course data, CMS records 438 students dropping out or being expelled and 12,460 students transferring out of the district during the year. These students' course data is wiped from the student information system at the end of the school year as a result of a business rule in the student information system updating procedure. There are another 9,407 students for whom I do not have any withdrawal data; of these, 8,032 were in eighth grade. These eighth-graders who disappear from the data are evenly distributed across the years 2008-2010.

Fourth, I divided the remaining students (those with a current year test score and course enrollment data, N=120,349) into three groups and summarize their characteristics in columns 3-5 of Table 4. Column 3 provides means and standard deviations for the 28,513 students who were missing a prior year test score. Columns 4-5 summarize the characteristics of students who have a score in the prior year and do not have one at t-2 (column 4) or do (column 5).

Columns 3-5 show that students who are missing one or two prior test scores differ significantly on observable characteristics from students who do have these scores. Students missing a score in the prior year score (column 3) score on average almost one-third of a standard deviation lower on the current year test than those who are not missing it (column 6). They are more likely to be special education students, older than their peers due to previous year retentions, and in their first year in the school. Their classmates scored on average one-fourth of a standard deviation lower than others in that grade and were far more likely to be special education students.

Among those students with scores from the prior years, those without a score from two years prior (column 4) differ significantly from those with a score from two years prior (column 5). They score significantly lower on the current year test as well as the prior year. They are more likely to be designated as having limited English proficiency (LEP), older than their peers, in their first year at the school and Hispanic. They are much less likely to be labeled as academically gifted. Many of these differences extend to their classes as well. The difference in their classroom mean math scores from the prior year is almost 0.15 standard deviations.

The differences in test scores among the students summarized in columns 3-5 extends beyond central tendencies. Figure 1 shows the kernel densities of current year test scores for the students summarized in columns 3-5. The distribution of current year scores for students missing scores from the last two years is well to the left of students who are not missing those scores. A similar pattern holds for students who have a score from the prior year but not two years prior.

The magnitude of the differences among these students has policy implications for the district's decisions on the how to handle missing data on prior test scores. All value-added models rely upon having at least two years of achievement data to estimate a student's growth, from which the teacher's effect is derived. An identification strategy for teacher effectiveness that excludes these students from the analysis leads to several potential problems. First, assuming that the district would impose a floor on the number of students a teacher must instruct in order to compute an effectiveness measure, excluding these students could lead to some teachers being left out of the analysis altogether. Second, some teachers will have a mix of students who are returning and who are new to the district. If a district excludes the new students from the teacher's calculation, it could be estimating the teachers' effect on only a relatively small proportion of their students.

The CMS data allows me to gauge the magnitude of these threats. Requiring each student to have two prior scores eliminates 156 teachers and 1,060 classrooms from the value-added calculation over three years. For teachers who do not drop out the sample, the requirement of students having two prior year scores results in a median loss of 4 students per year from a median teaching load of 56 students per year (three of the four grades in the sample are in middle school). I will compare estimates of teacher effects that include and exclude students without two prior scores as a way estimate the difference it will make for teachers who are not excluded from the sample.

One alternative to excluding students with missing data from the sample is an imputation procedure, and this has both statistical and practical concerns. Districts are likely equipped to handle simple single imputation. However, the multiple imputation technique recommended as a best practice in the statistics literature is more difficult to compute, especially in combination with more sophisticated estimators (e.g., multilevel models) (Rubin, 1996). Aside from the computational considerations, the rationale for imputing the prior values is hard for district analysts to explain to teachers, especially when high stakes are attached to the estimate of a particular teacher's estimate. In this study, I leave the issue of how to deal with missing data through imputation for further research.

I use as the estimation sample students who have a score in the prior year and who may or may not have two prior scores and who have non-missing values on other variables. These students are summarized in column 7 of Table 4.

METHODS

This section details the methods used to explore the three central lines of inquiry of this study. These lines of inquiry are:

1. What are the benefits and costs of various value-added models in terms of the identification and specification of teacher effects?
2. How serious are two often-cited threats to the validity of value-added estimates—ceiling effects in the test instrument and the sorting of teachers to students? What can be done to mitigate the risks they pose?
3. Are value-added estimates suitable for use in (a) considerations of a district's equitable allocation of its resources across students, and (b) personnel decisions?

INTRODUCTION

It can be useful to begin with the simple question that value-added estimates of teacher effectiveness are seeking to answer: what would happen to the test scores of the students in a given classroom if they had one teacher rather than another (Kane & Staiger, 2008)? To answer this question, a district would need to have multiple teachers teaching the same group of students in the same school at the same time of day. The district could then conclude that under *ceteris paribus* conditions, the difference in the student test scores at the end of the course would be the teacher's effect relative to the other teachers. This, of course, is impossible, but framing it this way points to what from a potential outcomes framework, Holland calls the fundamental problem of causal inference—someone cannot simultaneously receive the treatment and not receive the treatment at the same time (Holland, 1986).

One option would be to create an experiment.[2] The district could randomly assign teachers to students within the same school and grade. If the experimental conditions held, it could attribute the differences in the student test scores to the teacher. In this case, from a potential outcomes framework, the district would assume:

$$E\left[A_{ijt} \middle| \theta_t\right] = 0 \tag{1}$$

where A_{ijt} is the achievement of student *i* with teacher *j* at time *t* and θ is student's teacher at time *t*. From the perspective of the potential outcomes framework, the district would be looking at a specific teacher as a treatment effect (see Imbens & Wooldridge, 2009; Rubin, Stuart, & Zanutto, 2004). In Equation (1), it would assume that all differences among the students' test scores arose from the treatment.

The plausibility of this assumption is easily challenged. A long line of literature casts student achievement in terms of a production function in which the teacher's input is just one of many factors (Hanushek, 1986 ; Harris & Sass, 2006; Lazear, 2001; Todd & Wolpin, 2003). These factors include aspects of the child's neighborhood, home life, socio-economic status—factors largely beyond the control of the school district—as well as those factors that the school can control—the assigned school, peers and teacher.

Todd and Wolpin (2003) provide one of the seminal explanations of the educational production function and the assumptions required to identify the contribution

[2] Experimental estimates of teacher effects based upon random assignments of teachers to classrooms are both relatively rare and usually limited in scope. Researchers from Mathematica used random assignment to assess impact of Teach for America teachers, see Glazerman, S., Mayer, D., & Decker, P. (2006). Alternative routes to teaching: The impacts of teach for america on student achievement and other outcomes. *Journal of Policy Analysis and Management, 25*(1), 75-96. For a critique of the reluctance of school districts to use random assignment, see Cook, T. D. (2003). Why have educational evaluators chosen not to do randomized experiments? *Annals of the American Academy of Political and Social Science, 589*, 114-149..

of any one input. Adapting their argument, I specify a basic education production function as follows:

$$A_{ijt} = A_t[\mathbf{F}_i(t), \mathbf{Y}_{it}(t), \mathbf{Z}_i(t), \mathbf{C}_{-ijt}(t), \mathbf{S}_{-ijt}(t), \theta_{ijt}(t), c_i(t), e_i] \quad (2)$$

where *A* is the academic achievement of student *i* with teacher *j* at time *t*. **F** is a vector of student *i*'s family inputs to the student's achievement, **Y** is a vector of student inputs that are time-varying, **Z** is a vector of time-invariant student inputs, **C** is a vector of classroom level inputs; **S** is a vector of school inputs; θ is the teacher; *c* is the unobserved "endowment" or "heterogeneity" of the student, and *e* is a random error term. In this formulation, the *t* subscripts denote the presence of the input at time *t*, making explicit that the effect of the input could vary by time. For example, the student's unobserved endowment c_i is time invariant and yet its impact on student achievement could vary with time. In contrast, the student's classroom $\mathbf{C}_{\text{-ijt}}$ is both time-varying and its effect could depend on time, e.g., a student may be more influenced by the peer composition of the class in third grade than in the eighth.

For researchers using even the rich administrative data sets that have come to dominate the last ten years of research on value-added models, some of these inputs are often unobservable. How does a district know how to estimate the family's input into education save for a few proxies such as the decision to enroll in a magnet school or the student's eligibility for free or reduced lunch? To the extent that these unobserved inputs are correlated to the observed inputs and to the student's achievement, then estimates of the observed inputs, including the teacher effects, are likely to be inconsistent.

Turning to a regression framework for estimating Equation (2), I simplify the notation by temporarily collapsing inputs so that $\mathbf{F}, \mathbf{Y}, \mathbf{Z}, \mathrm{C}, \mathrm{S} \in \mathbf{X}$ to get:

$$\mathrm{A}_{ijt} = \mathbf{X}_{ijt}\alpha_1 + \mathbf{X}_{ij,t-1}\alpha_2 + \ldots + \mathbf{X}_{ij1}\alpha_\alpha + \theta_t\lambda_1 + \ldots\theta_1\lambda_\lambda + \beta \mathrm{c}_i + \varepsilon_{ijt}, \text{ for } t = 1,\ldots,T \tag{3}$$

where α_t is the estimated effect of the inputs **X** at time *t*, λ is the effect of the teacher at time *t*, β is the effect of the student's endowment and ε is a random error term. This formulation makes explicit the potential effects of prior inputs on the student's current achievement. A student's achievement at a given time, *t*, is an additive function of current and past realizations of the family, student, and school inputs, as well as a fixed contribution from the student endowment c_i and an error term that at this point I cannot assume is i.i.d.

For a number of reasons, estimation of Equation (3) is unfeasible. School districts do not observe the contemporaneous or lagged inputs from the child's family. In fact, school districts do not even approach observing a complete set of the prior schooling inputs. The risk of not accounting for these lagged inputs is that the current year inputs could be endogenous, rendering the estimates of the impact of contemporaneous inputs (e.g., the teacher effect) inconsistent and biased.

At this point, it may be useful to frame just what is at stake in the violations of the exogeneity assumption. The estimation of Equation (3) is a student-level estimation of the impact of the teacher effect on the student's achievement. Yet the parameter of interest for a district is not the particular impact of the teacher on that student's learning, but rather the impact of that teacher across all the teacher's students. By aggregating Equation (3) across the teacher's students it becomes possible to distinguish more readily between noise and bias. The noise at the level of Equation (3) may not result in bias when these are aggregated to the teacher. In this framework, I rework Equation (1) to the teacher level so that:

$$E\left[\sum_{j=1}^{J}\sum_{i=1}^{N}E\left[A_{ijt}\mid\theta_{ijt}\right]\right]=0,\ \text{for all i=1...N and j=1...J} \tag{4}$$

where *i* is student of teacher *j* at time *t*.

PREVIEW OF ANALYSES

The analyses in the study will proceed in three steps: (1) arriving at a preferred model for the estimation of teacher effects, (2) testing that based model against threats to validity common to most value-added models, and (3) exploring the policy implications of the resulting teacher effects. Table 1 summarizes the analyses.

TABLE 1: Sequence of Analyses.

Sequence of Analysis

- Arriving at a preferred model
    - Dealing with lagged score
    - Accounting for student heterogeneity
    - Accounting for school heterogeneity
    - Accounting for classroom heterogeneity
    - Modeling teacher effects
- Testing the preferred model
    - Prior Inputs
    - Ceiling Effects
    - Sorting of Teachers to Students
- Policy Implications
    - Distributing teachers to students
    - Personnel Decisions

ARRIVING AT A PREFERRED MODEL

In this section I build a preferred model for estimating teacher effects that I use to assess threats to the validity of the inferences about teacher effects and policy implications. In creating the preferred model, I cover four primary issues: (1) ways to handle the inclusion of the student’s prior test score, (2) options for controlling for student heterogeneity, (3) options for school heterogeneity, and (4) estimating teacher effects. In order to limit the number of analyses, I do not estimate every permutation of the available options. Instead, I consider only those options that seem likely to be consequential for the resulting effects.

Dealing with the Lagged Score

I begin building the preferred model by evaluating three common approaches to using the student's prior test score: as a lagged dependent variable, a gainscore and instrumenting for the lagged score with the twice lagged score. In evaluating these alternatives, I need to hold constant some of the other options (e.g., how to account for school heterogeneity) that will be discussed later in this section. In this first step, I estimate the teacher effects as fixed and conditioned on student characteristics. In the absence of a statistical test with which to compare the results of the models, I evaluate the approaches by the constraints on the sample imposed by the approach (e.g., needing three years of student data vs. two) and the extent to which their resulting estimates are correlated and. I turn now to a discussion of rationale for each option.

Models 1-2: Lagged Prior Score(s)

To capture the effect of past inputs on student achievement, many researchers include one or more lagged test scores in estimating some form of:

$$\mathrm{A}_{ijt} = \lambda_1 \mathrm{A}_{ij,t-1} + [\lambda_2 \mathrm{A}_{ij,t-2}] + \mathbf{X}_{ijt}\alpha + \theta_{ijt} + c_i + \varepsilon_{ijt}, \text{ for all } t = 2,...,T \qquad (5)$$

The premise is that the lagged score accounts for the accumulative effect of all prior inputs including the student's endowment, which is unobserved.

Although including the prior year score may help in mitigating the bias caused by being unable to account for prior inputs to the student's achievement, including it introduces three new threats to the validity of the estimated teacher effects. First, suppose it were the case that the unobserved endowment's effect on the score was not static over time. If the lagged score captures the unobserved endowment, then I am assuming that the correlation between the score and endowment is the same in all time periods. This assumption can be written as

$$\begin{aligned} &\mathrm{A}_{ijt} = \mathrm{A}_{ij,t-1} + \mathbf{X}_{ijt}\alpha_1 + \delta_1\theta_t + \beta_i \mathrm{c}_i + \varepsilon_{ijt} \\ &H_0 : \beta_i = \beta_j,\ \text{for i} \neq \text{j} \end{aligned} \tag{6}$$

It is conceivable that this would not be the case. Could smarter children grow faster than their peers, even conditional on the prior score? If so, then the coefficient on the lagged score will be biased upward (Andrabi, et al., 2009).

A second assumption of the models including lagged prior scores is that the effect of the student endowment is constant across time. This assumption can be expressed as:

$$\begin{aligned} &\mathrm{A}_{ijt} = \mathrm{A}_{ij,t-1} + \mathbf{X}_{ijt}\alpha_1 + \delta_1\theta_{ijt} + \beta_1 \mathrm{c}_i + \varepsilon_{ijt} \\ &\mathrm{A}_{ij,t-1} = \mathrm{A}_{ij,t-2} + \mathbf{X}_{ijt-1}\alpha_2 + \delta_2\theta_{ij,t-1} + \beta_2 \mathrm{c}_i + \varepsilon_{ij,t-1} \\ &\vdots \\ &\mathrm{A}_{ijT} = \mathrm{A}_{ijT} + \mathbf{X}_{ijT}\alpha_T + \delta_2\theta_{ijT} + \beta_T \mathrm{c}_i + \varepsilon_{ijT},\ \text{for all } t = 2,\ldots,T \\ &H_0 : \beta_1 = \beta_2 = \beta_T \end{aligned} \tag{7}$$

If this assumption does not hold, then the coefficient on the lagged score would need to vary by time or grade if it is to capture the impact of the student endowment on student achievement. I assume Equation (7) holds for the purposes of this study.

A third assumption with the inclusion of the lagged score is that the coefficient on the lagged score is constant across all students in the sample. One can allow the functional form of the lagged score to vary by including quadratic and cubic polynomials so that the coefficient could vary depending on where the student's prior score lies in the distribution of scores (see Figure 2 for the relationship between test scores at *t-2* and *t*, and *t-1* and *t).* Yet this still homogenizes the trajectories of the students. One way to mitigate this constraint is to include the second lagged score from *t-2*. In addition to allowing for more of a trajectory, having two prior scores should help to minimize the effects of measurement error in the prior scores. I estimate Equation (5) with the *t-2* score included as well. I report the results in column 2 of Table 5.

Model 3: IV Estimates

Models that include a lagged test score on the left-hand-side are subject to bias in the estimated teacher effect due to measurement error in the lagged score. The lagged test score is an additive function of the true score and measurement error. This measurement error could result from many things: the student could have had a bad testing day or the test just happened to have questions that the student was well-equipped to answer. The measurement error invites two potential problems. First, it increases the noise in the estimate of the coefficient on the lagged score and attenuates that coefficient. Second, it also renders OLS estimates of teacher effect estimates using a lagged score inconsistent by inducing correlation between the error term and the lagged score (Andrabi, et al., 2009; Harris & Sass, 2006).

Perhaps the most straightforward solution to the first problem—the attenuation of the coefficient on the prior score—would be to correct for the measurement issue by using a known estimate of the reliability of the assessment from the testing service. However, many districts may not have access to the reliability of the assessment In this study, I do not pursue solutions to using estimates of the reliability of the of the prior score in the estimation of the prior scores (Boyd, Lankford, Loeb, Wyckoff, & Grossman, 2008). I do pursue an approach often used to correct for the inconsistency wrought by having the lagged score correlated with the error term. This approach is to instrument for the score at *t-1* with that of *t-2*.

$$A_{ijt} = \lambda A_{ij,t-1} + \mathbf{X}_{ijt}\alpha \ + \theta_{ijt} + c_i + \varepsilon_{ijt}, \text{ for all } t = 3,\ldots,T \tag{8}$$

I estimate Equation (8) using a 2SLS estimator and report the results in column 3 of Table 5.

Model 4: Gainscore

As a final option, I estimate a gainscore specification. In this specification the dependent variable becomes the change rather than the level of the score. The intuition is that by moving the lagged score from the right-hand-side to the left, you solve the measurement error issue.

$$A_{ijt} - A_{ij,t-1} = \mathbf{X}_{ijt}\alpha \ + \theta_{ijt} + c_i + \varepsilon_{ijt}, \text{ for all } t = 2,\ldots,T \tag{9}$$

In order to understand the drawbacks of such an approach, it helps to consider the specific interpretation of the coefficient on the prior year test score. It can be construed as a measure of the decay of academic achievement from one year to the next.

$$A_{ijt} = \lambda A_{ij,t-1} + \mathbf{X}_{ijt} + \theta_{ijt} + c_i + \varepsilon_{ijt} \tag{10}$$

Jacob, Lefgren and Sims (2008) show that this coefficient could be the decay of long term learning but also measurement error and short-term cramming for test.

Under most plausible assumptions about the nature of learning, a district could assume that $0 < \lambda < 1$. If λ=0, then it is assuming complete decay such that no inputs from the prior year would have an impact on achievement in the current year. If λ=1, this implies that there is no decay in learning from one year to the next. The gainscore model effectively constrains λ to one. The implication of this constraint is that the effect of an input is independent of the time that it is applied, leading Andrabi et al. to conclude:

"...this implies that the effect of each input must be independent of when it is applied...For example, the quality of a child's kindergarten must have the same impact on their achievement at the end of age 5 as it does on their achievement at age 18” (2009, p. 8) .

Their critique is supported by empirical evidence that the effects of teacher effects do not in fact persist without decay (Andrabi, et al., 2009; Jacob, et al., 2008; Kane & Staiger, 2008).

While not disputing the implications of the complete decay assumption, Harris and Sass (2006) test the degree to which the constraint changes estimated teacher effects. They estimate specifications in which the lagged achievement variable's coefficient is constrained to various levels of decay. They find the teacher effects are highly correlated (r=0.88) regardless of the constraint. They conclude that the benefits of dealing with measurement error outweigh the cost of the complete decay assumption. I report the results of this estimation in column 4 of Table 5.

Accounting for Student Heterogeneity

The prior test score captures some but not all of the student heterogeneity. I begin by addressing the heterogeneity of a teacher’s students along four dimensions represented in Equation (2). These include their unobserved endowment or innate ability, *c*,; their unobserved family inputs such as the parental support of their education, **F**; time-invariant characteristics, **Z**, such as a student’s gender; and time-varying student characteristics, **Y**, such as their absence rate. The primary question for the district is the extent to which these variables explain enough of the student’s unobserved characteristics to mitigate substantially the risk that these variables could bias the estimates of the teacher effects.

Many of the common value-added models adjust for student characteristics that are both time-invariant and time-varying. These characteristics can include demographics (e.g., ethnicity, socio-economic status and gender), behavior (e.g., discipline and attendance), and enrollment patterns (e.g., between-school mobility). In the estimation of the models above that deal with the lagged score, I include sets of time-varying and time-invariant student covariates. I use F-tests of the joint significance of each set of characteristics a measure of their contribution to the model. I report the results in columns 1-4 of Table 5.

I discuss two sources of unobservable student heterogeneity. One of the most significant sources of variation among students in their achievement is the family level inputs that are largely unobservable to the district. We know empirically that these inputs matter and that they include factors ranging from the number of books in a household (Roland G. Fryer & Levitt, 2004; Roland G. Fryer & Levitt, 2006) to the child’s grandmother’s education (Phillips, Brooks-Gunn, Duncan, Klebanov, & Crance, 1998).

A second source of unobserved student heterogeneity is the student’s own innate ability. This endowment could affect both the level scores in a given year and the rate of change of scores across years. I assume that both sources of unobserved heterogeneity would influence a student’s achievement in a given year. The threat of these sources of unobserved student heterogeneity to estimates of teacher effects depends upon the extent to which they are also correlated with contemporaneous inputs to the student’s achievements. They might have two types of relationships to the current inputs that could classified as static and dynamic. Imagine that the student has parents who are very engaged in the student's education and provide outside-of-school opportunities for the child that would influence the child's achievement, perhaps even before entering school. I assume that this sort of parent input could be consistent over time and that its effect on the student's achievement is also similarly constant.

Parental input could also have a more dynamic nature. Suppose the student of highly engaged parents has an off year and receives an unusually low score in math. In this case, the parents may respond by securing a tutor for the child. If we imagine the mean parental involvement in the student's education over the years, then in this year, there would be a positive deviation from this mean that could easily be confused with the teacher's input in that year. How can the district account for the difference in the input of the tutor and the student's math teacher in that year?

There is little the district can do to account for time-varying unobservables such as the dynamic response by parents to inputs. Instead, I explore the potential of student fixed effects and first differences to capture student unobservables beyond those captured by the lagged scores and observable student characteristics in the previous models. The former identifies the impact of the inputs on a student's achievement by predicting deviations from the student's average academic achievement with the deviations of the inputs from their average for the student:

$$\overline{\mathrm{A}}_i = (\mathbf{X}_{ijt} - \overline{\mathbf{X}}_i)\alpha_2 + \left(\theta_{it} - \overline{\theta}_i\right)\lambda + \varepsilon_{ijt}, \text{ for all } t = 3, \dots, T \tag{11}$$

The latter identifies the effect of the inputs from changes in the inputs from one year to the next.

$$\Delta \mathrm{A}_{ijt,t-1} = \Delta \mathbf{X}_{ijt,t-1}\alpha_1 + \Delta\theta_t\lambda + \varepsilon_{ijt}, \text{ for all } t = 3, \dots, T \tag{12}$$

In this study, I estimate both student fixed effects and first differences and report the results in columns 5-6 of Table 5.

Approaches to estimating teacher effects that depend on panel data from students are likely to invite serial correlation into the error term. The intuition is that the unobserved endowment will be correlated with both the lagged score and the current score. If this were true, it would violate the assumption of strict exogeneity on which the panel approach to accounting for unobserved student heterogeneity depends. In this study, I test for the presence of serial correlation using Stata's –xtserial— program which is based upon a test developed by Jeffrey Wooldridge (Drukker, 2003). This test is predicated on the assumption that if the errors are uncorrelated, they should be correlated at -0.5 in a first difference estimation. If I find evidence of serial correlation, then it would suggest a first-difference approach rather than a student fixed effects approach to modeling unobserved student heterogeneity. I report the results of the serial correlation tests in the text.

After running the models reported in columns 1-6 of Table 5, I select a preferred model that I use in the subsequent analyses. I choose the model based upon the following criteria:

1. Tests of the joint significance of added variables and their contribution to the model fit.
2. When correlations of the estimated teacher effects from two or more models are high, I prefer the more parsimonious model.
3. Maximizing the number of students and teachers that can be included in the model.

Based upon these criteria, I choose a preferred model of student heterogeneity that I use as the preferred model upon to add ways to account for classroom heterogeneity.

Accounting for Classroom Heterogeneity

As with students, classrooms of students will also differ in their influence on student learning. The mechanism of the influence is primarily through peer effects (Hanushek & Rivkin, 2008; Hoxby, 2000; Hoxby & Weingarth, 2005). The importance of the classroom heterogeneity becomes clear in thinking about estimating, for example, a fifth-grade teacher's effect on her students' achievement in math in a given year. If she has one class, then a fixed effect for the class and a fixed effect for the teacher would pick up the same variation. They would be perfectly collinear.

It is easy enough to include classroom-level covariates on the right-hand-side of the teacher effect estimator. Most often these covariates are peer means. It is worth noting, however, that one issue complicating what would seem to be a straight-forward calculation is identifying the actual class. For example, using CMS data, I cannot reconstruct a class prior to 2007. Students could be with the same teacher at the same time but under different course names, e.g., a special education child might have a special education code and be in the same classroom. Section numbers were calibrated to the course code not to a physical location. The problem, of course, is that this is not modeling the student's production function precisely, but rather approximates it. In this study, I estimate the preferred model from Table 5 with classroom means, and I report the results in column 2 of Table 6.

Accounting for School Heterogeneity

There is consensus in the value-added literature that the majority of variation in teacher effectiveness is within rather than between schools (Aaronson, Barrow, & Sander, 2007; Goldhaber & Brewer, 1997). Even so, a nontrivial amount of the variation in teacher effects is across schools. It is difficult to isolate the teacher's effect from the school's on student achievement. Unless the district conditions its prediction of a student's achievement on the student's enrollment in this school, it is likely to conflate the effect of the program with the teacher's effect. One can easily imagine a number of observable and unobservable characteristics of the school—a new reading program or dynamic principal—that could similarly add noise to the estimate of the teacher's effect.

Researchers commonly respond to this issue of school heterogeneity by including either school level covariates (often peer means) or school fixed effects. Both approaches have the effect of constraining the comparison group for a given teacher to teachers in similar settings. Taking the more extreme case of school fixed effects, by removing the variation in student achievement associated with attending a specific school, the district constrains the estimate of the teacher's effectiveness to a comparison with other teachers in that school. Although this approach is likely to remove more noise from the estimate of the teacher's effectiveness, it also restricts the district's policy uses of the data. For example, it does not allow the district to measure the distribution of teacher effectiveness across schools. As a result, any attempt to compensate teachers for their performance on this metric would have to be within a school (e.g., the top 10% of teachers in each school will get a bonus). So there is a tradeoff for the district—does the benefit of reducing the noise in the teacher effect estimate outweigh the costs of restricting the use of the estimate? I attempt to explore the magnitude of the tradeoff by estimating models with and without school fixed effects (see columns 3-4 in Table 6), comparing, as before, the contributions of the school fixed effects to the model fit and the correlations of estimates under models with and without school fixed effects.

Modeling Teacher Effects

Once I develop a preferred model of student achievement that includes prior scores, student characteristics, classroom characteristics and school characteristics, I turn to estimating the teacher effects. In this section, I consider three issues regarding the teacher effect estimators: (1) fixed or random teacher effects, (2) adjusting the effects for sampling error, and (3) classroom-level shocks.

Form of Teacher Effects

Districts must decide whether to estimate the teacher effects as fixed or random. In much of the literature, the effects are estimated as fixed, but some estimate them as random (see Table 2). Only Harris and Sass (2006) discuss the methodological considerations of the choice. The assumption of the random effect approach is that the random teacher effects will be uncorrelated with the student, class and school characteristics that condition the expectation of a student's change in test scores. On the surface, this assumption seems untenable. There is strong evidence of sorting of teachers to students on observables (Bonesronning, Falch, & Strom, 2005; Boyd, Lankford, Loeb, & Wyckoff, 2005; Clotfelter, et al., 2005; Clotfelter, et al., 2006).

Given the implausibility of the assumption, why would a district consider estimating teacher effects as random? The advantage to the district is twofold: (1) the estimated effects are already shrunken by the sampling error and thus require no post-hoc transformation of the effects, and (2) the random effects allow the district a straightforward way to estimate effects for every teacher in the sample, i.e., the random procedure does not require a hold-out teacher (Mihaly, McCaffrey, Lockwood, & Sass, 2010). This simplifies the interpretation of the estimates for the district (i.e., a mean of zero is the average for the average teacher, not the hold-out teacher).

Adjusting for Sampling Error

As a set of estimates, the variation in teacher effects will include both estimation and sampling error. Standardized test scores can be noisy measures of achievement for individual students. The student may not be feeling well on a given test day and perform poorly. Or the student may have been lucky; the test covered items, such as a reading passage, that happened to sync with the students' own interests. The noise can result from classroom sources as well. The axiomatic "dog barking outside the classroom" or an air-conditioning unit that malfunctions on the day of the test are both examples of classroom-level shocks that could introduce noise to a student's test score.

The level of noise in individual students' test scores poses a threat to the validity of inferences about the effectiveness of teachers or schools in raising those scores. How much signal can a district wring from noisy student test scores? To the extent that this noise is random, one strategy is to assume that as student scores are aggregated across a teacher's class, the errors would wash out. One way to see this is to examine the relationship between the number of student contributing to a teacher effects and the variation of those teacher effects. In one of the canonical papers on the influence of sampling variation on accountability systems, Kane and Staiger (2002) show that schools in North Carolina with smaller numbers of students have effectiveness estimates at the extremes of the distribution. The magnitude of the estimates depends in part on the number of observations attributable to the teacher or school.

A common way to deal with both sources of error is through empirical Bayes or "shrinkage" estimators. The idea follows the logic prevalent in the multilevel model random effects literature (Rabe-Hesketh & Skrondal, 2008; Snijders & Bosker, 1999; West, Welch, & Galecki, 2007) in which you partition the residual variation from a student level regression of the current test score on prior year test scores and any covariates into those student, class and teacher level error. The shrinkage estimator multiplies the teacher effect by an estimate of its reliability as measured by the ratio of the signal of the teacher effect to the signal plus the noise. Thus estimates that have high amounts of noise are shrunken towards the population mean. In the literature, these empirical Bayes estimates are estimated primarily in two different ways.

The most common way is to estimate the teacher effects as fixed effects and then to apply a post-hoc shrinkage procedure (Harris & Sass, 2006; Kane & Staiger, 2008; Koedel & Betts, 2007). In this procedure, the author reports both unadjusted and adjusted variations of the teacher effects. The unadjusted are just the variation in the teacher effects. The adjusted variation of teacher effects is unadjusted variation of the teacher effects minus the sampling error. Typically, the literature uses the mean of the squared standard error of the teacher effects as the estimate of the sampling error.

Another way to generate the empirical Bayes estimates is to estimate them directly through a procedure such as Stata's –xtmixed— program. Here the estimates of the teacher effects are not estimated directly, but rather as predictions of the random effect that have been pre-shrunk. From a district's perspective, the key tradeoff is that estimates of the teacher effect can take substantially longer to complete since they are maximum likelihood estimates.

In this study, I estimate the proportion of the unadjusted variation in teacher effects that is attributable to sampling error and compare the resulting distributions of teacher effects from three specifications: (1) unadjusted from a teacher fixed effects approach, (2) the same, but adjusted using the procedure above for the adjustment, and (3) teacher random effects. I report the resulting distributions in columns 1-3 of Table 7. I graph the kernel density plots of the distribution of the effects in Figure 3. Finally, in Figure 4 I plot boxplots of the range of teacher effects by the number of students contributing to the estimated teacher effects as an attempt to gauge the sensitivity of the estimates to the number of students both for adjusted and unadjusted variations.

Classroom Shocks

A final consideration in estimating the teacher effects is the extent to which they may be confounded with unobservable classroom characteristics. Recall that even if one controls for classroom observables, you might still confound teacher and classroom effects if you have only one year of data for a teacher. A teacher might have had a particularly good match with the class in that year, or there might have been a classroom level shock. One way to handle this is to estimate multiple classes (in the case of elementary schools, this will be mean multiple years) and treat the classroom effect as a teacher-by-year effect nested within the teacher (Kane & Staiger, 2008; Rivkin, et al., 2005).

In this study, I test for the presence of a classroom level unobservable effect that is distinct from the teacher effect and report the results in column 4 of Table 7. I estimate the preferred model that emerges from the previous sections with and without the teacher-by-class effect. I use likelihood ratio tests to determine if the inclusion of this effect improves the fit of the model and the correlations of the teacher effects to test the practical significance of the difference.

TESTING THE PREFERRED MODEL

In this section of the study, I transition from comparing models on a number in terms of the specifications used to identify teacher effects to assessing threats to the validity that are common to all models. I examine: (1) the assumption that the prior test scores capture prior inputs to the student's production function, (2) the potential for ceilings and floors in the test score instrument to bias estimates of teacher effects and (3) the ways in which sorting of students across teachers within and between schools could bias teacher effect estimates.

Prior Inputs

A primary assumption of including the student score from t-1 is that it captures all prior inputs to the student's education production function. This assumption can be written as:

$$E\left[A_{ijt}\middle|\mathbf{X}_{ijt},\mathbf{X}_{ij,t-1},\ldots\mathbf{X}_{ijT},c_i\right]=E\left[A_{ijt}\middle|,\mathbf{X}_{ijt},A_{ij,t-1}\right],\text{ for all } t=2,\ldots,T \tag{13}$$

where **X** is the matrix of student inputs.

This assumption can be tested easily. I estimate the preferred model and test for whether the coefficients on the twice lagged inputs are jointly zero. If so, there is empirical support for the assumption that the lagged score has captured the impact of the prior inputs.

$$\begin{aligned} &A_{ijt} = A_{ij,t-1} + \mathbf{X}_{ijt}\alpha_1 + \mathbf{X}_{ij,t-2}\alpha_2 + \delta\theta_t + +\beta c_i + \varepsilon_{ijt}, \text{ for all } t = 2,\ldots,T \\ &H_0 : \alpha_2 = 0 \end{aligned} \tag{14}$$

Harris and Sass (2006, see Table 2) conduct a similar test and find no evidence of an impact for the prior inputs when the prior year score is included. As a sensitivity test, I add the student's score from t-2 and again test the joint significance of the prior inputs. I report the results in Table 9.

Ceiling Effects

One criticism of value-added models is that they will likely bias downward the effects of teachers who instruct students who are already at the high end of the distribution. The argument is that these students do not have "as far to grow" as those who are at lower ends of the distribution.

There are at least three threats to the validity of teacher effects for teachers whose students have scored at the high end of the distribution in the previous year. First, these teachers may focus on content that goes beyond the standard course of study and thus beyond the scale of scores. To the extent that these teachers' added value occurs beyond the range of knowledge assessed by the testing instrument, then these teacher effects will be biased downward. Koedel and Betts refer to this as a "lost information" problem (2009b, p. 7). This threat to the validity will not be addressed in this study. It is a question of the scope of the standard course of study.

The second threat to the validity concerns the assumption of interval scaling: a one-unit movement along the distribution of scores reflects the same magnitude of change in student achievement throughout the entire distribution. If this assumption does not hold, then equally effective teachers could have different value-added scores depending upon their students' scores in the previous year. For the purposes of this study, I assume interval scaling of the test score distribution.

A third and related threat to the validity is a function of the range of scores possible on the test. Students who scored high on the test in the previous year do not have as much "room to grow" as their peers and thus the potential value-added of their teachers is truncated or biased downward " (2009b, p. 7). The hypothesis is that if the teacher had a concentration of students whose previous year scores were at the top end of the distribution, then the teacher's effect would be biased downward. As Koedel and Betts point out, this hypothesis cannot be tested by examining the correlation between the previous year's score and the gains from that year because the presence of regression to the mean will induce a negative correlation. Prior scores are negatively correlated with gains. Nor can this hypothesis be tested by examining the correlation between prior scores and the teacher's current-year value-added estimate on the assumption that a negative correlation would be evidence of a ceiling effect bias. One could find little to no evidence of correlation but this would assume that teacher effectiveness is not sorted *a priori* by student ability.

There has been relatively little work done on the potential for ceiling effects to bias value-added estimates. Koedel and Betts (2009b) conduct simulations in which they right-censor the distribution of scores at various points and then test the effect of those ceilings on teachers' value-added estimates. They find that their estimates of teacher effects are robust to changes in the ceiling as they move it down to the 75th percentile (a skewness of -0.64). The correlations of teacher effects with a 75th percentile ceiling are correlated at 0.94 with the teacher effect estimates under no ceiling.

I cannot directly test the hypothesis that ceiling effects bias the estimates of teacher effects in the sample. However, following the empirical strategy of Koedel and Betts (2009b), I can explore the potential for bias given the distribution of student scores across teachers. First, I use kernel density plots of the actual and lagged scale scores for math in grades 5-8 to illustrate the extent to which the distribution of scores is negatively skewed in a way that might produce ceiling effects in the test. Second, following Koedel and Betts (2007), I divide students into deciles by their scores at *t-2* and plot their average gains from *t-1* to *t*. Smaller average gains in the upper tail of the distribution could indicate the presence of ceiling effects.

Third, I plot the value-added of teachers with by the percentage of their students in to the top and bottom 10% of the student distribution of scores at t-1. Smaller variation in the teacher effects for teachers with high proportions of previously high-achieving students could indicate that the test instrument is not picking up the full range of these teachers' contribution to their students' achievement.

Fourth, I analyze the proportion of students whose maximum gain on the test from t-1 to t will be smaller than the maximum teacher effect in that year. The intuition is that teachers with high concentrations of students at the top of the range could be at a disadvantage if the top teacher effect would be unattainable for them given their students. For example, suppose that a teacher at the 95th percentile of the effectiveness distribution improved their students' scores 0.4 standard deviations, or roughly 4 scale score points more than the average teacher. It is conceivable that an effect of this size could be impossible given the teacher's students' prior year scores.

Fifth, as a test of the effect of students at the high end of the distribution of scores from t-1 on the teacher effects, I re-estimate the base model with samples trimmed at the 99th, 95th and 90th percentiles of the distribution. I identify how many teachers are excluded from the calculation under each trimmed sample and the correlation of the effects from those that remain.

Finally, following a suggestion from Tim Sass (2010), I estimate the preferred model but with test scores in the current year normalized by the mean and standard deviation of the decile of the student's prior year score. The intuition behind this test is that the coefficients on the lagged score could vary by the position of the prior score's position in the distribution. By normalizing based upon the prior score, I would be modeling more flexibly the relationship between the prior score and the current score, perhaps even more so than including quadratic and cubic forms of the prior test (for an example of this technique, see LoGerfo, Nichols, & Reardon, 2006; Reardon, 2008). I compare the teacher effects estimated teacher effects from this model with those of the preferred model as a sensitivity test for the effect of the test score.

The potential for ceiling effects in test scores to bias estimates of teacher effectiveness depend largely on the extent of student sorting across teachers. In the next section, I outline how will I test the potential of such sorting to bias estimates of all teachers, not just those whose students scored at the top of the distribution in the prior year.

Sorting

The issue of sorting of teachers to students both across and within schools poses significant risks to inference about the effectiveness of teachers. It may be useful to observe these mechanisms through a measurement error framework—this time not in terms of the students' test scores, but rather in terms of the teacher's estimates. In this light, teacher effects are a function of their true effectiveness and an error component. We can never observe true effectiveness for it is always manifested in the context of confounding factors such as the characteristics of the students, classrooms and schools in which teachers demonstrate their effectiveness.

One approach to identifying the true effectiveness in the midst of sorting of teachers to students is to follow the strategy outlined above in the derivation of the preferred model where I control for sorting by conditioning teacher effects on observable student, classroom, and school characteristics. The presupposition of this approach is that true teacher effectiveness is distributed randomly across students conditional on the included characteristics. Hence, controlling for differences in students, classrooms and schools will reduce bias to the extent that these conditions influence the observed effects of teachers whose true effectiveness is not sorted across these student, classroom and school characteristics. This is the approach adopted in this study, and the sensitivity tests proposed in this section follow accordingly.

Yet before I proceed to the sensitivity tests, I want to note that it could also be the case that true teacher effectiveness is not distributed randomly across students. In this case, controlling for the student, classroom and school heterogeneity could conceivably bias the teacher effects. For example, suppose that truly highly effective teachers sorted themselves to schools with affluent children. In this case, true teacher effectiveness would be positively correlated with a school's socioeconomic status. By controlling for the school characteristics, I would bias downward the effectiveness of these teachers.

Because it cannot observe true teacher effectiveness, the district cannot assess the extent to which true teacher effectiveness is sorted across schools. However, the district can assess the benefits and costs of controlling for the student, classroom and school heterogeneity. Resuming the scenario above, suppose that the district's most effective teachers (true, not observed) sorted to affluent schools, and by controlling for the attributes of these schools in its value-added model, the district was in fact biasing downward these teachers' observed effect estimates to an undetermined extent. In the context of a pay for performance system, these teachers are already receiving non-pecuniary awards for being in these schools (e.g., increased parental involvement, less discipline problems, strong PTA support). There are already incentives now for these teachers to cluster to these schools. Even if the value-added model provided a slight disincentive to be at these schools, the district would need to decide if this is worse than the status quo in terms of increasing the probability that low-performing students in high-poverty schools have access to the most effective teachers?

Both across and within school sorting of teachers to students can threaten the validity of the value-added estimates. There is an extensive literature on the sorting of teachers across schools (Boyd, et al., 2005; Clotfelter & et al., 2004; Clotfelter, et al., 2005). Although there is less on the sorting of teachers to students within schools, the political science literature would suggest that within school assignment of teachers to students may be a principal's way of meting out rewards to favored teachers (Wilson, 1989). In fact, this type of sorting could be exacerbated by a compensation system which limits differential rewards based upon effectiveness among teachers. For example, in CMS, novice teachers are disproportionately assigned previously lower-performing student both across and within schools (Center for Educational Policy Research, 2010).

If this sorting occurred only on the basis of observable characteristics of the student, then the inclusion of these characteristics in the model to estimate teacher effects should mitigate the threat to validity of these estimates. However, if the sorting occurs on unobservable characteristics, then the threat is more pernicious. The threat of this type of sorting stems largely from the endogeneity of the student's inputs at time *t* to the student's unobserved endowment. For example, if the student's parents lobby for a particular teacher assignment and the human capital underlying the lobbying also predicts the student's score under that preferred teacher, then the estimate of that teacher's effect will be inconsistent.

One can imagine two types of within-school sorting of students to teachers. The first, and perhaps relatively easier one to account for, is that based on time-invariant student characteristics. If this were the case, then a model including student fixed effects should account for this sort of tracking. For example, this would remove the correlation between a student's assignment and the student's unobserved endowment. For the student fixed effect to mitigate the endogeneity created by the parental lobbying for the student's assignment, one would have to assume that this lobbying was constant across years. Or, as Rothstein (2009) notes, it would be, at least in terms of the student's unobserved characteristics, as if all decisions based on the child's placement were made at the beginning of the kindergarten and never changed throughout the student's education.

However, it seems equally likely that student assignment to teachers could result from dynamic rather than static processes. For example, consider the recent controversy around the publication of value-added estimates for teachers in the Los Angeles Unified School District by *The Los Angeles Times* (Song, Felch, & Smith, 2010). Might astute parents in Los Angeles examine the value-added estimates for their children's prospective teachers and then lobby the principals for assignment to these teachers? To the extent that the principals respond to this pressure, then the assignment is likely to be endogenous.

Rothstein (2009, 2010) proposes a simple falsification test for the presence of student sorting suggested by Todd and Wolpin (2003). He estimates the effects of 5th grade teachers in North Carolina on the 4th grade gains of their students. He finds almost as much variation in the effects of the 5th grade teachers as the student's 4th grade teacher although of course, at that point in time the 5th grade teachers had never taught the 4th grade students. Koedel and Betts (2009a) were able to replicate Rothstein's results on their own sample of students from San Diego. Both use the ratio of the variation of the effect of future teachers to the effect of current teachers as a measure of the size of the bias.

Nevertheless Rothstein's critique has left many researchers unconvinced (Hanushek & Rivkin, 2009). Suppose a class of 4th grade students has a highly effective teacher and as a result their test scores rise. Now suppose that their school sorts students based on their test scores at the end of the year. These students will be assigned to a certain fifth grade teacher because they had such gains in fourth grade. If one conducts Rothstein's falsification test and estimates the effect of the 5th grade teachers on the 4th grade gains, there will likely be a correlation that is induced by the sorting. After replicating Rothstein's findings, Koedel and Betts (2009a) conduct sensitivity tests and conclude that including multiple years of data for a teacher reduces the bias from the sorting significantly in all models and completely in the student fixed effect specification.

In this study, I estimate the extent of sorting of students to teachers across schools and the extent it biases estimates of teacher effects. First a I plot the means and standard deviations of the student scores from t-1 by classroom under three conditions: the actual sorting in the data, a simulation of random sorting within school year and grade, and a simulation of perfect sorting under the same strata. Comparing the means and standard deviations across these scenarios is one gauge of the extent of sorting.

Then, following an approach adopted by a several recent studies (Aaronson, et al., 2007; Hanushek & Rivkin, 2009; Koedel & Betts, 2009a), I identify classrooms that seem to approximate random assignment on observables. The strategy is to identify a subset of classes for which I cannot reject the null hypothesis of no sorting on observables. First, I regress the student test at *t-1* on a vector of indicators for class assignments in each school at *t*. Then, I test the hypothesis that the classroom indicators are jointly significant. Schools in which I fail to reject the null hypothesis will be placed into a sample of schools with random assignment on observables. Then, I rerun the basic specification on this restricted sample. The change in the variation of teacher effects in the restricted sample is an estimate of the effect of sorting.

Implications for Policy

Estimates of teacher value-added are not so interesting in and of themselves, but rather in the context of district policy. Many districts are interested in use of these estimates as one component of compensation reform (Podgursky & Springer, 2007; Springer, 2010). Others will be interested in using the data as a way of increasing the probability that low-performing students will receive high value-added teachers. For example, one could imagine a scenario where a district ceased to monitor the equitable distribution of inputs to the education production function that have trivial effects on student achievement (e.g., the number of VCRs in a building) and turned instead to monitoring the impacts of inputs such as high value-added teachers.

In this section, I explore two such policy issues. First, I assess the potential uses of value-added data to inform district policies of distributing teachers to students. Second, I evaluate the extent to which districts can reliably compare the value-added estimates of teachers for use in personnel decisions.

Distributing Teachers to Students

In this study, I explore the sorting of effective and ineffective teachers to students. Effective and ineffective teachers can be sorted to students across and/or within schools. If teacher effectiveness is primarily sorted across schools, then the district will likely try to incentivize effective teachers to switch schools to even out the distribution, assuming that the teacher's effectiveness is transferable. If teacher effectiveness is primarily sorted within schools, then the district can focus less on movement of teachers and more on matching teachers to students within those schools.

First, I use a variation decomposition approach to estimate the magnitude of the variation in teacher effectiveness within and between schools. I estimate the preferred model for math scores for grades 5-8 in 2010 with and without school fixed effects. The proportion of the variation in the teacher effects that remains after the inclusion of the school fixed effects is an estimate of the within-school variation. To show this graphically, I overlay kernel density plots of the distribution of teacher effects across and within schools. As a way to depict the across-school variation, I plot the range of teacher effects within each school in the district.

Once I establish the extent of across and within school variation in teacher effectiveness, I turn to estimating the extent to which these effective and ineffective teachers are distributed to specific types of students both across and within schools. To test the presence of the sorting, I divide students in grades 3-7 in 2009 into by-grade quartiles of achievement in 2009. Then, I estimate the across and within-school probabilities that these students are assigned in 2010 to teachers whose value-added as of 2009 was in the top 25% or bottom 25% of the district.

The sorting of teachers suggests that subject to the dynamics of the teacher labor market, a superintendent could assign or incent more effective teachers to move to schools with lower concentrations of effective teachers. There is evidence to suggest that teacher effectiveness is portable and not school-specific (Lockwood & McCaffrey, 2009; Sanders, Wright, & Langevin, 2010). If this were the case, moving the effective teachers to schools with less effective teachers could increase student achievement for those students in the receiving schools but could decrease the student achievement for those students in the schools from which the effective teachers are moved. The predicted general equilibrium benefits of this sort of policy would depend, in part, upon (1) the effectiveness of the teachers who replaced the transferring teachers, (2) whether effective teachers are effective across different student sub-types, and (3) whether different sub-types of students are equally responsive to an effective teacher. If a low-performing student benefits more from a highly effective teacher than a high-performing student, then one can imagine a scenario in which from a social welfare perspective, there is a net gain in student achievement. For example, the gains in student achievement for lower-performing students from having a top 25% teacher could be greater than the losses in student achievement from higher-performing students who move from having a top 25% to having an average teacher.

To investigate this possibility, I adapt an analysis from Aaronson, Barrow, and Sander (2007). I use the same groupings of students in grades 3-7 in 2009 as above and calculate the mean gains in their test scores from 2009 to 2010. Then I estimate teacher effects from the preferred model and report the standard deviation of the teacher effect in student level standard deviations. If I divide this standard deviation by the average gain for the group, I have an estimate of the proportion of average gain that is attributable to the teacher effects. If lower performing students are benefiting more from higher value-added teachers then there are possibilities that their gains could offset the losses from other students who are losing their better teachers.

Personnel Decisions

In addition to using value-added estimates to inform its policies of distributing teachers to students, a district may want to use the estimates as a measure of teacher performance in the context of evaluation, compensation and retention policies. The degree of stability of the estimates across time will inform the degree to which they can be used for personnel decisions.

Following much recent work (Aaronson, et al., 2007; Jacob, et al., 2008; McCaffrey, Sass, Lockwood, & Mihaly, 2009), I partition the variation in teacher effects from the preferred model into three parts

$$\begin{aligned}&\delta_{kt} = \theta_k + \xi_{kt} + \varepsilon_{kt}, \text{where} \\ &\text{var}(\theta_k) = \tau^2, \text{var}(\xi_{kt}) = \nu^2, \text{ and } \text{var}(\varepsilon_{kt}) = se_{kt}^2\end{aligned} \tag{15}$$

where θ_k is the part of the teacher's effect that is persistent across time, ξ_{kt} is the part of the teacher's effect that is specific to year t, and ε_{kt} is the sampling error. Then, τ^2 becomes the variation in the persistent part of teacher effectiveness (i.e., the between-teacher variation) and can be estimated as the correlation of teacher effects across time; ν^2 is the variation within teachers over time (i.e., the within-teacher variation); and se_{kt}^2 is variation in the sampling error which can be given by the mean squared error of the standard errors of the teacher effects.

Under this decomposition, I estimate the reliability of the teacher effect estimate (i.e., ratio of signal to noise) as:

$$\text{Reliability=}\frac{\tau^2 + \nu^2}{\tau^2 + \nu^2 + se_{kt}^2} \tag{16}$$

The reliability of the estimate is another way of thinking of the shrinkage factor in the empirical Bayes approach to shrinkage. It is that proportion of the variation in teacher effects that is not due to random or sampling error. The stability of the estimate is the proportion of the variation that is attributed to the time-persistent component of teacher effectiveness. It can be given as:

$$\text{Stability=}\frac{\tau^2}{\tau^2 + \nu^2 + se_{kt}^2} \tag{17}$$

From here, I estimate the between-teacher variation as

$$\text{Between Teacher} = \frac{\nu^2}{\tau^2 + \nu^2} \tag{18}$$

and the within-teacher component as:

$$\text{Within Teacher} = \frac{\tau^2}{\tau^2 + \nu^2} \tag{19}$$

The district could be interested in the latter two estimates as a way of informing its resource allocation between, for example, recruitment/deselection and professional development. For instance, if the proportion of within-teacher variation is low relative to that between teachers, then it could suggest that the professional development activities may have less return than recruiting and tenure policies.

The degree to which these measures of teacher effectiveness are consistent for a specific teacher over time is likely to influence the buy-in from the teacher and the measure's overall usefulness a policy tool. On the one hand, if the estimates are not sufficiently stable, then it is unlikely that the teachers and principals will see much of a signal in them and any use of the measures in an incentive capacity is likely to be undermined. On the other hand, if the measures are not sufficiently nimble or malleable, then it is likely that they will not pick up on changes in the effectiveness that stem from the teacher's effort to improve. This, too, would diminish the signal.

In this study, I follow the literature in assessing the stability of the estimates through (a) the correlation of the point estimates of teacher effectiveness and (b) transition matrices that record the quantile of a teacher's effectiveness in two successive time periods. I run two sets of analyses—one on the whole sample of teachers in the district and the other on a restricted sample of only those teachers present in the district in both time periods. The former is likely to be the more policy relevant to the district—it is what the teachers will see about their performance. The latter deals with selection and attrition dynamics that can add noise to the estimated correlation. For example, a given teacher's performance could look like it varies more than it does if the overall average teacher performance in the district is changing, i.e., the reference group changes

The pursuit of value-added estimates of teacher effectiveness is motivated by districts that seek to distinguish among their teachers' effect on student achievement for reasons as varied as professional development to compensation. Because much of the concern in the literature on value-added models of teacher effectiveness has been in estimating the variation in teacher effectiveness rather than the estimation of individual teacher effects, the issue of how districts should handle the imprecision of the effect estimates has been given less attention than perhaps it deserves (for a notable exception, see Lockwood, Louis, & McCaffrey, 2002; McCaffrey, et al., 2010). The general admonition the literature is to (a) include more than one year of data for the teacher and/or (b) to be cautious of dividing teachers into more than three groups—a large middle flanked by two smaller tails (Lockwood, et al., 2002; McCaffrey, et al., 2003).

Yet this issue requires more thought by the district. For example, in the literature, researchers typically construct 95% confidence intervals around the teacher effects. However, given the counterfactual in which a district has almost no information with which to distinguish teachers (Weisberg, et al., 2009), does it need to be 95% certain that the teacher's effect is above or below average? Or, how should a district balance the risks of committing Type I and Type II errors (McCaffrey, Han, & Lockwood, 2008)?

In this study, I investigate this issue in two ways. First, I estimate teacher effects for fifth grade and eighth-grade math teachers and plot the effects with varying confidence levels around them. I report how many teachers are significantly below or above the average based upon the confidence interval level.

Second, following a suggestion by Doug Staiger (2009), I experiment with an idea of estimating the probability that a teacher is in a given quantile. For example, imagine two teachers whose confidence intervals around their effects both cross zero, but one does just barely and the other straddles the line. Assuming a normal distribution of the error around the point estimate, we can be more confident that the former teacher is above the teach mean. In this approach, I leverage the assumption of normally distributed errors to estimate the probability that a teacher's effect is above or below the teacher mean.

LIMITATIONS

There are a number of limitations both to the study, specifically, and to the policy use of value-added models more generally that the methods described in this section do not solve.

Limitations of the Data

First, the teacher student matching in the data depends on what has been entered into the student information system. The student information is designed to capture at most two teachers who share equal responsibility for a student's instruction in a given subject. Consequently, if a school chooses a different instructional strategy—departmentalization or flexible grouping in which students are rearranged among the teachers episodically throughout the year—then the attribution of a single teacher to the student is likely to be invalid.

A similar concern arises with special student populations such as English language learners or exceptional children. For example, English language learner populations may receive instruction from a homeroom teacher as well as a resource teacher. The value-added models estimated in this study cannot handle this type of joint production. Rather, they ascribe all of the value-added to one teacher.

Another issue is the extent to which student achievement in a given subject is jointly produced by several teachers in the same term across subjects (Jackson & Bruegmann, 2009; Koedel, 2009). For example, it is easy to imagine that the Social Studies instructor could influence reading achievement. Although the data documents the course enrollments, the actual models do not allow for this. The implications are that the teacher effect estimates could be biased.

An additional limitation of the data lies in its provenance—hand-entered by a school-level administrator through the interface of the student information system. (This is not true of the test scores, but is true of most of the other variables, especially the control variables.) Sources of inaccurate data include:

1. The data is mis-entered.
2. The student information system is not designed to provide archival or retrospective data; it is designed for snapshots. As a result, records can be overwritten and this will be unobservable to the researcher. For example, the CMS student information system deletes the course records from the scheduling data base of a student who withdraws from school.

3. There can be incentives for school-level staff not to record certain types of data (e.g., discipline incidents that will reflect negatively on the school in district reports).
4. In at least the case of unexcused absences, the district allows schools to put students with large numbers of unexcused absences through a special program. When the student completes the program, the absences are erased from the student’s record. This is unobservable to the district office.

The potential problems with some of these measures points to a tradeoff for the district. On the one hand, the district may want to control for student heterogeneity by using as many of these measures as is feasible, especially if the district decides to refrain from including time-invariant variables such as gender or ethnicity in its models. Yet each new variable increases the risk of having right-hand-side variables measured with error.

I have limited the scope of the study to students with no missing data. As Table 4 makes clear, this eliminates a number of students who are likely to be different than those who remain in the sample. Although this type of constraint is most likely to affect individual teacher estimates, it is possible that the exclusion of the students with missing data could change the overall variation in teacher effects.

Limitations of Use of Value-added Measures for Policy

In addition to limitations posed by the data, there are limitations to the use of value-added measures for policy. First, the value-added methodology assumes that the tests on which they are based are good measures of student achievement. A number of testing experts challenge this assumption (Koretz, 2002; McCaffrey, et al., 2003) or emphasize how small changes in the scaling of the tests can produce substantial differences in the estimates for particular teachers (Ballou, 2009; Lockwood et al., 2007).

Second, value-added estimates provide measures of effectiveness for a limited number of students. Nationally, the oft-cited figure is that 69% of teachers do not teach in a subject or grade in which a test is administered that can be used to measure student achievement growth. In CMS, that number is 60%. Value-added covers one dimension of teaching for one subset of teachers in a limited number of subjects.

The options for policymakers are twofold: (1) add similar assessments of student learning in other grades and subjects and (2) use the value-added measures to learn what the effective and ineffective teachers are doing, taking advantage of the value-added measure as a validation of the effective practices. Then extrapolate the practices to teachers in non-assessed areas on the assumption that the practices are not subject-specific.

Finally, I do not attempt in this study to test some of the fundamental assumptions of value-added methodology, the types of assumptions outlined by Todd and Wolpin (2003). In particular, I assume that the effects of past inputs (including teachers) do not affect the student's current year achievement beyond what is captured by the prior test score. In addition, I also assume that the dynamic responses of students' families to prior education experiences are orthogonal to the observed inputs of the student's production function in a given year. I assume that the inputs to student achievement in a given year are additive rather than multiplicative. Section 4.3 outlines a number of assumptions required for consistent teacher effect estimates that I do not test in this study.

RESULTS

In this section, I summarize the results from deriving a preferred model, testing it against threats to its validity, and the exploring two potential policy uses.

ARRIVING AT A PREFERRED MODEL

Dealing with the Lagged Score

Columns 1-4 of Table 5 report the results from four prevalent ways in the literature to capture observed and unobserved prior inputs to the student's education production function through the use of one or more prior scores. Column 1 summarizes the model described in Equation (5) which includes one prior test score as a lagged dependent (the Lag (1) model). Column 2 adds another prior score from t-2 (hereafter, the Lag (2) model). Column 3 summarizes the model described in Equation (8), which instruments for the t-1 score with one from t-2 (the IV model). The model summarized in column 4 moves the prior score from t-1 from the right-hand-side to the left-hand-side so that the dependent variable is the change in scores from t to t-1 (the gainscore model).

Examining Table 5, the four models produce similar teacher effects. The standard deviation of the teacher effects (reported in student-level standard deviations of test scores) are consistently between 0.19-0.21 sds when the teacher effects are unadjusted for sampling error and 0.15-0.17 sds when the estimates are shrunken discussed previously. The estimated teacher effects from the models are correlated quite highly as well with the Lag (1), Lag (2), and IV estimators correlated from 0.96-0.99. Each of these is correlated with the noisier gainscore model at 0.84-0.86.

The striking similarity of the results places some of the statistical concerns among value-added researchers in perspective for the district. The primary justification for the

Lag (2) model is mitigate the effect of measurement error in the t-1 score and to allow the coefficient on the t-1 score to be conditional on the t-2 score rather than impose the same coefficient on each student. Yet this model correlates with the Lag (1) model at 0.96. Similarly, the rationale for the IV approach in column 3 is to handle the endogeneity created by the measurement error in the prior test score being on the right-hand-side. For the district, however, this approach results in almost identical teacher effects. It is harder to know what to conclude from the gainscore model results in column 4. It requires no additional complexity in estimation, and it remains highly correlated with the other models (0.84-0.86). From a statistical perspective, the decision comes down to a trade-off between the benefit of the gainscore model (e.g., no measurement error issues from having a lagged dependent variable) versus the cost of the assumption of complete decay of prior achievement discussed previously.

From an implementation perspective, districts may prefer the gainscore model and the Lag (1). Both are less complicated to compute than the IV model. More importantly, each of these models allows the district to include more students and teachers. Although the Lag (2) model results in estimates for only seven fewer teachers and 30 fewer classrooms over a three-year period, it results in 8,061 fewer student-by-year observations and eliminates 4,517 students from the estimation. As a result the district could be creating an incentive for teachers to ignore these students. As seen previously in Figure 1, these students are likely to be lower-performing than their peers and so the district could exacerbate its efforts to raise the achievement of its lowest-performing students. A larger issue of excluding teachers and students in the Lag (2) and IV models is that in NC, no fourth-grade students and teachers can be included in the estimation because there is no second-grade test.

Accounting for Student Heterogeneity

All of the models summarized in columns (1-4) assume that the student's prior score(s) also capture the effect of the student's unobserved endowment. Columns 5-6 of Table 5 go further through a student fixed effect [Equation (11)] and first difference approach [Equation (12)], respectively. The first difference approach in column 6 uses the same dependent variables as the gainscore model in column 4 but differs in transforming all the left-hand side variables into first differences, including the teacher effect. This means that the teacher effects it estimates are actually the difference in teacher effect from a student's teacher at t-1 to the teacher at t.

From a statistical perspective, these methods are correlated similarly less strongly (0.67-0.76) with the models in columns 1-4. On the hypothesis that noisier estimates of teacher effects, especially in the student fixed effects specification, could result in lower correlations, I checked the correlations of the estimates adjusted for sampling error and there was little difference in the magnitude of the correlations. In addition, both models seemed to decrease dramatically the impact of the time-varying student characteristics.

Both estimators also produced teacher effects with larger unadjusted standard deviations of teacher effects. The effect of adjusting for sampling error on the student fixed effect teacher estimates was more pronounced than the other models, which is consonant with the conventional wisdom in the value-added literature that the student fixed effects result in noisier estimates of teacher effects. The first difference estimator is the only one in which there was not an appreciable difference between the unadjusted and adjusted teacher effects.

The correlations of the teacher effects from the first difference and student fixed effects models are moderately correlated at 0.62. Given that the presence of serial correlation would indicate a preference for the first difference estimator, I conducted the test for serial correlation developed by Jeffrey Wooldridge and implemented using Stata's –xtserial— command (Drukker, 2003). The F-statistic for the test of the null hypothesis of no first order autocorrelation was 2027.44 resulting in a strong rejection of the null of no first-order autocorrelation. The implication is that the first difference estimator is the more appropriate estimator because the unobserved endowment will not be correlated with the lagged score.

In selecting a base model to move forward in the analysis, I chose the Lag (1) model over the first difference model for the ease of interpretation and computation. The first difference model requires estimating fixed effects for every combination of teacher from t and t-1 and extrapolating the teacher's value-added at t from those combinations. I move forward to look at modeling classroom and school heterogeneity with the base model summarized in column 1 of Table 5.

Accounting for Classroom and School Heterogeneity

Table 6 summarizes three models that attempt to take into account heterogeneity among classrooms and schools in estimating teacher effects. Column 1 brings forward the base model (Table 5, column 1) for controlling for student heterogeneity. Column 2 adds a vector of classroom means to the model in column 1. Column 3 adds an additional vector of school-level means. Column 4 excludes those school means and replaces them with school fixed effects.

In every specification, the additional characteristics were jointly significant at p<.001. The smaller F-statistics for the student controls indicate that some of the work being done by the student controls in the base model in column 1 was really the effect of the classroom and school composition (I excluded the individual student from the calculation of the classroom means). Nevertheless, the additional controls did little to change the overall fit of the model, consistently explaining 77% of the variation in student test scores. Nor did the presence of the classroom (column 2) and school covariates (column 3) change significantly the estimated teacher effects; they remained highly correlated at 0.96 and 0.89, respectively. The correlations in the school fixed effects models were significantly lower. This is to be expected, however, as this becomes the correlation of a teacher's effect relative to the district's teachers to the effect relative to other teachers in the teacher's school.

Although the correlations across the models were large, it may not be enough for the district to conclude that it is indifferent among the models. Even with the high correlations, it is possible that the teachers whose ratings moved significantly between the models could share characteristics that the district will need to heed from a policy perspective. A district would likely want to explore the cases of outlier individual teachers whose scores changed significantly between the models to identify patterns that could require a policy decision.

Modeling Teacher Effects

Up to this point, I have estimated the teacher effects as fixed effects pooled across all the teacher's classes. As discussed in Section 4.3.5, most researchers adjust these fixed effects estimates to account for the sampling variation and/or across-year or class variation in the effects. Table 7 compares the effects of two types of shrinkage on the distribution of teacher effects. Columns 1-2 are the unadjusted and shrunken standard deviations of teacher effects from the preferred model summarized in column 3 of Table 6. The estimates are shrunken by the procedure outlined previously. Column 3 replicates the model in column 2 but estimates the teacher effects as random effects rather than fixed. An assumption of the random effect estimator is that the random teacher effects are orthogonal to the inputs. In the teacher fixed estimates, there is evidence that this assumption is violated; the correlation between the teacher fixed effects and the left-hand-side variables is 0.11. Nevertheless, the teacher effects from these two estimators are correlated at 0.98.

Column 4 adds a classroom random effect to the model in column 3. Here, the classroom random effect is nested within the teacher random effect. In the case of teachers with only one class per year, the classroom random effect is equivalent to a teacher-by-year effect. Whether the teacher has one or many classes per year, the intent of the classroom random effect is to account for non-persistent variation at the classroom level—perhaps due to an especially good or poor match of the teacher and students in that specific class or perhaps a classroom-level shock out from the teacher's effect—from the teacher's persistent effect. In this case, including the classroom random effect reduces the variation in teacher-level effects slightly and a likelihood ratio test provides evidence that the inclusion of the classroom random effects improves the model fit. The correlation between the models with and without the classroom effects is strong. Figure 3 shows the effect of the shrinkage estimators on the distribution of teacher effects. As expected, shrinking the teacher effects (column 2 vs. column 1) narrows the distribution of teacher effects. Including the classroom random effects tightens the distribution still further.

A district faces a trade-off in deciding whether or not to shrink the estimates. On the one hand, the shrinkage estimators pull towards the mean teachers at the tails that might be their due to sampling variation. Figure 4 shows the sensitivity of the magnitude of teacher effects to the numbers of student attributed to the teacher across all years in the sample. Comparing Panels A-B, it is clear that the shrinkage estimators affect teachers with fewer numbers of students. These teachers could be novice teachers, elementary school teachers, or teachers in schools with smaller class sizes (e.g., a Title 1 school). By using a shrinkage estimator, the district could potentially underestimate the effect of a highly talented novice teacher or overestimate the effect of another teacher who is teaching a small number of students, in both cases by pulling the teacher toward the middle. In some ways, the district must decide whether it prefers avoiding a Type I error by choosing a shrinkage estimator versus a Type II error in which it fails to recognize a truly good or poor teacher.

The district faces a similar trade-off in deciding whether or not to include the classroom random effects. Choosing to include them helps buffer teachers from classroom-level shocks that could idiosyncratically change their effects (e.g., a problematic student that the teacher spends extraordinary amounts of time with.)

However, especially for teachers with only one class per year, using the classroom random effects could attenuate the true improvement of the teacher. For example, a portion of a second-year elementary school teacher's improvement will look like a classroom level shock and be partially netted out of the teacher effect. For teachers who are accustomed to quick changes year-to-year, this smoothing out of the effect could be frustrating.

In the end, I chose the model summarized in column 4 of Table 7 as the base model to be used throughout the rest of the dissertation. As the results in Table 7 demonstrate, there is no clear winner among the models there. They are all highly correlated. The advantages to the district of the teacher and class random effect model are the elimination for the need for post-estimation shrinkage of the teacher effects and the intuitive appeal of sweeping out non-persistent variation in a teacher's effect (i.e., the classroom random effect) which should lead to greater stability of the estimates.

Including Race Covariates

For many districts, the decision about whether to include the race of the student and/or the racial composition of the classroom and school is fraught with political implications. Many worry that including the race covariates is tantamount to having different expectations for students based upon their race. To test the effect of race covariates on the teacher effects from the preferred model, I estimate three additional variants that include student, classroom and school race covariates. Table 8 reports the results. Including race covariates did not seem to have an impact on results from the base model. It did not change the distribution of the teacher or classroom level effects. It does not reduce the student-level error. The teacher effects from the various models were correlated at no less than 0.988 with the base model. Likelihood ratio tests did indicate that the model fit improves with the presence of student and school-level race covariates. Classroom racial covariates were not statistically significant in any of the specifications. The results indicate an approach that districts may adopt to help mitigate potential conflicts over the inclusion of certain covariates: test for their effect before having a longer philosophical debate about their inclusion.

TESTING THE PREFERRED MODEL

With the preferred model established, I proceeded to test it for three particular threats to its validity—prior inputs to the education production function, ceiling effects in the test instrument and sorting of teachers to students. I am assuming that the analysis completed in this section would apply equally well to the other models considered in the preceding section. However, I do not test this assumption on the other models.

Testing the Prior Score Assumption

Before moving to testing the preferred model against two primary threats to its validity, I test the assumption that the student's prior score captures the effects of all prior educational inputs. The primary assumption behind the preferred model is that the prior score captures the impact of previous inputs to the student's educational production function. Table 9 summarizes tests of this assumption. To test the assumption that the score at t-1 captures all previous inputs from t-1 and t-2, I estimated the Lag (1) model and included classroom and school means of variables from t-1 and t-2. Wald tests of the joint significance of the t-2 variables provide strong evidence that observed characteristics at t-2 of the student predict the student scores at t conditional on the score and inputs from t-1. There is weaker evidence that the school characteristics from t-2 affect the score at t, but they are statistically significant at the 0.05 significance level. A likelihood ratio test suggests that including the t-2 inputs improves the fit of the baseline model.

I also tested the effect on inputs from t-2 of adding the t-2 score on the left-hand-side (column 3). The student, classroom and school inputs from t-2 remain statistically significant, and a likelihood ratio test provides evidence that this model fits the data better than the Lag (1) model. So in either case, the assumption that the score at t-1 and/or t-2 captures the effect of prior inputs appears to be violated. To test the significance of this violation, I calculated the correlation of the estimated teacher effects. The teacher effects from the model that included the t-1 score and t-2 inputs were correlated at 0.99 with the same model that excluded the t-2 inputs. The model that included the t-1 and t-2 score as well as the t-2 inputs was correlated at 0.96 with the model that excluded the t-2 score and inputs. The results were highly correlated, but there is evidence that the violation of the assumption would affect the estimates for some teachers.

Ceiling Effects

In this section, I report the results of the analyses concerning the potential of ceiling effects in the test instrument to bias the teacher effects for teachers with students who enter their classroom already at the high end of the distribution.

Figure 6 shows the distributions of student scale scores by grade for the current year and prior year. The extent of negative skewness could indicate the potential for a ceiling effect. Using one year of data (2010) as an example, we find that the greatest skewness in grades 5 & 8, although the magnitudes are relatively small. There seems to be little evidence of censored distributions either in the current or prior year test score.

As noted previously, comparing the gains from t-1 to t by level of score at t-1 points is more of a measure of regression to the mean rather than the effect of a ceiling. By comparing the gains from t-1 to t by score at t-2, we can see more clearly the potential for a ceiling effect. Figure 7 provides a second assessment of the potential for ceiling effects by using boxplots to compare the average gains in student achievement at time t with the gains the student made from t-2 to t-1. The shaded box represents the interquartile range (75th percentile is top, 25th percentile is bottom) of the average scale score gains for students in each decile. The vertical lines emanating from the shaded box are the upper and lower adjacent values, which by convention extend 1.5 times the difference in the 75th and 25th quartiles. The dots indicate values beyond the upper and lower adjacent values. Students in the bottom and top deciles have some of the highest average gains two years later. The variation of the gains among the deciles of prior achievement generally decreases as we move up the distribution of prior achievement. Students in the top 10% of achievement at t-2 have the highest average gains but also the smallest variation in those gains.

Figure 8 maps the relationship between the mean score at t-1 of a teachers' students and the teacher's value-added at t. There is no evidence of a linear relationship and this is not surprising because the classroom means of prior achievement were included as a covariate in our base model. However, the figure does show what seems to be smaller variation in the teacher effects when the mean prior achievement of the students is greater or less that one standard deviation below their peers.

To explore this further, Figure 9 plots the relationship of the proportion of a teacher's students in the top and bottom 10% in prior year achievement to the teacher's effect in year t. I weighted each teacher by the number of students the teacher instructed during the period. It is possible that teachers with high proportions of students in the top quantiles of prior achievement taught fewer students. If this were the case, the smaller variation of the estimated effects of these teachers could be due more to the shrinkage estimator than the possibility of a ceiling effect. This would be equally in the case of teachers with high proportions of students who are in the bottom quantiles.

Figure 9 provides mixed signals. On the one hand, there appear to be thresholds of proportions of students in the top quantiles above which teacher effects seem to move toward zero. It seems to be the case that many, but not all, of the teachers above this threshold have fewer students and effects generally closer to the mean. The pattern is similar in looking at the teacher effects of students with high proportions of students in the bottom 10% of the previous year, although it seems that these teachers generally have fewer students than their peers. In both cases, we see less variation in the teacher effects at the tails of the distributions of students and it seems that at least part of this is due to these teachers having fewer students.

To probe deeper, I examined how many students scored high enough on test at t-1 that it would be impossible for them to increase their scores by the amount of the effect of the highest value-added teachers. For example, for eighth grade math in 2010, the highest value-added teacher added approximate 3.5 scale score points. I counted how many students were within 3.5 scale score points in year t-1 of the maximum score in year t. The intuition is that if a teacher had a large proportion of these students, the teacher could not be the highest value-added teacher. In the sample, across all grades and years, 0.32% of the students could not have raised their scores by the size of the effect of the highest value-added teacher. Of these students, 15% hit the ceiling the next year. So across three years and five grades, 76 of 89,300 students hit the test score ceiling before they could move up by the amount of scale score points added by the highest value-added teacher.

As another test of the sensitivity of the teacher effects, I re-estimated the preferred model with samples trimmed at the 99th, 95th, and 90th percentiles of the student's prior achievement. The results are summarized in Table 10. The standard deviation of the teacher effects was unchanged across all specifications and the correlations of the teacher effects across the models was greater than 0.99. Trimming high-performing students did not exclude any classrooms or teachers from the sample.

The differences in the models for a district lay primarily in changes in the percentile ranking that are wrought from trimming the samples. Although the mean difference in percentile rankings for teachers in any of the trimmed sample models from the untrimmed sample was zero, there was some movement. For example, trimming the top 10% would result in approximately 95% of the teachers experiencing a percentile rank change of less than or equal to 7.8 percentile points. Depending on how the district grouped its teachers, this could be more or less significant.

As a final check on the impact of ceiling effects, I examined the impact of an alternative way of standardizing the student's current year score. Given the difference in gains among students in different deciles of the distribution of scores at t-1, I re-estimated the base model but standardized the score at time t with by the decile of the score at time t-1. Table 11 summarizes the results. I use this alternative normalization of the current year score in the models in columns 2, 4, and 6. In addition, to check the sensitivity of the ceiling effects to the number of students a teacher has, I re-estimate the model with one, two, and three years of data. The intuition, following Koedel and Betts (2009a), is that additional years of data should mitigate the potential of sorting to bias teacher effects (in this case by reducing the teachers' proportions of students at the tails of the distribution).

Table 11 shows that the alternative normalization increases the standard deviation of the teacher effects by one-third, and the classroom effects by an even greater amount. This could be a statistical artifact of the alternative normalization or possibly an indication that there is greater variation in teacher effects at different points of the distribution of prior student scores. Examining the correlations among the models, it seems that having multiple years of data for a teacher makes more of a difference for a teacher's effect than the alternative normalization. Across all teachers the effects generated from one-year versus three years are correlated at 0.69-0.76. The results of the alternative normalization are highly correlated for all teachers, with slightly weaker correlations for teachers with large numbers of students who were in the top decile of achievement in the prior year. Table 11 suggests that a district concerned about ceiling effects could mitigate any negative bias by including more years of teacher data as a way of smoothing out any shocks to the teachers' classroom composition due to sorting.

Sorting

In this section, I report the results of the analyses outlined above concerning the potential of sorting of students to teachers to bias the effects of some teachers. To begin the exploration of the potential of sorting of teachers to students to bias estimates of teacher effects, I considered three sorting scenarios that use the student's score at t-1 as the sorting criterion. The first is that of perfect sorting in which within each school-year-grade combination, students are sorted into classes strictly by their scores at t-1 while preserving the original class sizes. The second is similar to the first except that now students are sorted randomly into classes. The third scenario is the actual classroom assignments in the data.

I began by calculating the standard deviations and means of the teachers' students' scores at t-1 as a measure of the sorting of students to teachers in each of the three scenarios. Figure 11 shows the distribution of these means and standard deviations across teachers. Actual sorting in the sample results in average classroom standard deviations of prior test scores that are closer to distribution of standard deviations from the simulated random sorting than the perfectly sorted simulation. In Table 12, I re-estimated the preferred model using the perfectly and randomly sorted samples. The distribution of teacher effects was slightly larger in the simulated samples and this seemed to result from a slightly smaller variation in the classroom-level random effects.

Next, following Hanushek and Rivkin (2009), I created a subsample of classrooms that do not appear to be sorted on the student scores at t-1. I regressed the student scores at t-1 on each classroom within each year, school and grade. If the F-tests of joint significance of the classroom indicators failed to reject the null at p<.05, I considered that school-grade-year's classrooms to be not-sorted. This subsample included 15,159 student-year observations (approximately 17% of the full sample) containing 830 classrooms, 473 teachers, and 210 school-year-grades.

Then, I ran a series of regressions on these samples in which I estimated the effects of the teacher in time t and t+1 on student scores at time t. As discussed above, the intuition is that the student's teacher at t+1 should not have an effect on test scores from time t. Table 13 reports the results. Columns 1-2 summarize results from two regressions using the preferred model for the full subsample of students from 2008 and 2009 for which we have data on classroom assignments in the following year. Column 1 is the preferred model using the teacher and classroom at time t as the random effects. Column 2 repeats the estimation but substitutes the teacher and classroom at time t for those at time t+1. Columns 3-4 repeat the procedure but use the subsample of classrooms from Columns 1-2 that are sorted on student scores at t-1. Columns 5-6 replicate the analysis on the smaller subsample of classrooms that were not sorted on student scores at t-1.

In each sample, the teacher and classroom at t+1 did predict student achievement at t as would be expected if students sorted to classrooms in t+1 based on their scores at t. The variation in teacher effects at t+1 for the full and sorted samples was approximately 80% of that of the variation of teacher effects at t. Even in the non-sorted sample, the t+1 teacher effects had roughly 60% of the variation of the teacher effects at t. This is surprising because this is the sample in which we could reject sorting among classrooms based on student scores at t-1. This suggests that there could be sorting on other observables or unobservables that are biasing the teacher effects. Finally, a puzzling result is the coefficients at the student test scores at time t-1 for the non-sorted subsample in columns 5-6. Note that the coefficients on the score become insignificant both statistically and substantively.

In sum, the evidence for sorting bias is mixed. On the one hand, the simulations of perfect and random sorting resulted in slightly wider distributions of teacher effects. This seemed to be related to an accompanying narrowing of the distribution of the classroom effects. On the other hand, the replication of the Rothstein falsification test indicated that the effect of teachers at t+1 on student scores at t was significant, even in the non-sorted sample.

POLICY IMPLICATIONS

In this section, I turn to exploring two possible policy uses of value-added estimates—as a measure of equity for students and as an input into personnel decisions.

Distributing Teachers to Students

I began by estimating kernel densities of the distributions of the teacher effects across and within schools in Figure 13. The across and within school teacher effects are derived from model summarized in Table 7, column 4 which is the preferred model. The within school distribution teacher effects came from re-estimating the preferred model and adding school fixed effects. The standard deviation of teacher effects in the base model was 0.15 student level standard deviations. For the model with school fixed effects, the standard deviation of the teacher effects decreased by 20% to 0.12 standard deviations. Figure 14 provides boxplots of the teacher effects for all schools in the sample over the period 2008-2010, sorted from left to right by the mean teacher effect for each school. The between school variation in teacher effects is approximately 0.07 student-level standard deviations and the within school is 0.11. Both Figure 13 and Figure 14 provide evidence that there is substantially more variation in teacher effectiveness within schools than between them.

Knowing that there is substantial variation both within and between schools, a district will want to monitor the sorting of students to high- and low-value added teachers both within and across schools. The district will need to ensure that (1) high-value added teachers are at every school and (2) within the schools, these teachers are distributed across students in a way that meets the district's policy objectives.

Figure 15 shows the distribution of the district's top (Panel A) and bottom (Panel B) quartile value-added teachers across schools. Each circle represents one school weighted by the school's student enrollment during the period 2008-2010. There were a number of schools with no teachers in the district's top (Panel A) or bottom (Panel B) quartile of teachers during the period, and many of these schools had smaller student populations. There seemed to be a negative relationship between the proportion of teachers in the top 25% of the district and the proportion of the school's students qualifying for free or reduced lunch. There was almost no relationship between the proportion of bottom 25% teachers and the school's proportion of students eligible for free and reduced lunch status.

Following an analysis by Aaronson, Barrow and Sander (2007), I explored whether some types of students benefited more from good teaching than others. Table 14 summarizes the results. Each column represents a estimation of the base model for teacher effects restricted to the sample indicated by the column header. For example, column 1 reports results from students who scored in the bottom quartile in the previous year. The standard deviation of their test scores in scale score points was 6.4 and the mean gain in their scores from t-1 to t as 6.8. For their group, a one standard deviation change in teacher effectiveness resulted in a 0.13 standard deviation increase in their test scores, or 0.84 scale score points. The proportion of the mean gain in scores by students in the lowest quartile from t-1 that could be attributed to the teachers was then 0.84/6.8=.12. So, twelve percent 12% of their gain was associated with the difference in a teacher at the 50th percentile vs. 84th percentile of the distribution of teacher effectiveness.

Columns 1-4 of Table 14 show that the proportion of mean gains attributable to the teacher were the smallest for the students in the bottom quartile relative to the other quartiles. This is largely due to the largely average gains made by the lower quartile students that do not yield large variation in the effect of the teachers a fact borne out in Figure 10. The higher gains associated with this group could stem from some sort of mean reversion that is being netted out of the teacher effect. Indeed the proportion of the mean gain attributable to teachers in the top quartile is 2.5 times greater than that of the teachers of bottom quartile students. The results provide some evidence that in math students in the top 50% of performance coming into the year will benefit more from teaching than those in the bottom fifty percent.

Columns 5-8 show the results by ethnicity. I found more homogeneity across ethnic groups than achievement groups. The standard deviation of teacher effects was similar across groups as well as the proportion of the mean gain that could be attributable to the teacher. This suggests that teachers do not seem to matter more or less for different ethnic groups.

Table 15 extends the analysis further by summarizing the difference in probabilities that certain types of students are assigned a top or bottom quartile teacher from the prior year or a teacher whose prior year data is unobservable. Each column reports the results from a probit estimation of the probability of being assigned a top (columns 1-2) or bottom (columns 3-4) quartile value-added teacher from the prior year. Columns 5-6 estimate the probability of being assigned a teacher whose prior year value-added was unobservable (e.g., a novice teacher). The results in columns 1-3-5 reflect across and within school differences. Columns 2-4-6 reflect within school probabilities by including school fixed effects in the specifications for columns 1-3-5.

The samples for these estimations were smaller than the sample used throughout this study for estimating the base model. I used only 2009 and 2010 because I needed to have a value-added score from the teacher at t-1. The sample was further reduced in within-schools estimations in columns 2-4-6. Approximately 29% of the schools had no top quartile teachers and 39% had none in the bottom quartile, and 7% had no teachers who were missing a value-added score from the prior year, and as a result these schools were eliminated from the estimation.

Even with the reduced sample sizes, a district could learn a great deal from this simple analysis. There were no differences among the groups in their probability of being assigned a bottom quartile teacher and only a few in terms of the probability of being assigned a teacher with unobservable value-added data. There were differences in exposure to top quartile teachers. For example, white students were 5% more likely to have a top quartile teacher from the prior year and this discrepancy persisted, although with a smaller magnitude, when looking within schools. Whites were also 4.3% more likely to have top quartile teachers than Hispanic students but this difference eroded when looking within schools.

There are interesting differences when comparing ethnicities by quartile. Within schools, white students who scored in the top quartile in the previous year were 2.7% less likely to be assigned a top quartile teacher than black students who scored in the top quartile in the previous year, but 2.3% more likely to be assigned a teacher with a known value-added measure from the prior year. Conversely, within schools, white students from the top quartile were 2.3% more likely to be assigned a top quartile teacher than Hispanic students from the top quartile.

Perhaps the starkest discrepancies came from comparisons of black students from the top and bottom quartiles of achievement in the prior year. Black students from the top quartile were 6.8% more likely to be assigned a top quartile teacher than black students from the bottom quartile. Most of the difference was coming is coming from assignment patterns within schools. Similarly, black students from the top quartile were 3.2% less likely than black students from the bottom quartile to be assigned a teacher with missing prior year value-added data and this difference persisted within schools.

Personnel Decisions

A far more controversial policy use of teacher value-added estimates is in personnel decisions such as retention and compensation. In this section, I explore how useful these measures might be for districts.

I began by decomposing the variation in teacher effects between and within teachers. In Table 16, columns 1-2 show that the proportion of between-teacher variation in teacher effects was twice as large as that within teachers. This could suggest to the district that selection rather than professional development may be a greater lever for increasing the overall effectiveness of its teaching workforce. Following the decomposition outlined previously, I found that approximately 81% of the variation in teacher effects was attributable to either between- or within-teacher variation (the remaining portion to estimation error of the teacher effects). Fifty-one percent of the variation was persistent across years. The signal in the variation of teacher effects seemed to dwarf that of current evaluation systems in which 98% of teachers are rated as satisfactory or above.

Districts will also want to know to what extent the estimates of the teacher effects are stable across time. Table 17 provides an analysis based upon the movement of a teacher from one quartile to another from one year to the next. The transition tables in Panels A-C report the movement of teachers among quartiles from 2009 to 2010. Panel A places teachers in quartiles in 2009 and 2010 based upon single-year estimation of the teacher effects. Panel B places them in quartiles based upon an estimation using the teacher's pooled data from multiple years (i.e., the 2009 quartile is based upon available teacher data from 2008-2009; the 2010 quartile is based upon available teacher data from 2008-2010). Panel C replicated Panel B but restricted the sample to only teachers who had data in all years.

The stability of the estimates was weakest when the teacher's placement in a quartile depended on one year of data. The between year correlation was 0.48 with 8.9% of the teachers in the bottom quartile in 2009 moving to the top quartile in 2010 and 5.5% moving from the top to the bottom quartile in one year. Approximately 52% of the teachers in the top quartile stayed there a year later, and 42% of those in the bottom quartile stayed. The stability of the estimates increased dramatically when using more than one year of data in Panel B. Just 0.4% of the teachers moved from the top to the bottom quartiles or vice versa. Approximately 82% of the teachers who were in the top quartile stayed there, a similar percentage as for those who remained in the bottom quartile. The teacher effect point estimates were correlated at 0.94.

Panel C provides a check to see how much of the movement among quartiles could be a result of a selection effect of teachers moving in and out of the sample. The transition matrix from Panel B was replicated, but the quartiles were calculated using teachers who were present each year from 2008-2010. The results were similar to those of Panel B. Fewer teachers seemed to be moving more than one quartile and more are moving one quartile.

A chief use of the teacher effect estimates for a district may be to distinguish between the effect of teachers for the use of high-stakes rewards or sanctions. To that end, a district will need to grapple with the uncertainty around the estimates. Figure 16 provides a way of examining district options for the confidence intervals to categorize teachers into distinct groups based upon their effects. For example, in our sample a 95% confidence interval around the teacher effect estimates distinguished 14% of the teachers as above the mean and 13% teachers below. Some would argue that given the state of teacher evaluation in which the vast majority of teachers are deemed effective, that we do not need to be 95% certain that a teacher is above the mean for us to deem the teacher above average. The other panels in Figure 16 show how many teachers are distinct from the mean at 90%, 85%, and 80% confidence intervals. If a district moved to an 80% confidence interval, then it could place 44% of its teachers below or above the mean.

Some argue that this way of using the confidence interval ignores important information about the likelihood that a given teacher effect estimate is above or below the mean even if its confidence interval includes the mean. In Figure 17, I estimate the probability that a teacher's effect is above the mean on the assumption that the estimation error around the estimate is normally distributed. Using this approach, the district could be 90% or more confident that approximately 22% of its teachers are above the mean and about 80% or more confident that approximately 30% of its teachers are above the mean. In either case, this approach allows the district to expand the number of teachers it can label distinct from the average. It could also be applied to other thresholds, such as the probability that a teacher is in the top quartile.

DISCUSSION

The potential of value-added models for use in measuring teacher effectiveness should be evaluated in light of the current state of teacher evaluation in which nearly 98% of teachers nationally are rated as satisfactory or above despite large differences in their impact on student achievement. Value-added measures are not perfect measures. They cover only teachers who teach subjects and grades for which there are standardized tests. They assume that the standardized tests are telling the district something about the learning that happened in the classroom. The imprecision of the estimates give pause to some.

And yet these limitations should be seen in context. Many of the other measures proposed such as classroom observations or student work samples are also imperfect. A classroom observation that occurs maybe four times for a total of one or two hours over the course of 180 instructional days has confidence intervals around it as well. They are subject to factors outside the control of the teacher, too, such as the subjectivity of the observer. They are stable across time only because there is little variation in the results, period. Each measure is going to have shortcomings. For any proposed measure, the district must ask whether it adds information to what it currently knows about its teachers.

Many districts have concluded that value-added measures of teacher effectiveness provide more information about teachers than they have presently. The decision to pursue value-added measures by a district raises a number of decisions for the district. These decisions often have neither a wrong or right answer, but instead reflect a trade-off among viable policy options. The central questions addressed in this study are

1. What are the benefits and costs of various value-added models in terms of the identification and specification of teacher effects?
2. How serious are three often-cited threats to the validity of value-added estimates—influence of prior inputs, ceiling effects in the test instrument and the sorting of teachers to students? What can be done to mitigate the risks they pose?
3. Are value-added estimates suitable for use in (a) considerations of a district's equitable allocation of its resources across students, and (b) personnel decisions?

Benefits and Costs of Various Models

The results of the study may provide some comfort to districts overwhelmed by some of the statistical arguments within the academic community on the proper specification of the value-added models. In most cases, the resulting value-added estimates seemed to differ very little across specifications. In terms of modeling student heterogeneity, the results were largely insensitive as to whether the district includes one or two prior scores as predictors of the current year score. Instrumenting for the score at t-1 with the score at t-2 in an effort to deal with measurement error also resulted in little difference in the effect estimates. Student fixed effect and gainscore models were less correlated with the results from the simple lagged score models largely due to the noisier estimates created by these estimators.

Given the similarity of the results, a district is freer to choose a model that allows more students and teachers to be included in the estimation. The models that required a second lagged score eliminate a significant portion of students who are missing that for two reasons. First, some students will be missing those prior scores because they are new to the district and they are likely to be lower performing on average. Second, for example, in Charlotte-Mecklenburg Schools, the decision to require two prior lagged scores would eliminate all fourth-graders from the estimation (there is no second grade test). Note that the student fixed effects models also restrict the number students and teachers included. The identifying variation comes from students who switched teachers implying again that fourth grade students who are not repeating the grade with a different teacher will not be included in the estimation. So the district can choose an approach that is both simpler to calculate (e.g., the Lag (1) model that requires only one prior test score) and includes more students and teachers and which will result in very similar results to the more complicated models.

The district faces similar flexibility in choosing how to measure classroom and school heterogeneity. The models with additional classroom- and school-level characteristics covariates correlated strongly with the model that included only student level characteristics. As one would expect, adding school fixed effects did change the results significantly because the teacher effects were being identified only within schools, making across-school estimates impossible. The trade-off for the district is that only the school fixed effects can handle the unobserved school characteristics (such the effects of a great principal) that could otherwise be included in the teacher's effect estimates, giving some teachers an advantage over others just based on the school they served. The general flexibility provided by the models allows districts to work with their stakeholders, primarily teachers, in deciding which classroom and school characteristics to include.

Similarly, the results of estimating different forms of teacher effects yielded very similar results. The estimates from fixed and random teacher effects yielded almost identical estimates, perhaps due to the large sample sizes. The result of the teacher random effects, especially when adding the classroom random effects, was a normal distribution that was much tighter around the mean. The tighter distribution makes it more likely that the district will fail to identify teachers who are significantly above or below the mean of teacher effectiveness (the tails shrink toward the mean). At the same, time, it is likely to improve the inter-temporal stability of the estimates which could result in more teacher buy-in for the use of the value-added data.

THREATS TO THE VALIDITY OF INFERENCES

After establishing a preferred model, I tested it for its sensitivity to three potential threats to the validity of its estimates.

Influence of Prior Inputs

I did find evidence that a student's prior score was not capturing the effect of all prior inputs. This violates a central assumption of value-added models that incorporate lagged scores. It was unclear, however, how this violation resulted in different teacher effects for a district. Conditional on the prior score, including the prior inputs did not measurably change the resulting teacher effects.

Ceiling Effects

A common fear of teachers is that value-added measures penalize teachers whose students who enter the year at the upper end of the prior test distribution. In the study, I conducted several tests on the potential severity of the ceiling effect to bias the teacher effect estimates and the results were mixed. On the one hand, I found little evidence of right-censored distributions in examining current and prior year scores. Further, students in the top decile at t-2 had the highest median gain from t-1 to t but the smallest variation in those gains. I found no relationship between the mean scores of a teacher's students at t-1 and the teacher's effect on those students at time t. I estimated the number of students who hit the ceiling in year t before they could have yielded their teacher the highest teacher effect that year was only 76 of 89,300 students. And I estimated teacher effects trimming the sample of the students at the 99th, 95th and 90th percentile of the students' scores at t-1. The correlations between the models were extremely high, although there was some movement in the rankings for some teachers.

On the other hand, a few of the tests did indicate that ceiling effects could be biasing the results. Teachers with large concentrations of students from the top quartile or decile of the prior year score did have smaller variation in their effects. One explanation is that for high performing students the tests offers less range for the for the students to score (i.e., they are going to be scoring high anyway so that the margin of effect for the teacher is much smaller, perhaps even depending on how the students answer very few questions).

I also estimated a series of models that normalized the student's score at t by the decile of their score at t-1. The correlations of the effects were high for all teachers (0.87) and slightly lower for teachers with high concentrations of previously high (0.84) or low (0.81) students. The results suggest that the district that wanted to minimize any potential ceiling effects might be better served by focusing on including more than one year of data for a teacher in the estimation. This could reduce the teacher's proportion of the number of students at the upper and lower ends of the distribution.

In sum, the risk of ceiling effects biasing teacher effects in this district was minimal and districts could address the minimal risk by including more than one year of teacher data in the estimations.

Sorting of Teachers and Students

The threat that the sorting of students to teachers could bias teacher effects has received considerably more attention than ceiling effects. To test the severity of any bias due to sorting, I focused on sorting based on student scores at t-1. I found that the variation in teacher effects was very similar (0.15-0.16 sds) over simulated conditions of sorting that yielded very different within-class variation in student prior test scores.

At the same time, I did find evidence that sorting on test scores did seem to lead to the seeming impossibility of the teacher at t+1 having an effect on the student at t, both in samples sorted and not sorted on prior test scores. This would suggest that some sort of sorting on unobservables that are not being captured in the student's prior test scores is occurring and that the teacher results could be biased by this sorting. To the extent that students were sorted based on their score at time t to their classroom at t+1, then we would expect that at time t+1, the teacher will look as if they had an impact on the score at t. The results of this study are less reassuring than recent work would indicate (Koedel & Betts, 2009a). There are limits to testing the severity of sorting bias in the observational data. The real test of the threat of sorting will come from experiments using random assignment of teachers such as current national studies being conducted by the Gates foundation and Harvard's Center for Education Policy Research.[3]

[3] See the Measuring Effective Teaching project at http://www.metproject.org/ and the Harvard at http://www.gse.harvard.edu/ncte/default.php.

POLICY IMPLICATIONS

Finally, I turned to two possible uses of the value-added estimates for the district.

Distributing Teachers to Students

I found that value-added estimates can provide districts a great deal of information that can inform its policies to ensure that students get access to its most effective teachers. A district wanting to ensure that every student has access to its highly effective teachers will need to ensure both that every school has these teachers and that within these schools every student has access to them. I found evidence that in CMS schools there was significant variation in high value-added teachers across schools and that schools with larger proportions of students eligible for free or reduced lunch also tended to have fewer high value-added teachers.

When examining differences in the probabilities that certain types of students were more or less likely to be assigned highly effective teachers, I found evidence of differential rates of exposure. In general whites were more likely to be assigned high value-added teachers than blacks both across schools and within schools. The starkest difference in the exposure rates came in the significantly higher likelihood that black students from the top quartile of prior performance had high-value added teachers than black students from the lower quartile. The majority of this difference was happening within schools. Interestingly, there were no real differences in exposure to low value-added teachers between any of the groups.

Personnel Decisions

Given the rather uniform distribution their teacher evaluation scores, many districts will consider using value-added estimates as a criterion for high-stakes personnel decisions such as compensation, tenure or layoffs. Districts may have more confidence in value-added measures if their results are somewhat stable over time and if the measures are precise enough to distinguish between teachers. The results would seem encouraging to districts. Including more than one year of data for teacher radically improves the stability of the estimates from year to year. The teacher's effect in year t is correlated with the effect at t+1 at 0.94 when using multiple years of data to estimate the effect at year t. Prior performance on value-added is a strong predictor of future performance.

The results on the precision of the estimates were not as strong. A district must consider how confident it needs to be to designate a teacher as above or below average. A district that requires a 95% confidence interval is going to be able to pinpoint approximately 27% of its teachers as distinct from the mean. Of course, the district has to ask whether it needs to be this sure given its current information on its teachers. The results indicate that estimating the probability that a teacher is above the mean or in a certain quartile could allow the district to identify additional teachers.

The degree of imprecision in the value-added estimates will need to be viewed by districts in the context of the other measures it uses. In many cases, districts lack the comparable reliability statistics for other measures it uses such as classroom observations. Does the district have reason to believe that these measures are more or less precise than the value-added estimates? Can the imprecision of the value-added estimates be offset by including other measures or is the result of combining noisy measures just more noise?

Further Research

The results of this study suggest fruitful avenues of research that could be helpful for districts. First, districts could benefit from further help with handling missing data on prior student performance. One approach would be the creation of within-state longitudinal data systems so that students transferring into a district from within the state would have their test scores follow them. Another approach would be technical expertise on how to include imputation techniques in a district's value-added approach. The vexing problem for districts is the large number of students who are missing data on prior achievement and seem to score systematically lower than their peers once they are in the district.

A second line of research would aid districts in using value-added estimates to help teachers improve their instruction. Some districts may use value-added only as a sorting mechanism. Others will want to use them as diagnostically as possible on the assumption that moving the entire distribution of teachers will result in larger student achievement gains than lopping off the bottom tail of the distribution and replacing it. Value-added estimates in and of themselves do not provide teachers data on why their students are scoring higher or lower than expected with an average teacher. One way to improve the estimates might be to investigate whether or not teachers' value-added varies by student-type. For example, teachers could see that they are doing well with their previously high-performing students but not as well with their lower-performing students. One could imagine any number of student subgroups for which a district could calculate a teacher's value-added in hopes of providing deeper insight into the teacher's effectiveness.

Finally, districts could use more research such as the randomized assignment studies mentioned above to understand more fully to what extent the sorting of teachers to students could bias teacher effect estimates.

TABLES

TABLE 2: Review of Studies.

			Teacher Effects (SD)
Study	Data	Estimator(s)	Math
(Rothstein, 2010)	North Carolina, 1998-2001, Grade 3-5	Correlated Random Effects	0.15
(McCaffrey, et al., 2009)	Florida, 2001-2005, Grades 3-8	Fixed Effects with Shrinkage	0.21
(Hanushek & Rivkin, 2009)	Texas Unspecified District, 1996-2001, Grades 4-8	Fixed Effects with Shrinkage	0.17
(Kane & Staiger, 2008)	Los Angeles, 1997-2007, Grades 2-5	Variety	0.23
(Jacob, et al., 2008)	Anonymous, 1999-2005, Grades 3-6	Fixed Effects with Shrinkage	0.29
(Aaronson, et al., 2007)	Chicago, 1997-1999, Grade 9	Fixed Effects with Shrinkage	0.15
(Koedel & Betts, 2007)	San Diego, 1999-2002, Grades 3-5	IV estimator	0.26
(Harris & Sass, 2006)	Florida, 2000-2004, Grades 6-8	Variety	n/a
(Rivkin, et al., 2005)	Texas, 1993-1998, Grades 4-7	Fixed Effects	0.11
(Rockoff, 2004)	New Jersey county, 1990-2001, Grades K-6	Random Effects	0.10

Notes: Many of these estimates a variety of models in a comparison framework. Where possible, I chose models closest to those incorporated into this study.

TABLE 3: Exposure to Multiple Classes and Teachers in Math within Year.

Math Classes in 2010	# of Different Classes			
# of Different Teachers	1	2	3	Total
1	51.1	0.0	0.0	51.1
2	37.5	10.8	0.2	48.4
3	0.1	0.3	0.1	0.5
Total	88.6	11.2	0.3	100.0
# of students	42552			

Notes: CMS Grades 5-8 (2010). Figures are cell percentages.

TABLE 4: Sample Selection.

	Missing		Has Prior Score at				Final
	Score	Course	t-1=0	t-1=1			Sample
				t-2=0	t-2=1	t-2=0|1	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Math Z-score	.	0.922	-0.377	-0.241	-0.0158	-0.0359	-0.0288
	(.)	(0.787)	(1.002)	(0.967)	(0.967)	(0.969)	(0.967)
Math Z-Score at t-1	-1.561	1.011	.	-0.343	-0.0355	-0.0628	-0.0572
	(0.591)	(0.719)	(.)	(0.972)	(0.963)	(0.968)	(0.966)
Math Z-Score at t-2	-1.585	0.978	-0.750	.	-0.0481	-0.0481	-0.0435
	(0.573)	(0.723)	(1.019)	(.)	(0.960)	(0.960)	(0.959)
Male	0.551	0.508	0.533	0.508	0.505	0.505	0.504
	(0.497)	(0.500)	(0.499)	(0.500)	(0.500)	(0.500)	(0.500)
LEP	0.166	0.114	0.171	0.178	0.133	0.137	0.136
	(0.372)	(0.318)	(0.376)	(0.382)	(0.339)	(0.344)	(0.343)
Academically Gifted	0.0705	0.199	0.0519	0.0290	0.166	0.154	0.156
	(0.256)	(0.399)	(0.222)	(0.168)	(0.372)	(0.361)	(0.363)
Exceptional Child	0.290	0.0831	0.203	0.0968	0.0819	0.0832	0.0801
	(0.454)	(0.276)	(0.402)	(0.296)	(0.274)	(0.276)	(0.271)
Age on Jan 1 (Years) Centered by Grade	0.180	0.0104	0.131	0.0302	-0.0329	-0.0272	-0.0289
	(0.660)	(0.586)	(0.635)	(0.565)	(0.489)	(0.496)	(0.496)

TABLE 4 (continued)

	Missing		Has Prior Score at				Final
	Score	Course	t-1=0	t-1=1			Sample
				t-2=0	t-2=1	t-2=0|1	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Proportion of Days Absent at t-1	0.0562	0.0455	0.0568	0.0425	0.0358	0.0364	0.0363
	(0.0686)	(0.0566)	(0.0728)	(0.0428)	(0.0357)	(0.0364)	(0.0362)
Proportion of Days in Out-of-School Suspension at t-1	0.00984	0.00614	0.0101	0.00570	0.00456	0.00466	0.00462
	(0.0370)	(0.0288)	(0.0385)	(0.0249)	(0.0196)	(0.0202)	(0.0200)
Proportion of Days in In-School Suspension at t-1	0.00198	0.00143	0.00196	0.00132	0.00145	0.00144	0.00145
	(0.00859)	(0.00720)	(0.00895)	(0.00624)	(0.00701)	(0.00695)	(0.00697)
Repeating Grade	0.0700	0.0497	0.0644	0.0198	0.0161	0.0165	0.0166
	(0.255)	(0.217)	(0.245)	(0.139)	(0.126)	(0.127)	(0.128)
First Year in School	0.449	0.321	0.619	0.402	0.358	0.362	0.366
	(0.497)	(0.467)	(0.486)	(0.490)	(0.479)	(0.481)	(0.482)
Moves Between Schools w/i School Year	0.0698	0.0403	0.0779	0.104	0.0663	0.0696	0.0683
	(0.311)	(0.231)	(0.318)	(0.358)	(0.295)	(0.301)	(0.298)
Multi-racial	0.0382	0.0390	0.0423	0.0491	0.0354	0.0366	0.0365
	(0.192)	(0.194)	(0.201)	(0.216)	(0.185)	(0.188)	(0.188)
White	0.314	0.442	0.300	0.265	0.331	0.325	0.326
	(0.464)	(0.497)	(0.458)	(0.441)	(0.470)	(0.468)	(0.469)

TABLE 4 (continued)

	Missing		Has Prior Score at				Final
	Score	Course	t-1=0	t-1=1			Sample
				t-2=0	t-2=1	t-2=0|1	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Native American	0.00629	0.00535	0.00684	0.00624	0.00520	0.00529	0.00534
	(0.0791)	(0.0729)	(0.0824)	(0.0788)	(0.0719)	(0.0726)	(0.0729)
Hispanic	0.168	0.126	0.172	0.187	0.144	0.148	0.148
	(0.374)	(0.332)	(0.377)	(0.390)	(0.352)	(0.355)	(0.355)
Asian	0.0403	0.0508	0.0471	0.0450	0.0417	0.0420	0.0422
	(0.197)	(0.220)	(0.212)	(0.207)	(0.200)	(0.201)	(0.201)
African-American	0.433	0.337	0.432	0.447	0.442	0.443	0.442
	(0.495)	(0.473)	(0.495)	(0.497)	(0.497)	(0.497)	(0.497)
Classroom Mean: Math Z-Score at t-1	-0.377	.	-0.251	-0.168	-0.0175	-0.0309	-0.0273
	(0.409)	(.)	(0.540)	(0.616)	(0.667)	(0.664)	(0.667)
Classroom Mean: Male	0.611	.	0.543	0.515	0.505	0.506	0.505
	(0.194)	(.)	(0.164)	(0.144)	(0.142)	(0.142)	(0.141)
Classroom Mean: Limited English	0.174	.	0.176	0.162	0.141	0.143	0.143
Proficiency	(0.189)	(.)	(0.170)	(0.149)	(0.135)	(0.136)	(0.136)
Classroom Mean: Academically Gifted	0.0213	.	0.0694	0.102	0.150	0.145	0.147
	(0.0698)	(.)	(0.141)	(0.173)	(0.213)	(0.210)	(0.211)

TABLE 4 (continued)

	Missing		Has Prior Score at				Final
	Score	Course	t-1=0	t-1=1			Sample
				t-2=0	t-2=1	t-2=0|1	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Classroom Mean: Exceptional Child	0.610	.	0.260	0.106	0.0873	0.0889	0.0855
	(0.429)	(.)	(0.364)	(0.175)	(0.146)	(0.149)	(0.140)
Classroom Mean: Age on Jan 1	0.215	.	0.0802	0.0101	-0.0186	-0.0161	-0.0180
(Years) Centered by Grade	(0.289)	(.)	(0.239)	(0.177)	(0.167)	(0.168)	(0.166)
Classroom Mean: Proportion of Days	0.0465	.	0.0390	0.0354	0.0339	0.0340	0.0340
Absent at t-1	(0.0252)	(.)	(0.0184)	(0.0131)	(0.0116)	(0.0117)	(0.0114)
Classroom Mean: Proportion of Days	0.00993	.	0.00688	0.00524	0.00422	0.00431	0.00430
in Out-of-School Suspension at t-1	(0.0186)	(.)	(0.0143)	(0.0104)	(0.00868)	(0.00885)	(0.00869)
Classroom Mean: Proportion of Days	0.00209	.	0.00169	0.00155	0.00128	0.00130	0.00132
in In-School Suspension at t-1	(0.00401)	(.)	(0.00329)	(0.00292)	(0.00257)	(0.00260)	(0.00260)
Classroom Mean: Repeating Grade	0.0283	.	0.0230	0.0192	0.0157	0.0160	0.0162
	(0.0697)	(.)	(0.0534)	(0.0408)	(0.0374)	(0.0377)	(0.0376)
Classroom Mean: Moves Between	0.137	.	0.106	0.0826	0.0692	0.0704	0.0693
Schools w/i School Year	(0.229)	(.)	(0.172)	(0.118)	(0.108)	(0.109)	(0.106)
Class Size	17.92	.	23.83	24.19	24.65	24.61	24.67
	(15.40)	(.)	(14.55)	(5.728)	(5.658)	(5.666)	(5.538)

TABLE 4 (continued)

	Missing		Has Prior Score at				Final
	Score	Course	t-1=0	t-1=1			Sample
				t-2=0	t-2=1	t-2=0|1	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Classroom Mean: Multi-racial	0. 0260	.	0.0324	0.0367	0.0376	0.0375	0.0375
	(0.0520)	(.)	(0.0454)	(0.0443)	(0.0427)	(0.0428)	(0.0427)
Classroom Mean: White	0.186	.	0.253	0.294	0.324	0.322	0.323
	(0.220)	(.)	(0.262)	(0.278)	(0.289)	(0.288)	(0.288)
Classroom Mean: Native American	0.00578	.	0.00569	0.00569	0.00542	0.00544	0.00550
	(0.0256)	(.)	(0.0203)	(0.0169)	(0.0164)	(0.0164)	(0.0164)
Classroom Mean: Hispanic	0.174	.	0.177	0.168	0.150	0.152	0.152
	(0.177)	(.)	(0.158)	(0.142)	(0.135)	(0.135)	(0.134)
Classroom Mean: Asian	0.0318	.	0.0402	0.0431	0.0448	0.0447	0.0448
	(0.0749)	(.)	(0.0632)	(0.0553)	(0.0563)	(0.0562)	(0.0562)
Classroom Mean: African-American	0.576	.	0.491	0.453	0.438	0.439	0.438
	(0.255)	(.)	(0.256)	(0.253)	(0.263)	(0.263)	(0.262)
School Mean: Math Z-Score at t-1	-0.103	0.0427	-0.0890	-0.0479	0.0104	0.00517	0.00739
	(0.470)	(0.484)	(0.472)	(0.476)	(0.481)	(0.480)	(0.479)
School Mean: Reading Z-Score at t-1	-0.107	0.0337	-0.0958	-0.0531	0.00799	0.00255	0.00486
	(0.459)	(0.464)	(0.460)	(0.459)	(0.463)	(0.463)	(0.461)

TABLE 4 (continued)

	Missing		Has Prior Score at				Final
	Score	Course	t-1=0	t-1=1			Sample
				t-2=0	t-2=1	t-2=0|1	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
School Mean: Male	0.511	0.507	0.511	0.508	0.504	0.504	0.140
	(0.0493)	(0.0413)	(0.0456)	(0.0408)	(0.0405)	(0.0406)	(0.0978)
School Mean: LEP	0.152	0.139	0.154	0.149	0.140	0.140	0.0290
	(0.103)	(0.0941)	(0.104)	(0.101)	(0.0988)	(0.0990)	(0.0430)
School Mean: Academically Gifted	0.122	0.154	0.124	0.132	0.152	0.150	0.151
	(0.106)	(0.115)	(0.106)	(0.108)	(0.117)	(0.116)	(0.117)
School Mean: Exceptional Child	0.101	0.0869	0.0975	0.0904	0.0870	0.0873	0.0871
	(0.0955)	(0.0505)	(0.0799)	(0.0484)	(0.0325)	(0.0342)	(0.0314)
School Mean: Proportion of Days in In-School Suspension at t-1	0.00165	0.00149	0.00162	0.00147	0.00136	0.00137	0.00139
	(0.00201)	(0.00168)	(0.00195)	(0.00167)	(0.00160)	(0.00160)	(0.00159)
School Mean: Proportion of Days in Out-of-School Suspension at t-1	0.00624	0.00501	0.00600	0.00507	0.00441	0.00447	0.00450
	(0.00997)	(0.00756)	(0.00929)	(0.00717)	(0.00610)	(0.00621)	(0.00608)
School Mean: Proportion of Days Absent at t-1	0.0380	0.0370	0.0378	0.0372	0.0362	0.0363	0.0363
	(0.00891)	(0.00737)	(0.00831)	(0.00674)	(0.00652)	(0.00654)	(0.00646)
School Mean: First Year in School	0.403	0.406	0.408	0.405	0.388	0.389	0.391
	(0.160)	(0.139)	(0.160)	(0.159)	(0.161)	(0.161)	(0.159)

TABLE 4 (continued)

	Missing		Has Prior Score at				Final
	Score	Course	t-1=0	t-1=1			Sample
				t-2=0	t-2=1	t-2=0|1	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
School Mean: Moves Between Schools	0.229	0.222	0.228	0.213	0.184	0.187	0.179
w/i School Year	(0.218)	(0.200)	(0.212)	(0.185)	(0.174)	(0.175)	(0.154)
School Mean: Student in Different	0.352	0.352	0.354	0.356	0.348	0.348	0.187
School This Year	(0.195)	(0.185)	(0.194)	(0.197)	(0.196)	(0.196)	(0.175)
Proportion of Economically	0.555	0.481	0.548	0.524	0.503	0.505	0.504
Disadvantaged Students	(0.262)	(0.259)	(0.264)	(0.265)	(0.260)	(0.261)	(0.260)
School's Student Enrollment	779.5	877.5	796.2	819.0	801.3	802.9	813.3
	(366.9)	(352.4)	(365.5)	(367.0)	(361.8)	(362.3)	(359.3)
School Mean: Multi-racial	0.0368	0.0365	0.0368	0.0371	0.0379	0.0379	0.0377
	(0.0154)	(0.0141)	(0.0150)	(0.0142)	(0.0153)	(0.0152)	(0.0150)
School Mean: White	0.293	0.360	0.300	0.320	0.334	0.333	0.334
	(0.259)	(0.266)	(0.262)	(0.266)	(0.264)	(0.264)	(0.263)
School Mean: Native American	0.00533	0.00503	0.00531	0.00533	0.00513	0.00515	0.00518
	(0.00402)	(0.00360)	(0.00394)	(0.00396)	(0.00391)	(0.00391)	(0.00389)
School Mean: Hispanic	0.159	0.144	0.160	0.157	0.147	0.148	0.148
	(0.103)	(0.0952)	(0.104)	(0.103)	(0.103)	(0.103)	(0.102)

TABLE 4 (continued)

<table>
<tr><td></td><td colspan="2">Missing</td><td colspan="4">Has Prior Score at</td><td>Final</td></tr>
<tr><td></td><td>Score</td><td>Course</td><td>t-1=0</td><td colspan="3">t-1=1</td><td>Sample</td></tr>
<tr><td></td><td></td><td></td><td></td><td>t-2=0</td><td>t-2=1</td><td>t-2=0|1</td><td></td></tr>
<tr><td></td><td>(1)</td><td>(2)</td><td>(3)</td><td>(4)</td><td>(5)</td><td>(6)</td><td>(7)</td></tr>
<tr><td>School Mean: Asian</td><td>0.0424</td><td>0.0447</td><td>0.0435</td><td>0.0443</td><td>0.0454</td><td>0.0453</td><td>0.0453</td></tr>
<tr><td></td><td>(0.0259)</td><td>(0.0262)</td><td>(0.0268)</td><td>(0.0273)</td><td>(0.0276)</td><td>(0.0276)</td><td>(0.0273)</td></tr>
<tr><td>School Mean: African-American</td><td>0.464</td><td>0.410</td><td>0.454</td><td>0.436</td><td>0.430</td><td>0.431</td><td>0.430</td></tr>
<tr><td></td><td>(0.225)</td><td>(0.226)</td><td>(0.226)</td><td>(0.225)</td><td>(0.230)</td><td>(0.229)</td><td>(0.228)</td></tr>
<tr><td>2008</td><td>0.351</td><td>0.387</td><td>0.356</td><td>0.355</td><td>0.294</td><td>0.300</td><td>0.304</td></tr>
<tr><td></td><td>(0.477)</td><td>(0.487)</td><td>(0.479)</td><td>(0.479)</td><td>(0.456)</td><td>(0.458)</td><td>(0.460)</td></tr>
<tr><td>2009</td><td>0.344</td><td>0.310</td><td>0.330</td><td>0.349</td><td>0.347</td><td>0.348</td><td>0.356</td></tr>
<tr><td></td><td>(0.475)</td><td>(0.463)</td><td>(0.470)</td><td>(0.477)</td><td>(0.476)</td><td>(0.476)</td><td>(0.479)</td></tr>
<tr><td>2010</td><td>0.305</td><td>0.303</td><td>0.313</td><td>0.296</td><td>0.358</td><td>0.353</td><td>0.340</td></tr>
<tr><td></td><td>(0.461)</td><td>(0.460)</td><td>(0.464)</td><td>(0.457)</td><td>(0.479)</td><td>(0.478)</td><td>(0.474)</td></tr>
<tr><td>Grade 5</td><td>0.271</td><td>0.177</td><td>0.262</td><td>0.269</td><td>0.279</td><td>0.278</td><td>0.264</td></tr>
<tr><td></td><td>(0.444)</td><td>(0.381)</td><td>(0.440)</td><td>(0.443)</td><td>(0.449)</td><td>(0.448)</td><td>(0.441)</td></tr>
<tr><td>Grade 6</td><td>0.261</td><td>0.182</td><td>0.261</td><td>0.250</td><td>0.267</td><td>0.266</td><td>0.272</td></tr>
<tr><td></td><td>(0.439)</td><td>(0.386)</td><td>(0.439)</td><td>(0.433)</td><td>(0.443)</td><td>(0.442)</td><td>(0.445)</td></tr>
<tr><td>Grade 7</td><td>0.226</td><td>0.140</td><td>0.236</td><td>0.263</td><td>0.264</td><td>0.264</td><td>0.270</td></tr>
<tr><td></td><td>(0.418)</td><td>(0.347)</td><td>(0.425)</td><td>(0.440)</td><td>(0.441)</td><td>(0.441)</td><td>(0.444)</td></tr>
</table>

TABLE 4 (continued)

	Missing		Has Prior Score at				Final
	Score	Course	t-1=0	t-1=1			Sample
				t-2=0	t-2=1	t-2=0|1	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Grade 8	0.242	0.501	0.241	0.219	0.189	0.192	0.194
	(0.428)	(0.500)	(0.428)	(0.413)	(0.392)	(0.394)	(0.395)
Number of Student (x) Year Observations	18285	21893	28513	8171	83665	91836	89300
Number of Distinct Students	16103	21512	23790	8171	46198	50830	48552
Number of Distinct Classrooms	1708	0	4270	3627	4687	4710	4458
Number of Distinct Teachers	994	0	1408	1192	1348	1351	1144

Notes: CMS Grades 5-8 (2008-2010). All Z-Scores are normalized by grade and year to mean of zero and unit variation.

TABLE 5: Accounting for Student Heterogeneity.

	Lag (1)	Lag (2)	IV	Gain-score	Student Fixed Effects	First Difference
	(1)	(2)	(3)	(4)	(5)	(6)
Teacher Effect (SD)						
Unadjusted	0.20	0.19	0.21	0.19	0.34	0.24
Adjusted	0.17	0.16	0.17	0.15	0.21	0.23
Student-Level Error (SD)	0.47	0.43	0.47	0.51	0.39	0.41
Student Characteristics:						
Time-Varying	204.21	188.46	1340.63	175.8	47.92	40.22
Time-Invariant	442.79	173.69	2111.27	31.03	~	~
R^2	0.77	0.80	0.73	0.11	0.15	0.41
Number of Student (x) Year Observations	89300	81303	81303	89300	89300	83607
Number of Distinct Students	48552	44063	44063	48552	48552	45415
Number of Distinct Classrooms	4458	4443	4443	4458	4458	4455
Number of Distinct Teachers	1144	1140	1140	1144	1144	1144

Correlations of Teacher Effects Between Models	(1)	(2)	(3)	(4)	(5)	(6)
Lag (1)	1					
Lag (2)	0.96	1				
IV	0.99	0.96	1			
Gainscore	0.86	0.85	0.84	1		
Student Fixed Effects	0.64	0.66	0.64	0.75	1	
First Differences	0.69	0.70	0.68	0.77	0.62	1

Notes: CMS, Grades 5-8, (2008-2010). All specifications include student-level characteristics and grade-by-year fixed effects. Time-varying student characteristics include their prior year proportion of days absent of days enrolled in the prior year, proportion of days spent in out-of-school suspension in prior year, proportion of days spent in in-school suspension in prior year, whether the student is repeating the grade, enrolled in the school for the first time, and the number of moves between schools in the current year. Time-invariant characteristics include the student's gender, limited English proficiency, designation as academically gifted or special education student. In models that include one or more prior year math scores, the functional form of the prior scores is allowed to vary by up to a cubic. Teacher effects are estimated as fixed effects and are

TABLE 5 (continued)

reported in standard deviations of student test scores. These standard deviations are reported in two forms--unadjusted for sampling error and adjusted as empirical Bayes estimates. The figures for each set of control variables are the F-statistics from a Wald test of the joint significance of the control variable, except in column 3 where they are chi-square test statistics due to the IV estimator used. All tests reject the null of that the controls are jointly equal to zero at p<.001. Models summarized in columns 1-2 and 4 are estimated using the -felsdvreg_dm- command (Mihaly, et al., 2010) in Stata. Model in column 3 estimated using Stata's –xtivreg- command. Model in column 6 estimated using the –fese- command (Nichols, 2008) in Stata. The correlations of the models use the teacher estimates that have been unadjusted for sampling error.

TABLE 6: Accounting for Classroom and School Heterogeneity.

	Base	Class Cov	School Cov	School FE
	(1)	(2)	(3)	(4)
Teacher Effect (SD)				
Unadjusted	0.20	0.19	0.20	0.33
Adjusted	0.17	0.16	0.17	0.12
Student-Level Error (SD)	0.47	0.46	0.46	0.46
Student Controls	264.67	255.3	256.02	254.62
Classroom Controls		45.23	45.84	43.29
School Controls			19.34	
School Fixed Effects				3.27
R^2	0.77	0.77	0.77	0.77
Number of Student (x) Year Observations	89300			
Number of Distinct Students	48552			
Number of Distinct Classrooms	4458			
Number of Distinct Teachers	1144			
Correlations of Teacher Effects Between Models	(1)	(2)	(3)	(4)
Base (Lag1)	1			
Classroom Covariates	0.96	1		
School Covariates	0.89	0.94	1	
School Fixed Effects	0.54	0.60	0.58	1

Notes: CMS Grades 5-8 (2008-2010). All specifications include student-level characteristics (excluding ethnicity) described summarized in Table 4 and grade-by-year fixed effects. Classroom controls include the classroom means of all the student level characteristics as well as the class size. School controls include the school means of all the student level characteristics as well as the school size, the school's mean reading achievement from the prior year, and the school's proportion of students who qualify for free or reduced price lunch. All models include a student's prior year math score, the functional form of which is allowed to vary by up to a cubic. The prior year score is interacted with the student's grade. Teacher effects are estimated as fixed effects and are reported in standard deviations of student test scores. These standard deviations are reported in two forms--unadjusted for sampling error and adjusted as empirical Bayes estimates. The figures for each set of control variables are the F-statistics from a Wald test of the joint significance of the control variable, except in column 3 where they are chi-square test statistics due to the IV estimator used. All tests reject the null of that the

TABLE 6 (continued)

controls are jointly equal to zero at p<.001. Models summarized in columns 1-3 are estimated using the -felsdvreg_dm- command (Mihaly, et al., 2010) in Stata. The model in column 6 estimated using the –fese- command (Nichols, 2008) in Stata. The correlations of the models use the teacher estimates that have been unadjusted for sampling error.

TABLE 7: Form of Teacher Effects.

	Fixed		Random	
	Unadj	Adj		Class RE
	(1)	(2)	(3)	(4)
Teacher Effect (SD)	0.20	0.17	0.16	0.15
Classroom Effect (SD)	n/a	n/a	n/a	0.10
Student-Level Error(SD)	0.46	0.46	0.47	0.46
R^2	0.77	0.77	0.77	0.78
Chi-Squared Statistic from LR test				921.85***
Number of Student (x) Year Observations	89300			
Number of Distinct Students	48552			
Number of Distinct Classrooms	4458			
Number of Distinct Teachers	1144			
Correlations of Teacher Effects Between Models	(1)	(2)	(3)	(4)
Teacher Fixed Effects (Unadjusted)	1			
Teacher Fixed Effects (Shrunken)	0.99	1		
Teacher Random Effects	0.94	0.96	1	
Teacher + Class Random Effects	0.91	0.93	0.99	1

Notes: CMS Grades 5-8, (2008-2010). All specifications include student, classroom and school characteristics (excluding ethnicity) summarized in Table 4 and grade-by-year fixed effects. All models include a student's prior year math score, the functional form of which is allowed to vary by up to a cubic. The prior year score is interacted with the student's grade. Teacher effects in model summarized in columns 1-2 are estimated as fixed effects. The teacher effects in columns 3-4 are predictions of the teacher random effects using Stata's –xtmixed- command. The model in columns 1-2 is replicated from Table 6, column 3. ***p<.001

TABLE 8: Effect of Race Controls.

		Race Controls		
	Base	Student	Class	School
	(1)	(2)	(3)	(4)
Teacher Effect (SD) Adjusted	0.15	0.15	0.15	0.15
Classroom-Level Effect (SD)	0.10	0.10	0.10	0.10
Student-Level Error (SD)	0.46	0.46	0.46	0.46
Controls for Racial Composition				
Student		664.65***	672.41***	652.58***
Classroom			18.81	13.82
School				29.08***
Likelihood Ratio Test		603.69***	18.77	28.86***
Number of Student (x) Year Observations	89300			
Number of Distinct Students	48552			
Number of Distinct Classrooms	4458			
Number of Distinct Teachers	1144			
Correlations of Teacher Effects Between Models	(1)	(2)	(3)	(4)
Base	1			
Student	0.998	1		
Classroom	0.995	0.999	1	
School	0.988	0.992	0.994	1

Notes: CMS Grades 5-8 (2008-2010). Each column describes maximum likelihood estimates of the distribution of teacher effects under different covariate specifications. Column 1 replicates the preferred model summarized in Table 7, column 4. Race covariates are indicators for student's self-reported ethnicity. The figures in columns 2-4 for each type of covariate (student, classroom, school) are the chi-square statistics for the joint significance of those controls. The likelihood ratio test row reports the chi-square statistics for the likelihood ratio test of the model versus the model in the prior column. ***p<.001

TABLE 9: Testing Assumption of Prior Year Score.

	Base	Lag (1)	Lag (2)
	(1)	(2)	(3)
Teacher Effect (SD) Adjusted	0.14	0.14	0.14
Student-Level Error (SD)	0.45	0.45	0.42
Chi-Square Statistics from t-1			
Student Characteristics	137.92***	83.95***	112.60***
Classroom Characteristics	29.38***	339.9***	65.51***
School Characteristics	29.30***	74.78***	69.31***
Chi-Square Statistics from t-2			
Student Characteristics		47.22***	24.07**
Classroom Characteristics		40.39***	42.06***
School Characteristics		13.84*	35.21***
Likelihood Ratio Test		158.51***	4459.77***
Number of Student (x) Year Observations	34507	34813	34507
Number of Distinct Students	24946	25165	24946
Number of Distinct Classrooms	2407	2409	2407
Number of Distinct Teachers	626	627	626
Correlations of Teacher Effects Between Models	(1)	(2)	(3)
Base	1		
Lag (1)	0.99	1	
Lag (2)	0.96	0.97	1

Notes: CMS, Math, Grades 5-8, (2008-2010). All specifications include the student, classroom and school characteristics from the preferred model summarized in Table 7, column 4. Column 1 is the preferred model. Column 2 replicates the base and adds all lagged student, class and school inputs from the prior two years. Column 3 replicates column 2 and adds the student's prior score from t-2.
*p<.05, **p<.01, ***p<.001.

TABLE 10: Teacher Effects with Varying Right-Censored Student Distributions.

	Base	<99	<95	<90
	(1)	(2)	(3)	(4)
Teacher Effect (SD) Adjusted	0.15	0.15	0.15	0.15
Percentile Rank Change from Base Model				
Mean		0	0	0
Standard Deviation		2.00	2.51	3.88
Minimum		-11	-11	-45
Maximum		17	16	20
Number of Student (x) Year Observations	89300	87179	84583	80909
Number of Distinct Students	48552	48033	47076	45387
Number of Distinct Classrooms	4458	4457	4457	4457
Number of Distinct Teachers	1144	1144	1144	1144
Correlations of Teacher Effects Between Models	(1)	(2)	(3)	(4)
Base	1			
<99	0.99	1		
<95	1.00	1.00	1	
<90	1.00	1.00	1.00	1

Notes: CMS Grades 5-8 (2008-2010). Column 1 replicates the preferred model summarized in Table 7, column 4. Columns 2-4 estimate the same model but with samples trimmed at the top tail of the student test score distribution at t-1. The percentile rank changes refer to the differences in teacher percentile ranks under the different specifications.

TABLE 11: Ceiling Effects.

	Years Contributing to Teacher Effect					
	2008		2008-09		2008-10	
	(1)	(2)	(3)	(4)	(5)	(6)
Teacher Effect (SD)	0.15	0.21	0.15	0.20	0.15	0.20
Classroom Effect (SD)	0.07	0.11	0.09	0.18	0.10	0.17
Student-Level Error (SD)	0.45	0.56	0.45	0.58	0.45	0.56
Student Current Score Standardized by:						
All Students t-1 Scores	x		x		x	
Decile of t-1 Scores		x		x		x
Number of Student (x) Year Observations	19961	19957	38936	38931	56687	56682
Number of Distinct Students	19961	19957	30601	30605	39037	39041
Number of Distinct Classrooms	940	940	1868	1868	2766	2766
Number of Distinct Teachers	484	484	484	484	484	484

Correlations of Teacher Effects Between Models

All Teachers	(1)	(2)	(3)	(4)	(5)	(6)
(1)	1					
(2)	0.85	1				
(3)	0.83	0.70	1			
(4)	0.70	0.77	0.85	1		
(5)	0.76	0.64	0.93	0.80	1	
(6)	0.65	0.71	0.80	0.92	0.87	1

Teachers in Top Decile of Prior Student	(1)	(2)	(3)	(4)	(5)	(6)
(1)	1					
(2)	0.76	1				
(3)	0.81	0.60	1			
(4)	0.70	0.84	0.77	1		
(5)	0.69	0.58	0.90	0.79	1	
(6)	0.62	0.79	0.67	0.93	0.83	1

TABLE 11 (continued)

Teachers in Bottom Decile of Prior Student	(1)	(2)	(3)	(4)	(5)	(6)
(1)	1					
(2)	0.88	1				
(3)	0.83	0.72	1			
(4)	0.69	0.87	0.74	1		
(5)	0.69	0.65	0.89	0.74	1	
(6)	0.60	0.74	0.68	0.90	0.81	1

Notes: CMS Grades 5-8 (2008-2010). All models use the preferred model's estimator summarized in Table 7, column 4. Columns 1-2 restrict the sample to one year (2008); columns 3-4 restrict sample to two years (2008-2009); and columns 5-6 use the full three-year sample (2008-2010). Columns 2-4-6 normalize the student test scores at t using the mean and standard deviation of the decile of the students' scores at t-1. In the correlation tables, "Teachers in Top Decile of Prior Student Achievement" refers to teachers whose mean student test scores at t-1 was in the top decile of all teachers in the given year.

TABLE 12: Teacher Effects under Sorting.

	Sorting		
	Actual	Perfect	Random
	(1)	(2)	(3)
Teacher Effect (SD) Adjusted	0.147	0.158	0.161
	(0.00442)	(0.00425)	(0.00423)
Classroom Effect (SD)	0.101	0.0952	0.0759
	(0.00260)	(0.00297)	(0.00360)
Student Level Error (SD)	0.456	0.456	0.459
	(0.00111)	(0.00118)	(0.00120)
SD of Student Scores (t-1) by Class	0.61	0.16	0.82
Number of Student (x) Year Observations	89300		
Number of Distinct Students	48552		
Number of Distinct Classrooms	4458		
Number of Distinct Teachers	1144		

Notes: CMS Grades 5-8 (2008-2010). All models use the preferred model's estimator summarized in Table 7, column 4. Column 1 replicates the preferred model which is estimated under the actual degree of sorting of students to teachers. Column 2 simulates perfect sorting of students to teachers within year, school, and grade by prior test score. Column 3 simulates random assignment of student to teachers within year, school and grade. Standard errors in parentheses.

TABLE 13: Testing for Effects of Next Year's Teacher on Current Year Gains.

	Full		Sorted		Not Sorted	
	t	t+1	t	t+1	t	t+1
	(1)	(2)	(3)	(4)	(5)	(6)
Teacher Effect (SD)	0.15	0.12	0.15	0.12	0.13	0.08
Classroom Effect (SD)	0.09	0.24	0.08	0.22	0.10	0.25
Student Level Error (SD)	0.46	0.44	0.45	0.44	0.46	0.44
Math Score at t-1	0.532***	0.461***	0.554***	0.498***	0.046	-0.028
	(0.05)	(0.05)	(0.05)	(0.05)	(0.23)	(0.25)
Math Score at t-1 Squared	0.005	-0.010***	0.002	-0.004	0.006	-0.022***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)	(0.005)
Math Score at t-1 Cubed	-0.0389***	-0.033***	-0.039***	-0.032***	-0.038***	-0.033***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)
Number of Student (x) Year Observations	40707	40707	32042	32042	8665	8665
Number of Distinct Students	28845	28845	24731	24731	8641	8641
Number of Distinct Classrooms	2538	2365	1998	2274	540	940
Number of Distinct Teachers	1011	592	724	546	391	304

Notes: CMS Grades 5-7 (2008-2009). All models use the preferred model's estimator summarized in Table 7, column 4. Columns 1-2 use the subsample of students from the sample used throughout the study that have classroom assignments at t+1. Columns 3-4 use a subsample of the group in columns 1-2 who are in classrooms at t that are sorted on student test scores at t-1. Columns 5-6 are a subsample of the group in columns 1-2 who are in classrooms at time t that are not sorted on the scores at t-1. Columns 1, 3, and 5 report the effects of teachers and classrooms at time t on student scores at time t. Columns 2, 4, and 6 report the effects of teachers and classrooms at time t+1 on student scores at time t. Standard errors in parentheses. ***p<.001

TABLE 14: Heterogeneous Impacts of Effective Teachers on Student Sub-Types.

	Quartile of Scores at t-1				Ethnicity			
	Q1 (low)	Q2	Q3	Q4	White	Black	Hispanic	Asian
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
SD of Student Scale Scores at t	6.4	6.0	5.6	5.8	8.3	8.4	8.6	9.5
Mean Scale Score Gain from t-1 to t	6.8	4.1	3.0	2.5	5.1	5.8	5.9	5.3
SD Teacher Effect in Student Level SDs	0.13	0.16	0.16	0.13	0.14	0.15	0.14	0.12
SD Teacher Effect in Student Level Scale Scores	0.84	0.96	0.88	0.75	1.20	1.23	1.24	1.15
Proportion of Mean Gain Associated with 1 SD Change in Teacher Quality	0.12	0.23	0.29	0.30	0.23	0.21	0.21	0.22
Number of Student (x) Year Observations	22468	22343	22313	22176	29121	39445	13226	3769
Number of Distinct Students	15004	16969	16141	14539	15874	21426	7188	2076
Number of Distinct Classrooms	3512	3866	3658	2723	3511	4311	3668	2139
Number of Distinct Teachers	1095	1115	1082	990	1013	1138	1045	862

Notes: CMS Math Grades 5-8, (2008-2010). All models use the preferred model’s estimator summarized in Table 7, column 4. Each column represents an estimation of teacher effects that has been restricted to the sample described in the column header. The proportion of the students' mean gain in test scores that is associated with a one standard deviation change in teacher quality is calculated by dividing the standard deviation of the teacher effect in scale score points by the mean scale score gain from t-1 to t.

TABLE 15: Difference in Probabilities of Being Assigned Effective Teacher.

	p(Top 25%=1)		p(Bottom 25%=1)		p(Missing VAM=1)	
	Across & Within	Within	Across & Within	Within	Across & Within	Within
	(1)	(2)	(3)	(4)	(3)	(4)
White - Black (All Quartiles)	0.050 #	0.012 *	0.003	-0.008	-0.001	0.001
	0.028	0.005	0.025	0.005	(0.001)	0.005
White - Hispanic (All Quartiles)	0.043 #	0.010	0.022	-0.003	0.005	0.005
	0.026	0.006	0.024	0.006	0.012	0.007
Black - Hispanic (All Quartiles)	-0.008	-0.001	0.019	0.005	0.007	0.004
	0.014	0.006	0.012	0.006	0.009	0.006
White Top Q - Black Top Q	0.002	-0.027 *	0.010	0.002	0.025	0.023 *
	0.03	0.01	0.03	0.01	0.02	0.01
White Top Q - Hispanic Top Q	0.043	0.023 #	0.028	0.001	0.018	0.015
	0.030	0.013	0.027	0.013	0.018	0.014
White Top Q - White Bottom Q	0.028	0.013	0.031	0.011	0.001	0.006
	0.030	0.013	0.026	0.013	0.021	0.014
Black Top Q - Black Bottom Q	0.068 **	0.064 ***	0.002	-0.013	-0.032 #	-0.021 #
	0.024	0.011	0.025	0.010	0.017	0.011

TABLE 15 (continued)

	p(Top 25%=1)		p(Bottom 25%=1)		p(Missing VAM=1)	
	Across & Within	Within	Across & Within	Within	Across & Within	Within
	(1)	(2)	(3)	(4)	(3)	(4)
Hispanic Top Q -Hispanic Bottom Q	0.007	-0.008	0.012	0.012	-0.015	-0.008
	0.030	0.015	0.023	0.015	0.020	0.016
Number of Student Observations	58521	52740	58521	48961	62109	61539
Number of Distinct Classrooms	2858	2566	2858	2332	3061	3028
Number of Distinct Teachers	881	735	881	650	968	952
Number of Distinct Schools	139	99	139	85	139	129

Notes: CMS Math Grades 5-8, (2009-2010). Each column summarizes a probit estimation of the probability that students of different ethnicities and quartiles of achievement from t-1 were assigned to teachers whose value-added at t-1 was in the top quartile of the district (columns 1-2), the bottom (columns 3-4), or whose value-added from the prior year was missing (columns 5-6). Columns 1-3-5 present the probability across and within schools, while columns 2-4-6 use school fixed effects to estimate the within school probability. Robust standard errors are in parentheses.
#p<.1 ,*p<.05, **p<.01, ***p<.001.

TABLE 16: Decomposition of Teacher Effects.

Teacher Variation (Proportion)			
Between	Within	Reliability	Stability
(1)	(2)	(3)	(4)
.63	.37	.81	.51

Notes: CMS Math Grades 5-8, (2008-2010). Based upon estimation of an alternative specification of the base model. Under this random effects specification, teacher-by-year effects, rather than classroom random effects, are nested within teacher effects. Columns 1-2 decompose the variation in teacher effects (net of estimation error) due to between and within teacher differences. Column 3 indicates the proportion of the variation in teacher effects due to the between and within teacher variation. Column 4 indicates the proportion of the between teacher variation of the total teacher variation.

TABLE 17: Teacher Effect Transition Matrices.
Panel A: 2009 v. 2010

		Quartile in 2010			
		(1)	(2)	(3)	(4)
Quartile in 2009	(1)	41.8	28.8	20.5	8.9
	(2)	28.9	26.3	21.7	23.0
	(3)	20.8	22.7	31.2	25.3
	(4)	5.5	14.5	28.5	51.5
					N=617

Correlation of Teacher Effect Point Estimates= 0.48

Panel B: <=2009 v. <=2010

		Quartile in 2010			
		(1)	(2)	(3)	(4)
Quartile in 2009	(1)	83.7	14.0	1.9	0.4
	(2)	13.3	68.9	14.8	3.0
	(3)	1.1	16.7	67.4	14.8
	(4)	0.4	0.8	17.1	81.7
					N=1055

Correlation of Teacher Effect Point Estimates= 0.94

Panel C: <=2009 vs. <=2010 for all teachers in sample all three years

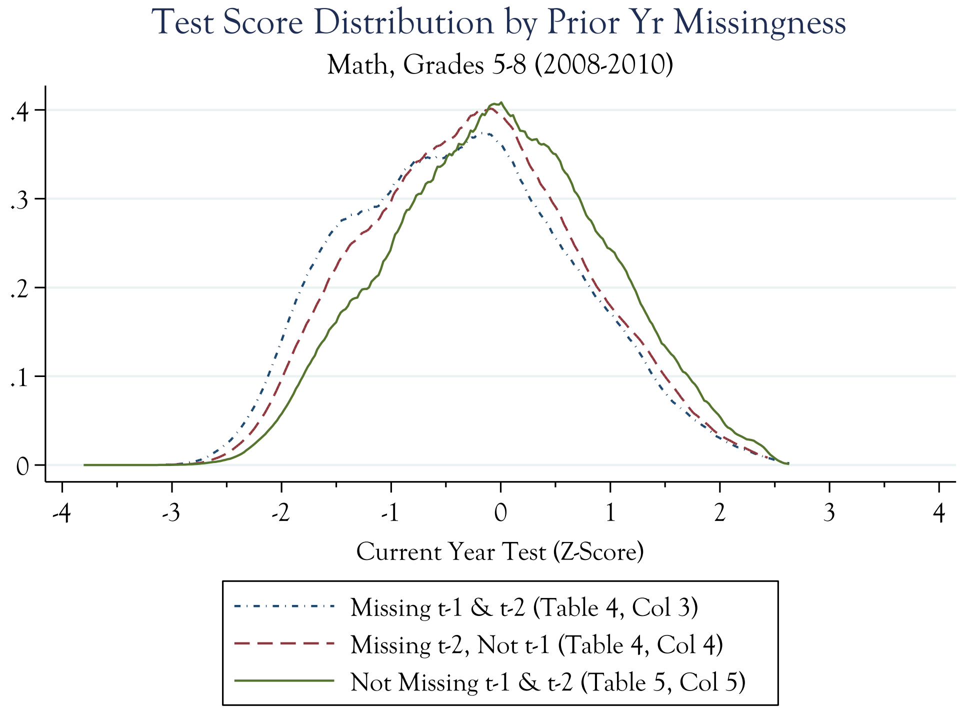
		Quartile in 2010			
		(1)	(2)	(3)	(4)
Quartile in 2009	(1)	78.7	19.7	1.6	0.0
	(2)	19.0	62.7	17.5	0.8
	(3)	2.4	15.7	63.0	18.9
	(4)	0.0	0.8	18.3	81.0
					N=506

Correlation of Teacher Effect Point Estimates= 0.94

Notes: CMS Grades 5-8, 2008-2010. Each panel provides a transition matrix for the teachers in a given quartile in 2009 who are in a given quartile in 2010. All figures are row percentages. Panel A uses quartiles based upon single year estimates of teacher effects in 2009 and 2010. Panel B uses multi-year estimates of teacher effects up to and including 2009 and then up to and including 2010. Panel C replicates Panel B but restricts the sample to those teachers who were present each year 2008-2010. Teacher effects are calculated using the base model described in Table 7, column 4.

FIGURES

FIGURE 1: Kernel Density of Current Test Score by Prior Year Score Missingness.



Notes: CMS Grades 5-8 (2008-2010). Current year test scores normalized by grade and year to have mean of zero and unit variation.

FIGURE 2: Functional Form of Prior Score.

Panel A: Score at t and t-1

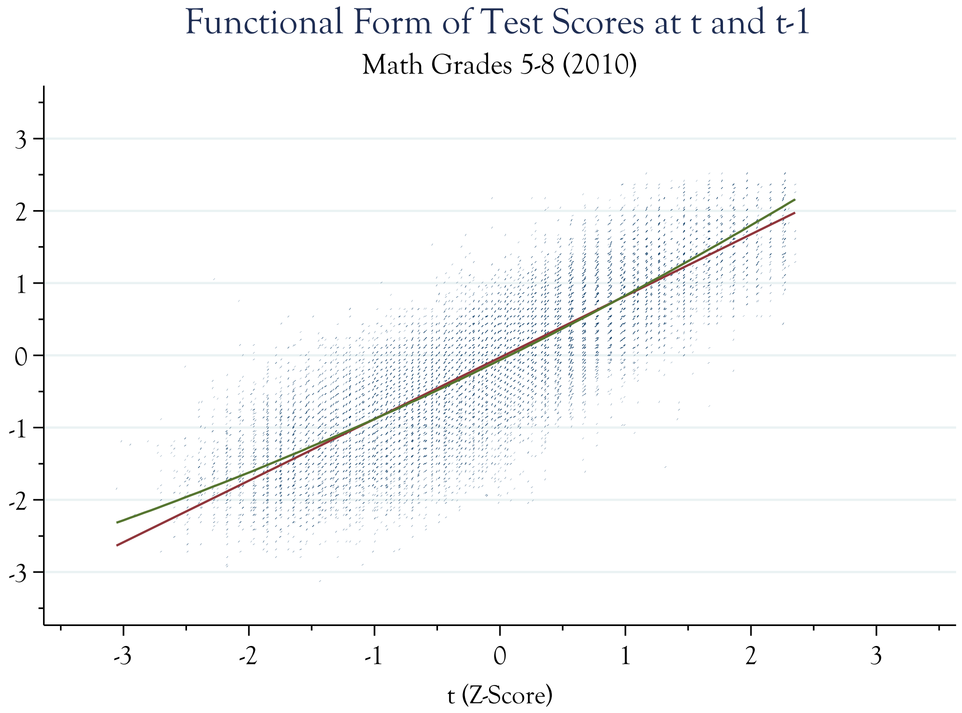
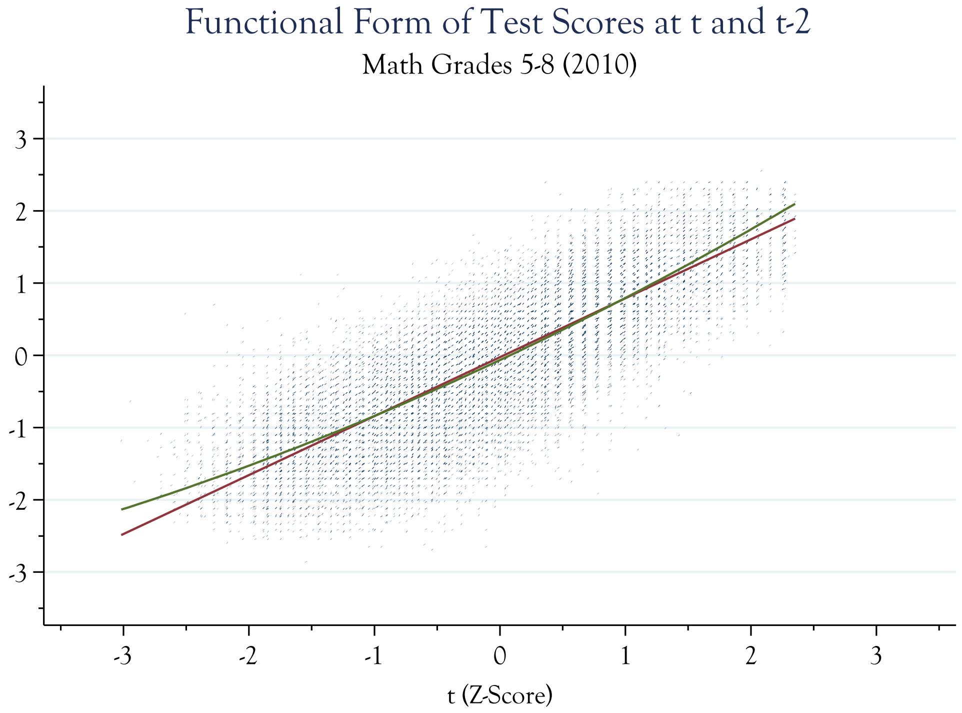


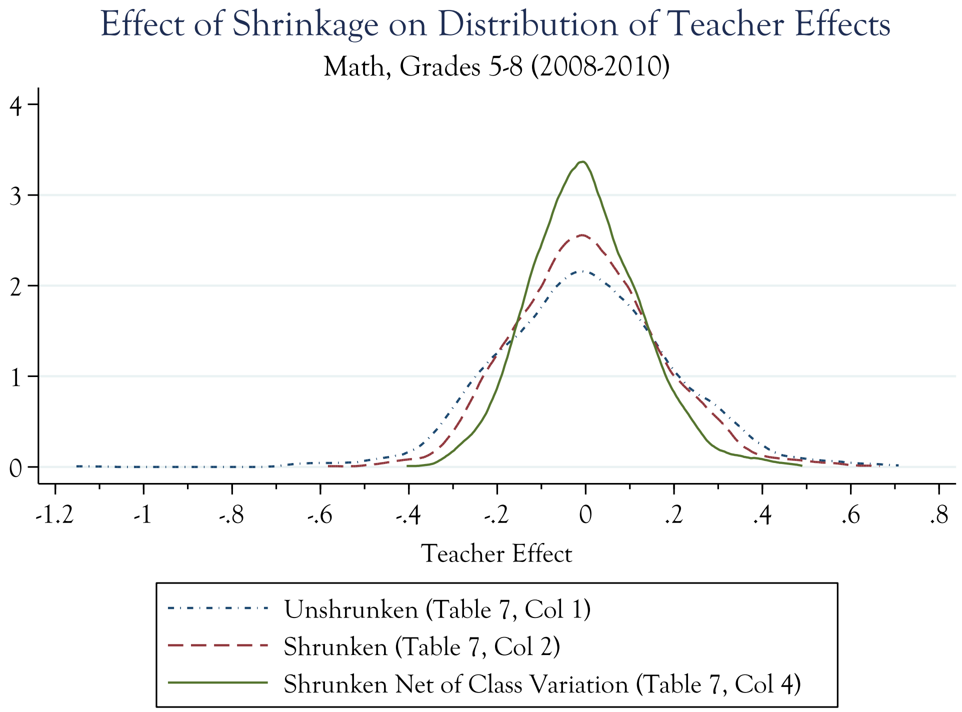
FIGURE 2 (continued)

Panel B: Score at t and t-2



Notes: CMS Grades 5-8 (2010). Test scores normalized by grade and year to have mean of zero and unit variation.

FIGURE 3: Teacher Effects and Shrinkage.



Notes: CMS Grades 5-8 (2008-2010). The distributions of the teacher effects come from the models estimated in Table 7.

FIGURE 4: Sampling Error and Teacher Effects.

Panel A: Unshrunken Estimates

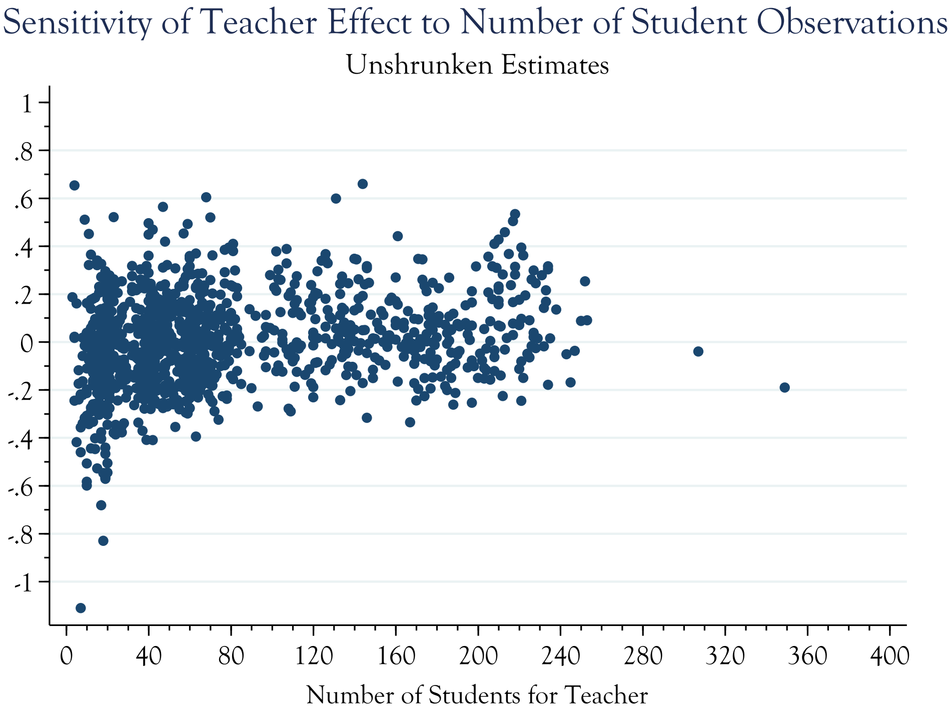
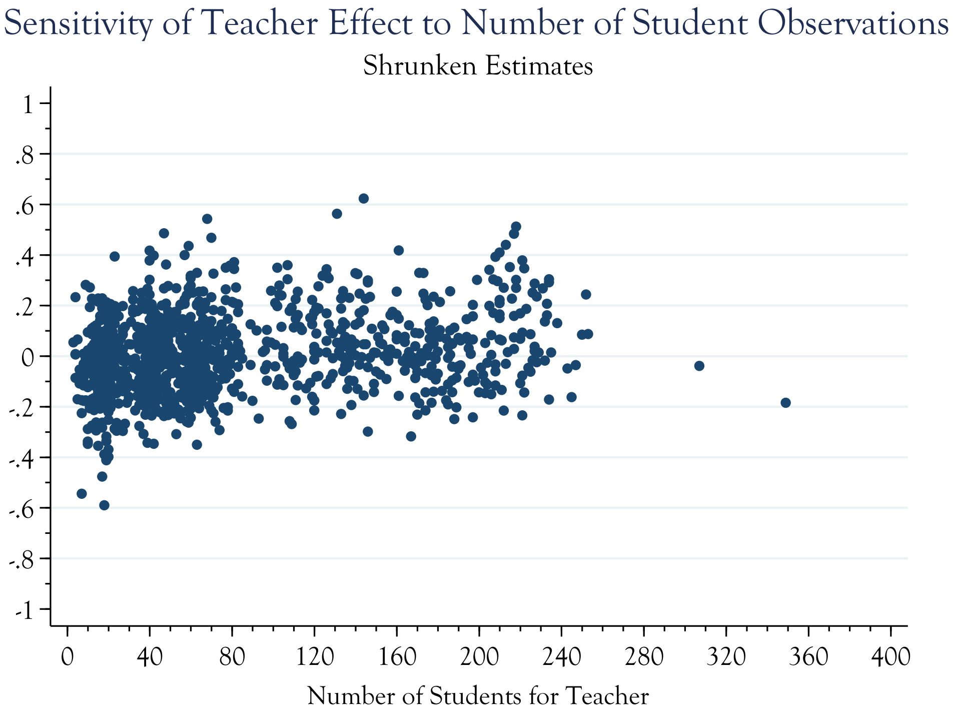


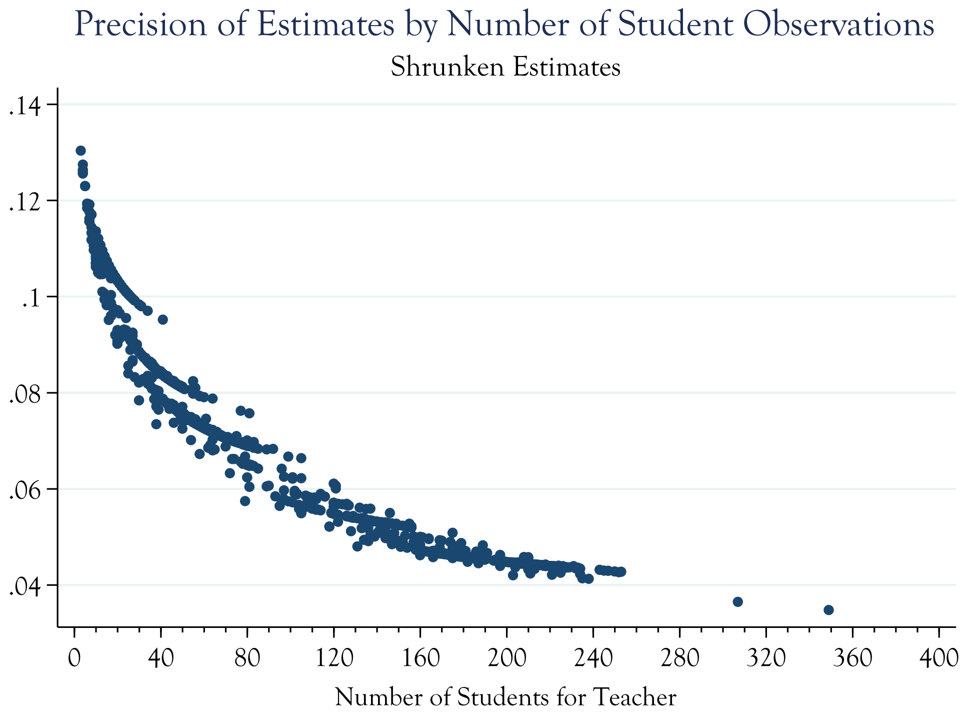
FIGURE 4 (continued)

Panel B: Shrunken Estimates



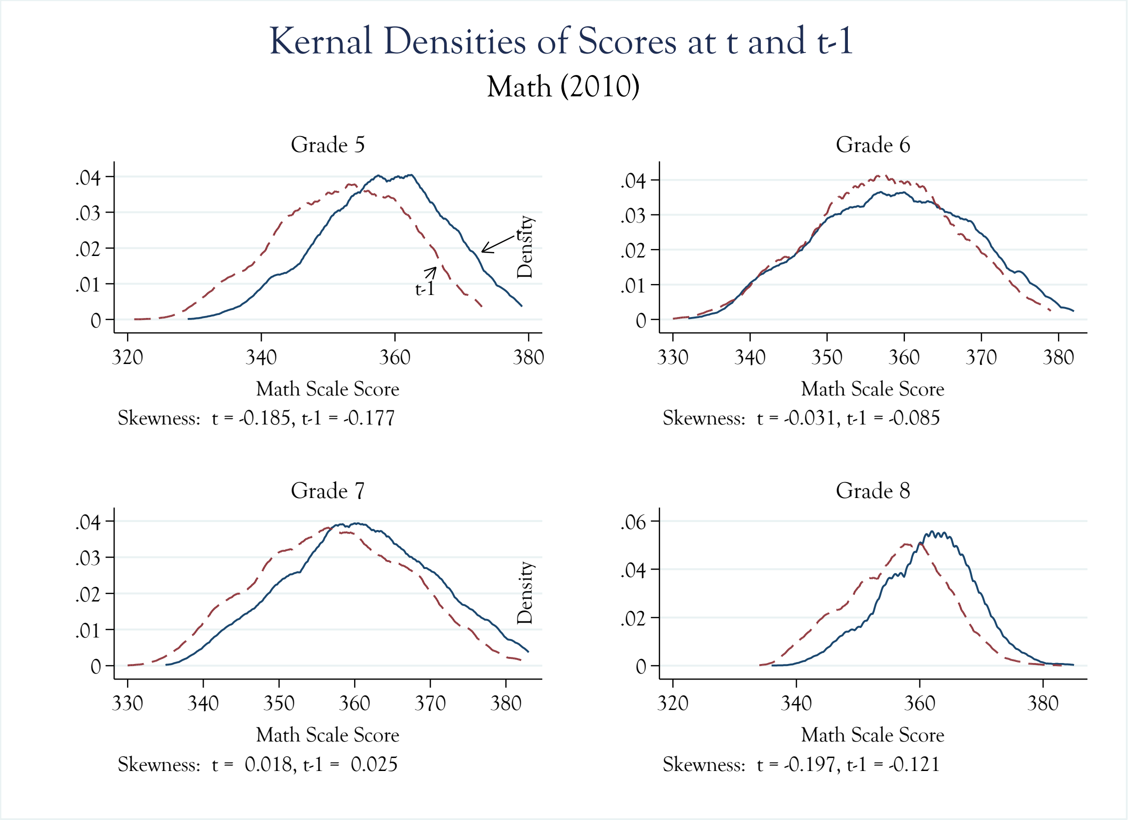
Notes: CMS Grades 5-8 (2008-2010). The effects plotted in Panels A-B come from the model summarized in Table 7 columns 1-2, respectively.

FIGURE 5: Precision of Estimates by Number of Observations.



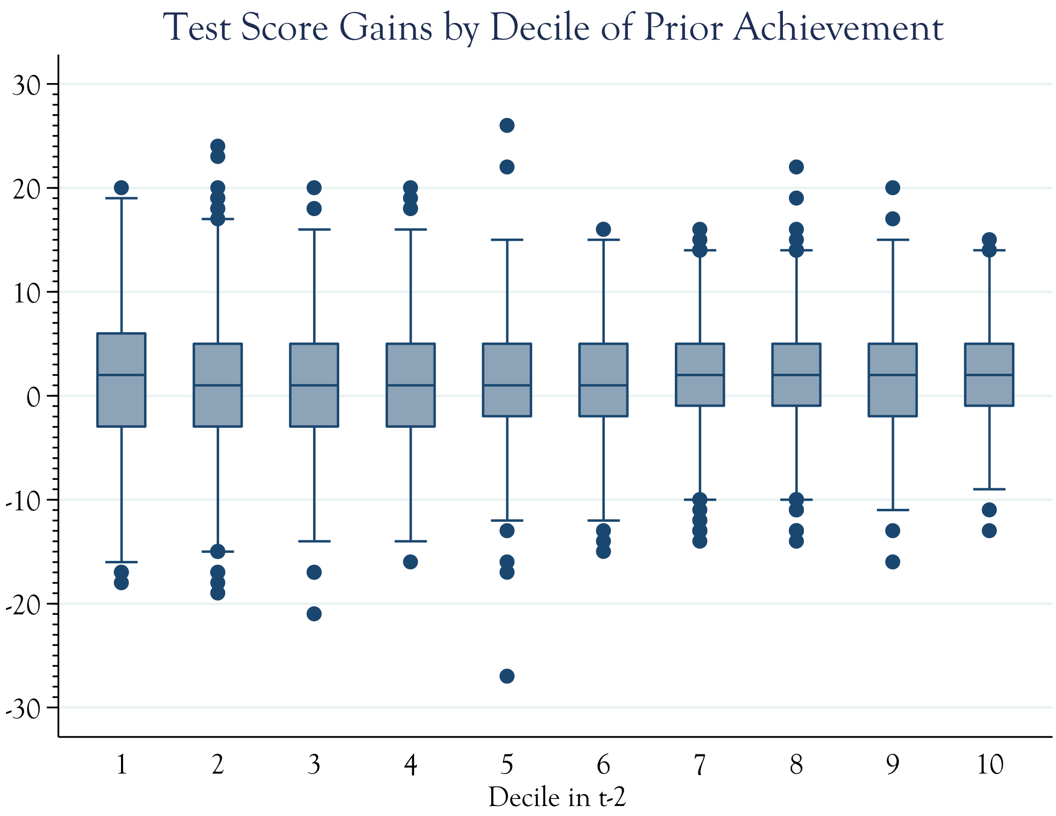
Notes: CMS Grades 5-8 (2008-2010). The standard errors are from the teacher effects generated by the model summarized in Table 7, column 4.

FIGURE 6: Skewness of Current and Prior Year Scores.



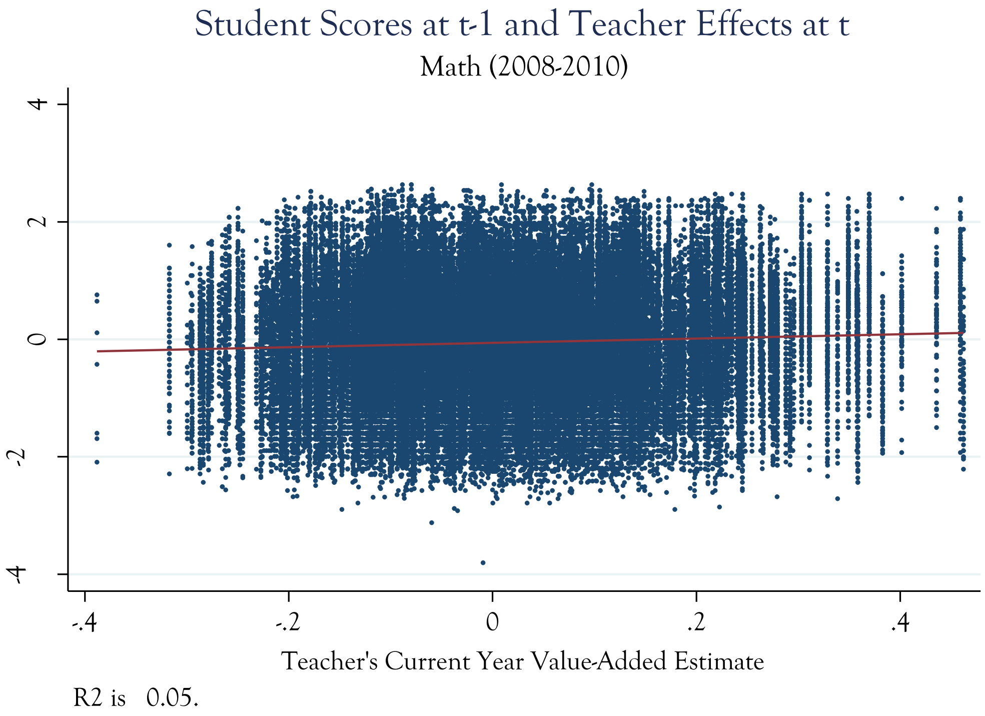
Notes: CMS Grades 5-8 (2010).

FIGURE 7: Average Test Score Gains by Prior Test Scores.



Notes: CMS Grades 5-8 (2008-2010). The elements of the figure are boxplots of the range of test scores changes from t-1 to t by decile of student test scores at t-2. (Decile 10 is top.) The shaded portion of the box represents the interquartile range. The whiskers represent the range of adjacent values, and the dots represent students outside the range of adjacent values.

FIGURE 8: Relationship of Prior Score to Current Year Teacher Effect.



Notes: CMS Grades 5-8 (2008-2010). Teacher effects are generated by the preferred model summarized in Table 7, column 4.

FIGURE 9: Teacher Effects and the Proportion of Students Top and Bottom Deciles.

Panel A: Top 10% of Students

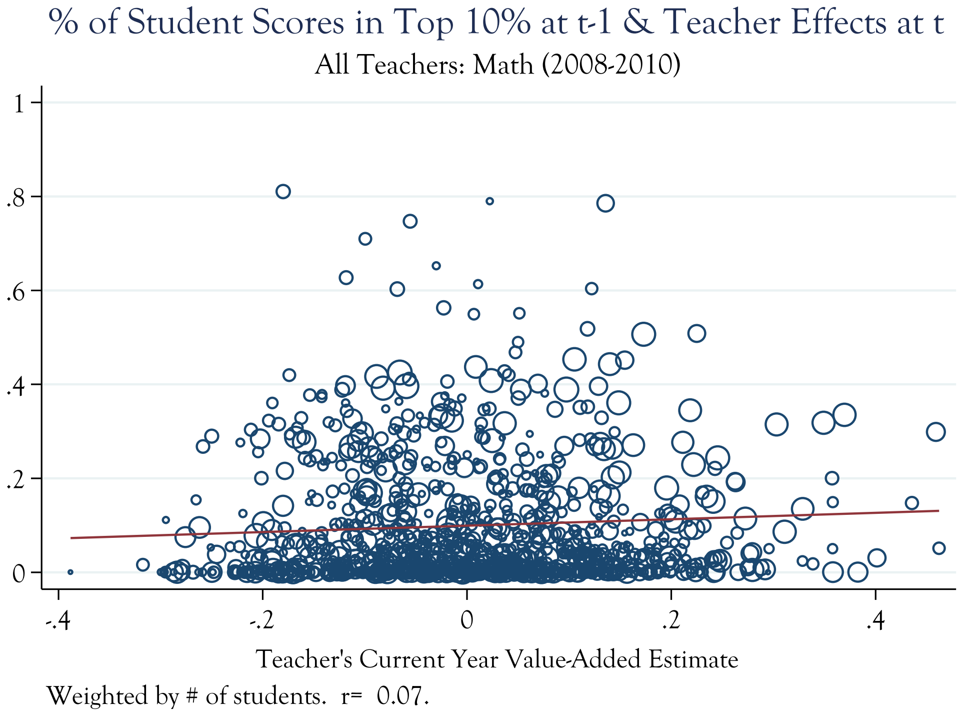
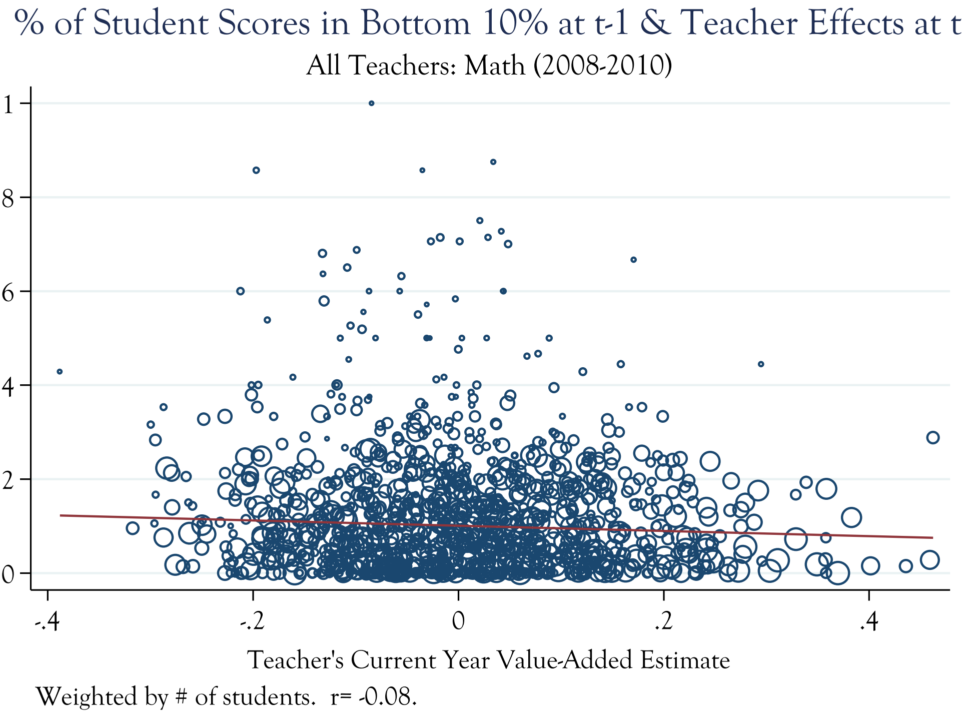


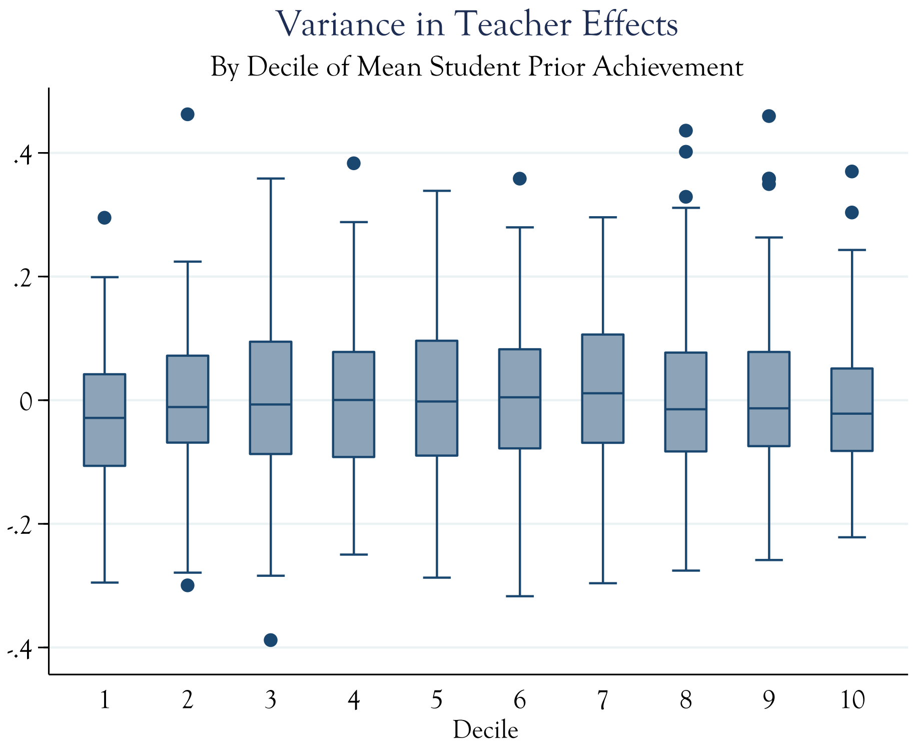
FIGURE 9 (continued)

Panel B: Bottom 10% Students



Notes: CMS Grades 5-8 (2008-2010). Teacher effects are based on preferred model summarized in Table 7, column 4.

FIGURE 10: Variation of Teacher Effects Across Student Deciles.



Notes: CMS Grades 5-8 (2008-2010). Teacher effects are based on preferred model summarized in Table 7, column 4. The elements of the figure are boxplots of the range of teacher effects by decile of teacher mean student test scores at t-2. (Decile 10 is top.) The shaded portion of the box represents the interquartile range. The whiskers represent the range of adjacent values, and the dots represent students outside the range of adjacent values.

FIGURE 11: Classroom Sorting.

Panel A. Standard Deviation of Student Scores at t-1 by Class

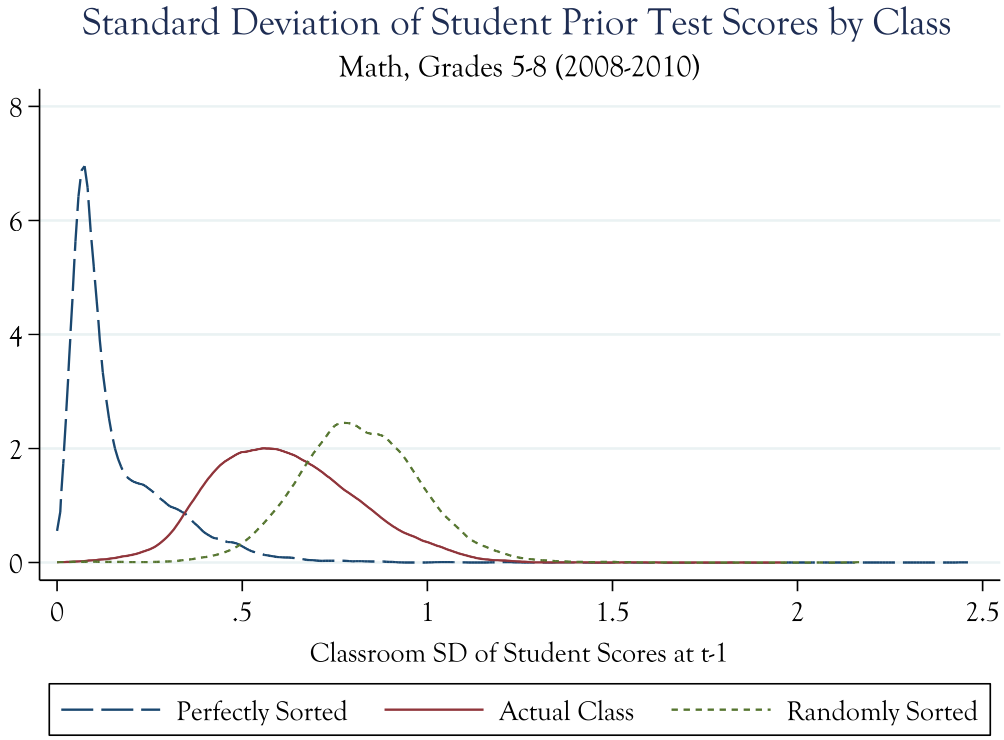
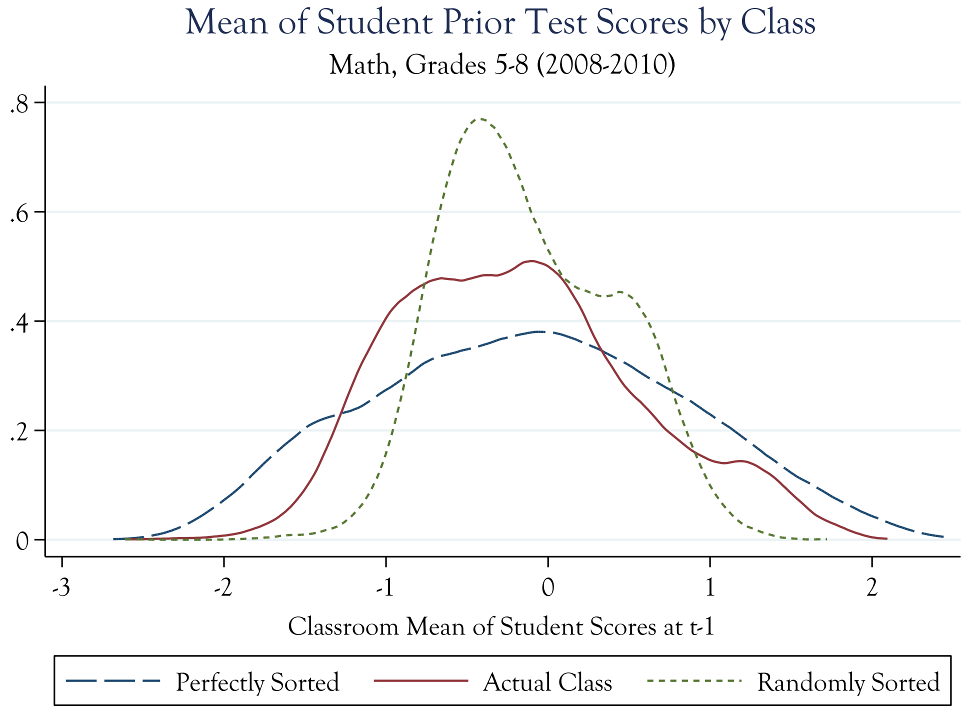


FIGURE 11 (continued)

Panel B. Classroom Means of Student Test Scores at t-1



Notes: CMS Grades 5-8 (2008-2010). Panels A-B depict the kernel densities of classroom prior test score standard deviation and means under three conditions of sorting. The perfectly sorted sample is a simulation in which students are sorted within year, school, and grade by prior test score. The randomly sorted sample simulates random assignment of teachers to students within these same strata. The actual sample reflects the extant sorting in the data.

FIGURE 12: Sorting by Subsample.

Panel A: Teacher's SD Student Test Scores at t-1

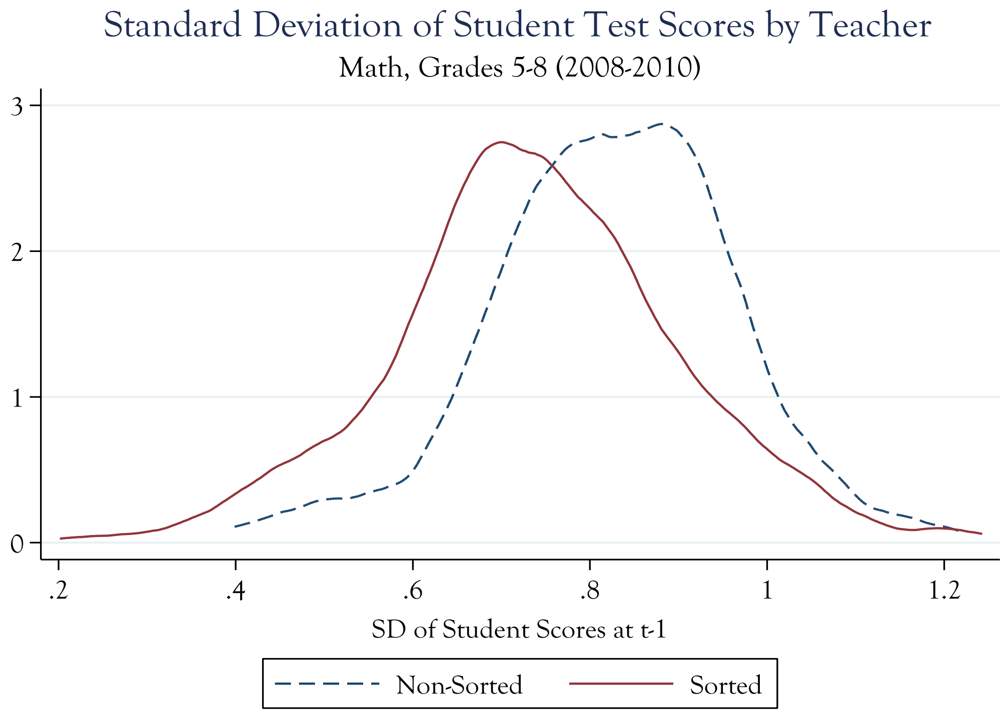
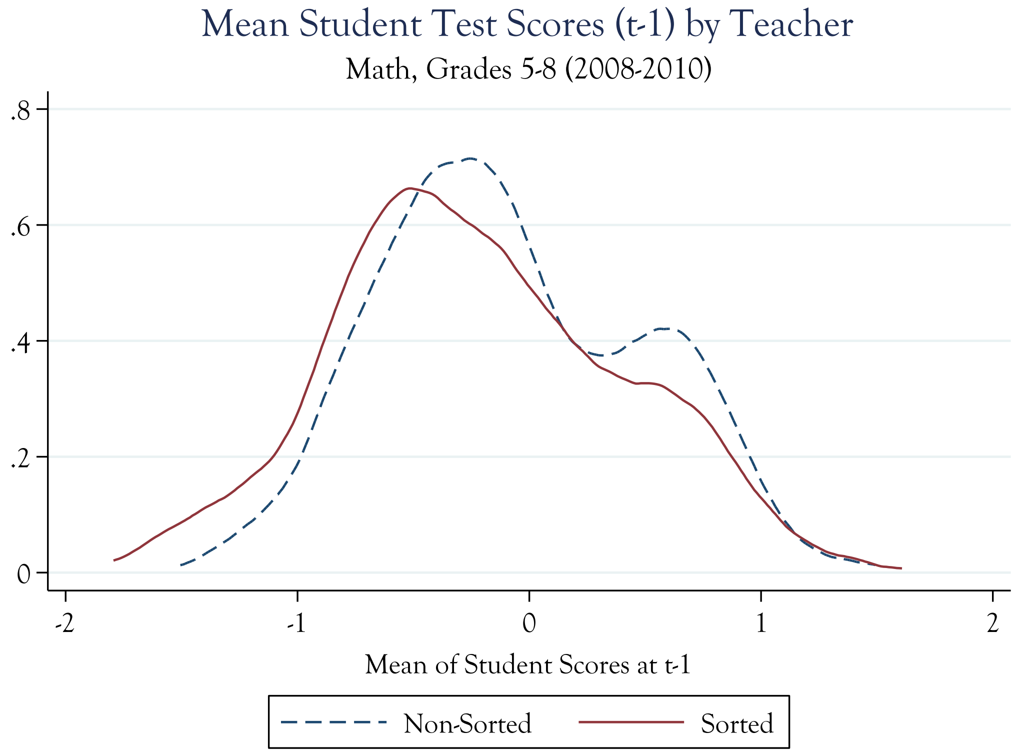


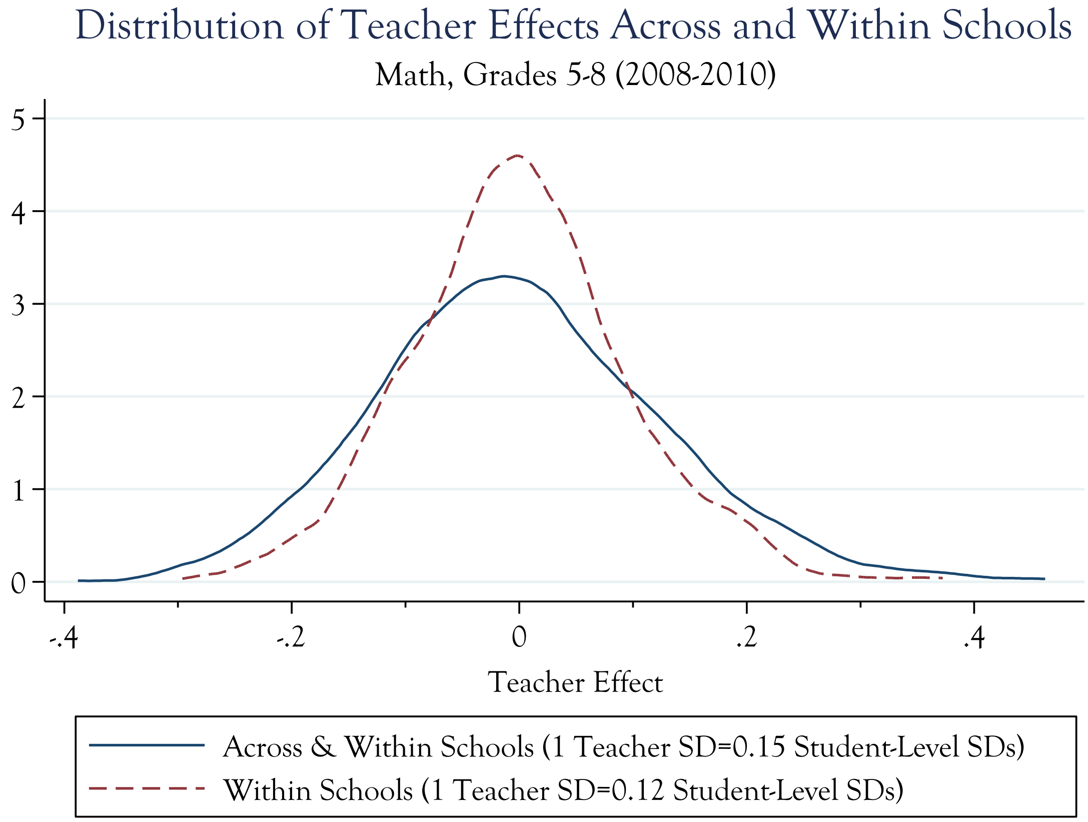
FIGURE 12 (continued)

Panel B: Teacher's Mean Student Test Scores at t-1



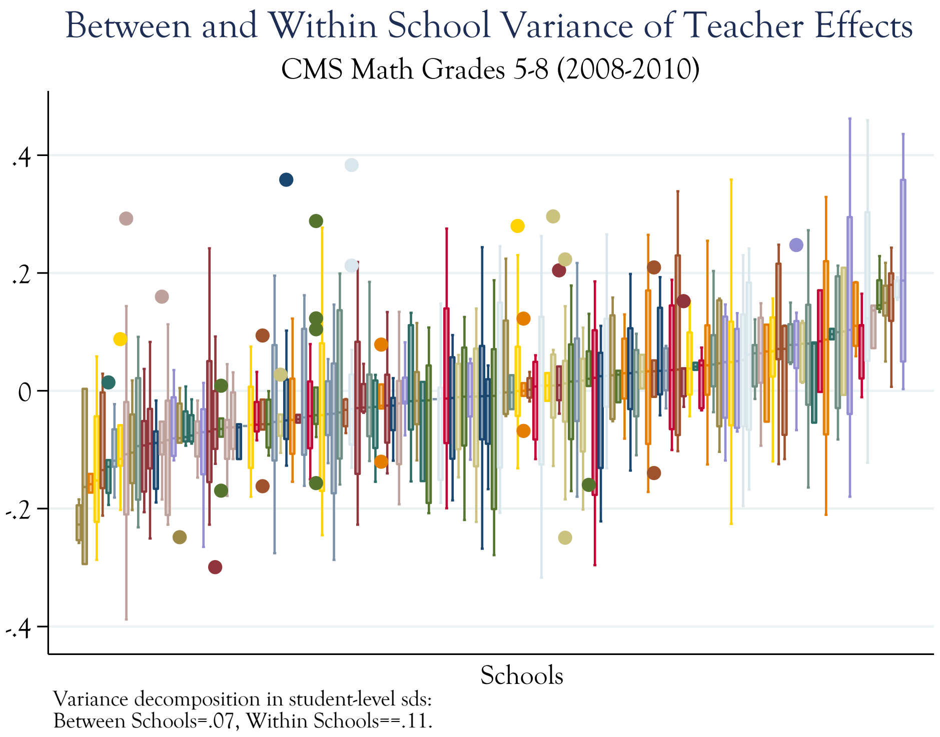
Notes: CMS Grades 5-8 (2008-2010). Panels A-B depict the kernel densities of the standard deviations and means of the teachers' students test scores at t-1. The non-sorted and sorted samples refer to the sample described in Table 12, columns 3-4 and 5-6, respectively.

FIGURE 13: Distribution of Teacher Effects Across and Within Schools.



Notes: CMS Grades 5-8 (2008-2010). The teacher effects plotted in the density curve for across and within school are derived from the preferred model summarized in Table 7, column 4. The teacher effects for the within-school distribution are derived from an estimation of the preferred model that includes school fixed effects.

FIGURE 14: Between and Within-School Variation in Teacher Effects.



Notes: CMS Grades 5-8 (2008-2010). Teacher effects derived from the preferred model preferred model summarized in Table 7, column 4. The between and within school variation is calculated using Stata's -xtsum- command. Each vertical line is a boxplot of the teacher effects for one school over the period 2008-2010).

FIGURE 15: Distribution of High and Low Value-Added Teachers Across Schools.

Panel A: Distribution of Top 25% Teachers Across Schools.

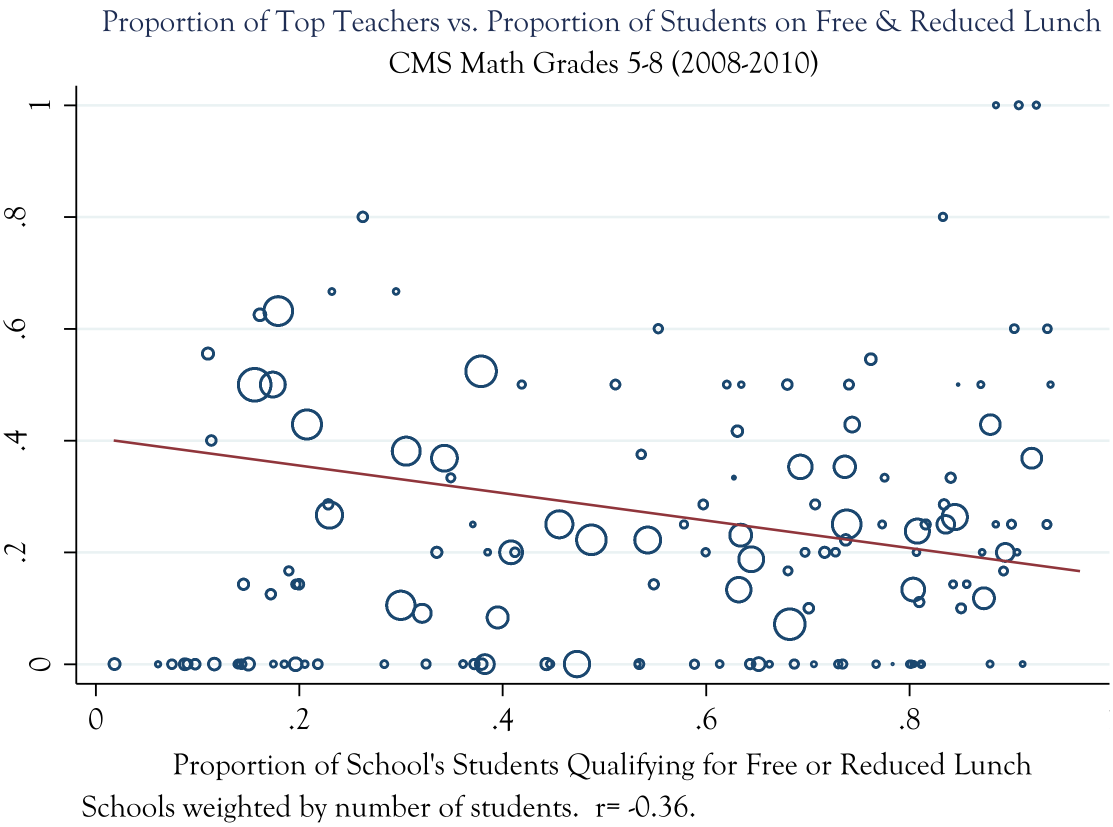
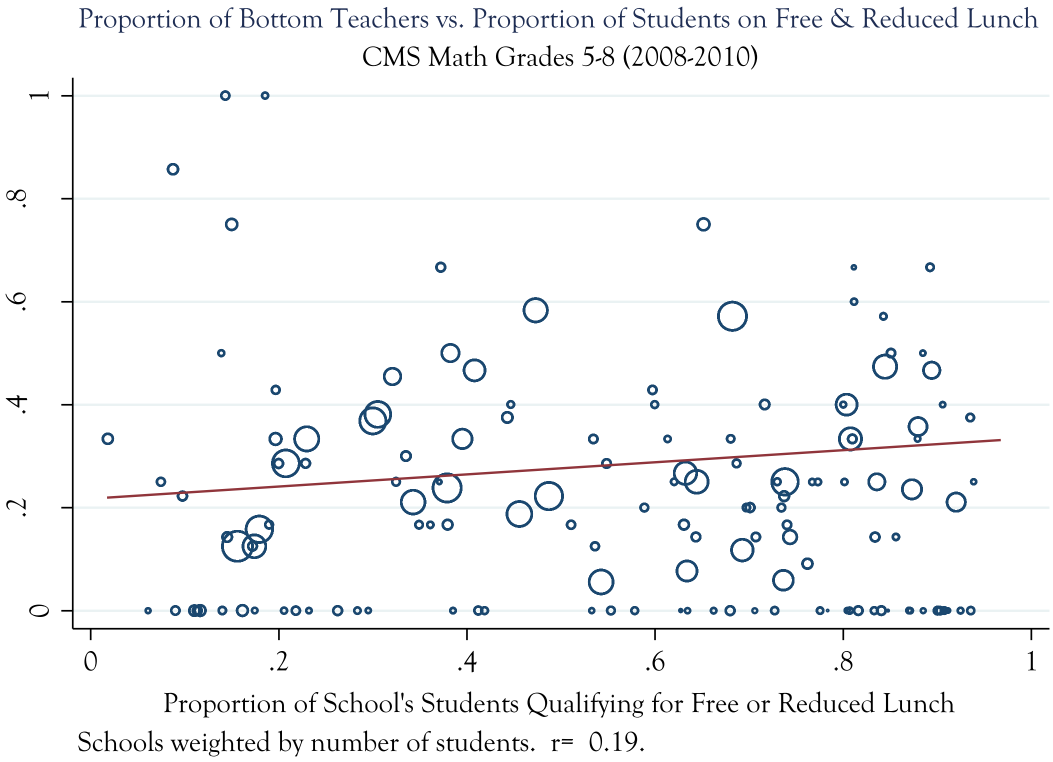


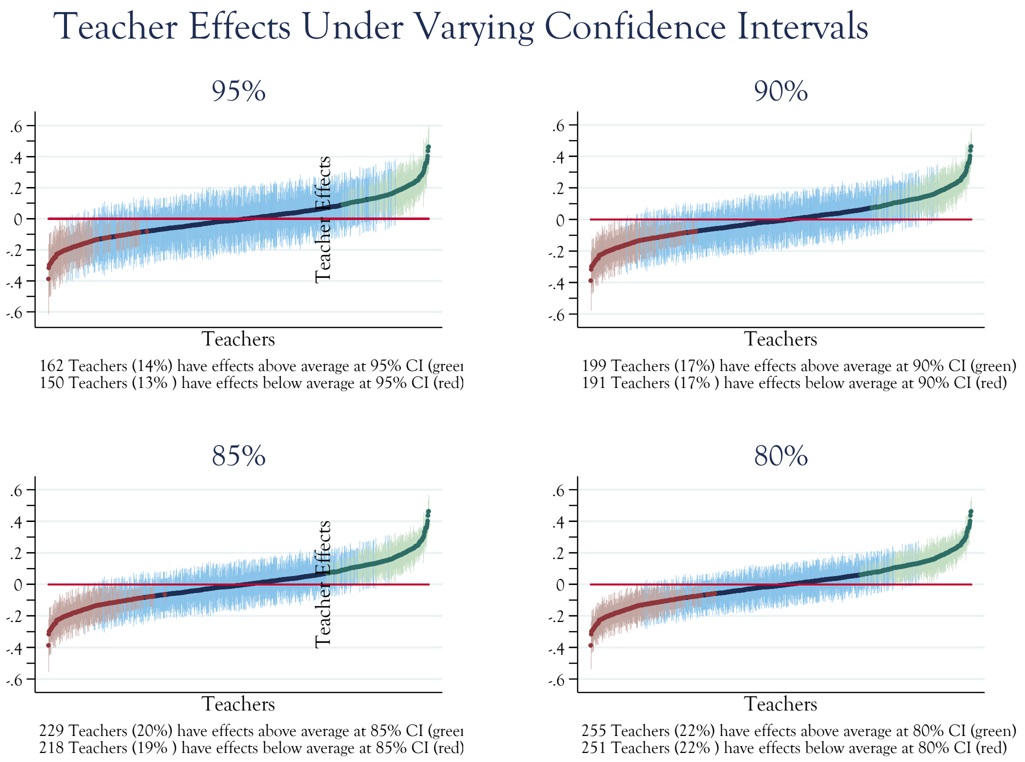
FIGURE 15 (continued)

Panel B: Distribution of Bottom 25% Teachers Across Schools.



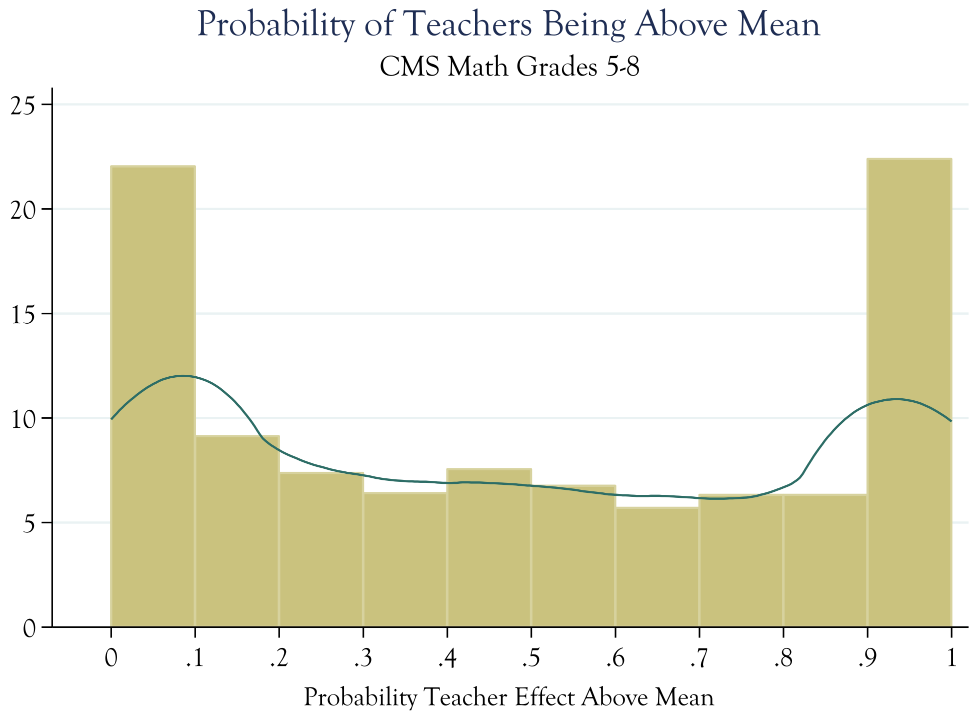
Notes: CMS Grades 5-8 (2008-2010). Panels A-B plot the proportion of a school's teachers in the top and bottom quartile of teacher effects by the school's percentage of students qualifying for free and reduced lunch in 2010. The teacher effects are derived from the preferred model.

FIGURE 16: Teacher Effects under Varying Confidence Intervals.



Notes: CMS Grades 5-8 (2008-2010). The teacher effects are derived from the preferred model.

FIGURE 17: Probabilities of Teachers in Given Quantile.



Notes: CMS Grades 5-8 (2008-2010). The teacher effects are derived from the preferred model. The probabilities are calculated under the assumption that the error around the teacher effect estimate is normally distributed. A kernel density curve is superimposed on the histogram.

REFERENCES

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*(1), 95-135.

Andrabi, T., Das, J., Khwaja, A. I., & Zajonc, T. (2009). Here today, gone tomorrow? Examining the extent and implications of low persistence in child learning. *HKS Working Paper Series, No. RWP09-001*.

Angrist, J. D., & Lang, K. (2004). Does school integration generate peer effects? Evidence from Boston's Metco program. *The American Economic Review, 94*(5), 1613-1634.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E. H., Ladd, H. F., Linn, R., et al. (2010). *Problems with the use of student test scores to evaluate teachers* (No. 278): Economic Policy Institute.

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy, 4*(4), 351-383.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37.

Bonesronning, H., Falch, T., & Strom, B. (2005). Teacher sorting, teacher quality, and student composition. *European Economic Review, 49*(2), 457-483.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008a). Measuring effect sizes: The effect of measurement error. *CALDER Working Paper Series, 19*.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008b). Teacher preparation and student achievement. *National Bureau of Economic Research Working Paper Series, No. 14314*.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2005). The draw of home: How teachers' preferences for proximity disadvantage urban schools. *Journal of Policy Analysis and Management, 24*(1), 113-132.

Boyd, D., Lankford, H., Loeb, S., Wyckoff, J., & Grossman, P. (2008). *Measuring effect sizes, the effect of measurement error*. Paper presented at the National Conference on Value-Added Modeling.

Cantrell, S., Fullerton, J., Kane, T. J., & Staiger, D. O. (2008). National board certification and teacher effectiveness: Evidence from a random assignment experiment. *National Bureau of Economic Research Working Paper Series, No. 14608*.

Card, D., & Krueger, A. B. (1996). School resources and student outcomes: An overview of the literature and new evidence from North and South Carolina. *The Journal of Economic Perspectives, 10*(4), 31-50.

Center for Educational Policy Research. (2010). *Teacher employment patterns and student results in Charlotte-Mecklenburg Schools*. Cambridge, MA: Harvard.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2010). How does your kindergarten classroom affect your earnings? Evidence from Project Star. *National Bureau of Economic Research Working Paper Series, No. 16381*.

Chudowsky, N., Koenig, J. A., Braun, H. I., National Research Council (U.S.). Center for Education., & National Academy of Education. (2010). *Getting value out of value-added : Report of a workshop*. Washington: National Academies Press.

Clotfelter, C. T., & et al. (2004). Do school accountability systems make it more difficult for low-performing schools to attract and retain high-quality teachers? *Journal of Policy Analysis and Management, 23*(2), 251-271.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review, 24*(4), 377-392.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources, 41*(4), 778-820.

Cook, T. D. (2003). Why have educational evaluators chosen not to do randomized experiments? *Annals of the American Academy of Political and Social Science, 589*, 114-149.

Corcoran, S. P., Evans, W. N., & Schwab, R. M. (2004). Women, the labor market, and the declining relative quality of teachers. *Journal of Policy Analysis and Management, 23*(3), 449-470.

Drukker, D. M. (2003). Testing for serial correlation in linear panel-data models. *Stata Journal, 3*(2), 168-177.

Fryer, R. G. (2011). Teacher incentives and student achievement: Evidence from New York City Public Schools. *National Bureau of Economic Research Working Paper Series, No. 16850*.

Fryer, R. G., & Levitt, S. D. (2004). Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics, 86*(2), 447-464.

Fryer, R. G., & Levitt, S. D. (2006). The black-white test score gap through third grade. *Am Law Econ Rev, 8*(2), 249-281.

Glazerman, S., Loeb, S., Goldhaber, D. D., Raudenbush, S. W., & Whitehurst, G. J. (2010). *Evaluating teachers: The important role of value-added*. Washington, DC: The Brookings Brown Center Task Group on Teacher Quality.

Glazerman, S., Mayer, D., & Decker, P. (2006). Alternative routes to teaching: The impacts of Teach for America on student achievement and other outcomes. *Journal of Policy Analysis and Management, 25*(1), 75-96.

Godwin, R. K., & Kemerer, F. R. (2002). *School choice tradeoffs: Liberty, equity, and diversity* (1st ed.). Austin: University of Texas Press.

Godwin, R. K., Leland, S. M., Baxter, A. D., & Southworth, S. (2006). Sinking Swann: Public school choice and the resegregation of Charlotte's public schools. *Review of Policy Research, 23*(5), 983-997.

Goldhaber, D. D. (2007). Everyone's doing it, but what does teacher testing tell us about teacher effectiveness? *Journal of Human Resources, 42*(4), 765-794.

Goldhaber, D. D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources*, *32*(3), 505-523.

Goldhaber, D. D., & Hansen, M. (2010). Using performance on the job to inform teacher tenure decisions. [Article]. *American Economic Review, 100*(2), 250-255.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research, 66*(3), 361-396.

Guryan, J. (2004). Desegregation and black dropout rates. *The American Economic Review, 94*(4), 919-943.

Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources, 14*(3), 351-388.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature, 24*(3), 1141-1177.

Hanushek, E. A. (1996). Measuring investment in education. *Journal of Economic Perspectives, 10*(4), 9-30.

Hanushek, E. A. (2003). The failure of input-based schooling policies. *The Economic Journal*, *113*(485), F64-F98.

Hanushek, E. A. (2005). The economics of school quality. *German Economic Review, 6*(3), 269-286.

Hanushek, E. A. (2006). Chapter 14 school resources. In E. H. a. F. Welch (Ed.), *Handbook of the economics of education* (Vol. Volume 2, pp. 865-908): Elsevier.

Hanushek, E. A. (2010). The economic value of higher teacher quality. *National Bureau of Economic Research Working Paper Series, No. 16606*.

Hanushek, E. A., & Rivkin, S. G. (1997). Understanding the twentieth-century growth in u.S. School spending. *Journal of Human Resources, 32*(1), 35-68.

Hanushek, E. A., & Rivkin, S. G. (2008). Harming the best: How schools affect the black-white achievement gap. *National Bureau of Economic Research Working Paper Series, No. 14211*.

Hanushek, E. A., & Rivkin, S. G. (2009). Do disadvantaged urban schools lose their best teachers?

Hanushek, E. A., & Rivkin, S. G. (2010a). Constrained job matching: Does teacher job search harm disadvantaged urban schools? *National Bureau of Economic Research Working Paper Series, No. 15816*.

Hanushek, E. A., & Rivkin, S. G. (2010b). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2), 267-271.

Harris, D. N. (2011). *Value-added measures in education : What every educator needs to know*. Cambridge, MA: Harvard Education Press.

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

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945-960.

Hoxby, C. (2000). Peer effects in the classroom: Learning from gender and race variation. *National Bureau of Economic Research Working Paper Series, No. 7867*.

Hoxby, C., & Weingarth, G. (2005). Taking race ouf of the equation: School reassignment and the structure of peer effects. Cambridge, MA: Department of Economics, Harvard University.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47*, 5-86.

Ishii, J., & Rivkin, S. G. (2009). Impediments to the estimation of teacher value added. *Education Finance and Policy, 4*(4), 520-536.

Jackson, C. K., & Bruegmann, E. (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics, 1*(4), 85-108.

Jacob, B. A., Lefgren, L., & Sims, D. (2008). The persistence of teacher-induced learning gains. *National Bureau of Economic Research Working Paper Series, No. 14065*.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2006). What does certification tell us about teacher effectiveness? Evidence from new york city. *National Bureau of Economic Research Working Paper Series, No. 12155*.

Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives, 16*(4), 91-114.

Kane, T. J., & Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. *National Bureau of Economic Research Working Paper Series, No. 14607*.

Koedel, C. (2009). An empirical analysis of teacher spillover effects in secondary school. *Economics of Education Review, 28*(6), 682-692.

Koedel, C., & Betts, J. (2007). Re-examining the role of teacher quality in the educational production function. National Center on Performance Incentives, Vanderbilt, Peabody College.

Koedel, C., & Betts, J. (2009a). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the rothstein critique. *Working Papers*.

Koedel, C., & Betts, J. (2009b). Value-added to what? How a ceiling in the testing instrument influences value-added estimation. *National Bureau of Economic Research Working Paper Series, No. 14778*.

Koretz, D. M. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *The Journal of Human Resources, 37*(4), 752-777.

Krueger, A. B. (2003). Economic considerations and class size. *Economic Journal, 113*(485), F34-63.

Ladd, H. F., Hansen, J. S., & National Research Council (U.S.). Committee on Education Finance. (1999). *Making money matter: Financing America's schools*. Washington, D.C.: National Academy Press.

Lazear, E. P. (2001). Educational production. *The Quarterly Journal of Economics, 116*(3), 777-803.

Lazear, E. P. (2003). Teacher incentives. *Swedish Economic Policy Review, 10*(2), 179-214.

Lockwood, J. R., Louis, T. A., & McCaffrey, D. F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics, 27*(3), 255.

Lockwood, J. R., & McCaffrey, D. F. (2009). Exploring student-teacher interactions in longitudinal achievement data. *Education Finance and Policy, 4*(4), 439-467.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47-67.

LoGerfo, L., Nichols, A., & Reardon, S. F. (2006). *Achievement gains in elementary and high schools*. Washington, D.C.: Urban Institute.

McCaffrey, D. F., Han, B., & Lockwood, J. R. (2008). From data to bonuses: A case study of the issues related to awarding teachers pay on the basis of their students' progress. *National Center on Peformance Incentives Working Paper Series, 2008-14*.

McCaffrey, D. F., Han, B., & Lockwood, J. R. (2010). Turning student test scores into teacher compensation systems. In M. G. Springer (Ed.), *Performance incentives : Their growing impact on american k-12 education* (pp. 113-148). Washington, D.C.: Brookings Institution Press.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572-606.

Mihaly, K., McCaffrey, D. F., Lockwood, J. R., & Sass, T. R. (2010). Centering and reference groups for estimates of fixed effects: Modifications to felsdvreg. *Stata Journal, 10*(1), 82-103.

Millman, J., & Darling-Hammond, L. (Eds.). (1990). *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers*. Newbury Park, Calif.: Sage Publications.

Neal, D. (2011). The design of performance pay in education. *National Bureau of Economic Research Working Paper Series, No. 16710*.

Nichols, A. (2008). Fese: Stata module calculating standard errors for fixed effects. .

No child left behind act of 2001, 2001-PL107-110 (2001).

North Carolina Department of Public Instruction (2010). *Fiscal year 2010-2011 North Carolina public school salary schedules*.

Phillips, M., Brooks-Gunn, J., Duncan, G. J., Klebanov, P., & Crance, J. (1998). Family background, parenting practices, and the black-white test score gap. In C. Jencks & M. Phillips (Eds.), *The black-white test score gap* (pp. 103-148). Washington, D.C.: Brookings Institution Press.

Podgursky, M. J., & Springer, M. G. (2007). Teacher performance pay: A review. *Journal of Policy Analysis and Management, 26*(4), 909-950.

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata* (2nd ed.). College Station, Tex.: Stata Press Publication.

Reardon, S. F. (2008). Differential growth in the black-white achievement gap during elementary school among initially high- and low-scoring students. Institute for Research on Education and Policy & Practice.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy, 4*(4), 492-519.

Reber, S. J. (2007). School desegregation and educational attainment for blacks. *National Bureau of Economic Research Working Paper Series, No. 13193*.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review, 94*(2), 247-252.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2008). Can you recognize an effective teacher when you recruit one? *National Bureau of Economic Research Working Paper Series, No. 14485*.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*(4), 537-571.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics, 125*(1), 175-214.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*(434), 473-489.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*(1), 103.

Sanders, W. L., Wright, S. P., & Langevin, W. E. (2010). The performance of highly effective teachers in different school environments. In M. G. Springer (Ed.), *Performance incentives : Their growing impact on american k-12 education* (pp. 171-190). Washington, D.C.: Brookings Institution Press.

Sass, T. (2010). In A. D. Baxter (Ed.). Charlotte, NC.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London ; Thousand Oaks, Calif.: Sage Publications.

Song, J., Felch, J., & Smith, D. (2010, Aug 22). Grading the teachers. *Los Angeles Times*.

Springer, M. G. (2010). *Performance incentives : Their growing impact on american k-12 education*. Washington, D.C.: Brookings Institution Press.

Springer, M. G., Ballou, D., Hamilton, L., Le, V.-H., Lockwood, J. R., McCaffrey, D. F., et al. (2010). *Teacher pay for performance: Experimental evidence from the project on incentives in teaching*. Nashville: National Center on Performance Incentives at Vanderbilt University.

Staiger, D. O. (2009). In A. D. Baxter (Ed.).

Staiger, D. O., & Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives, 24*(3), 97-118.

Steele, J. L., Hamilton, L. S., & Stecher, B. M. (2010). *Incorporating student performance measures into teacher evaluation systems*: RAND Corporation.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal, 113*(485), F3-33.

Todd, P. E., & Wolpin, K. I. (2004). *The production of cognitive achievement in children: Home, school and racial test score gaps*: Penn Institute for Economic Research Department of Economics University of Pennsylvania PIER Working Paper Archive.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New York: The New Teacher Project.

West, B. T., Welch, K. B., & Galecki, A. T. (2007). *Linear mixed models: A practical guide using statistical software*. Boca Raton, FL: Chapman & Hall/CRC.

Wilson, J. Q. (1989). *Bureaucracy: What government agencies do and why they do it*. New York: Basic Books.

Yeh, S. S., & Ritter, J. (2009). The cost-effectiveness of replacing the bottom quartile of novice teachers through value-added teacher assessment. *Journal of Education Finance*, *34*(4), 426-451.