

***DRAFT Report– Please do not cite or distribute without
permission of the authors***

How well do standards-based teacher evaluation scores identify high-quality teachers?

A multilevel, longitudinal analysis of one district

By

Sarah Archibald

March 19, 2007

DRAFT

Paper prepared for the American Educational Research Association Annual Meeting, April 9-13, 2007 in Chicago, IL. The research reported in this paper was supported by a grant from the U.S. Department of Education, Office of Educational Research and Improvement, National Institute on Educational Governance, Finance, Policymaking and Management, to the Consortium for Policy Research in Education (CPRE) and the Wisconsin Center for Education Research, School of Education, University of Wisconsin-Madison (Grant No. OERI-R308A60003). The opinions expressed are those of the authors and do not necessarily reflect the view of the National Institute on Educational Governance, Finance, Policymaking and Management, Office of Educational Research and Improvement, U.S. Department of Education, the institutional partners of CPRE, or the Wisconsin Center for Education Research.

Abstract

States and districts functioning under the federal No Child Left Behind (NCLB) legislation and state standards-based reforms are being measured and potentially sanctioned by their ability or inability to staff every classroom with a highly qualified teacher, ensure that all students make adequate yearly progress (AYP), and submit a plan to address inequitable distributions of teacher quality. All of this emphasis on teacher quality, combined with a general agreement among education researchers with various backgrounds about the importance of teacher quality to student learning, has generated a need for states and districts to identify teachers who are capable of helping students learn to expected levels. One such means of identifying teachers that holds some promise is standards-based teacher evaluation systems. Previous research using a two-level hierarchical linear model has shown that they are valid, reliable means of identifying high-quality teachers (Milanowski & Kimball, 2005; Holtzapple, 2003).

This study improves upon previous studies by analyzing the effect of teachers with varying standards-based teacher evaluation scores in the context of a three-level model with controls for student-, teacher- *and* school-level characteristics. This model more accurately estimates the impact of higher or lower teacher evaluation scores on student achievement. It also accounts for the probability that school-level characteristics play a role in determining how effective teachers are in facilitating student learning, as well as the likelihood that teachers are not evenly distributed across schools, as the previous two-level model assumes. In addition, using three years of data, this study evaluates the stability of standards-based teacher evaluation scores, an important indicator of how well they measure teacher quality.

Introduction

Across the nation, state standards-based reform efforts are underway to raise the level of student achievement, particularly the achievement of students below state-designated proficiency levels, many of whom are students in poverty. With the passage of the No Child Left Behind (NCLB) Act of 2001, the federal government has also become more deeply involved in seeking to boost student achievement, particularly the achievement of students in poverty, students learning English, and students with disabilities. This focus on outcomes forces states to think in terms of the *quality* of inputs rather than merely on the *level*. In the pre-standards-based reform policy environment, states needed to be concerned about the *level* of education inputs for equity reasons, but the *quality* of inputs was largely left to local control. This left room for great variation in the quality of inputs – and outcomes – from one district to another. Disparities in such inputs as per-pupil spending and teacher quality are well documented, both within states and within districts, with low-income, high-minority populations being shortchanged (Condron & Roscigno, 2003; Darling-Hammond, 2000; Ferguson, 1991; Lankford, Loeb, & Wyckoff, 2002; Peske & Haycock, 2006; Roza & Hill, 2004; Sanders & Rivers, 1996).

One of the most critical inputs in the education system is the classroom teacher, but the complex nature of teachers' interactions with students makes it difficult to define and identify teacher quality in a reliable way. Recognizing that the current policy context intensifies this need, this paper analyzes how well one method of evaluating teachers identifies high quality teachers¹.

¹ For the purposes of this study, high-quality teachers are defined as those with the higher evaluation scores that are correlated with higher levels of student achievement.

Actors in many different facets of the education system use different methods of identifying quality teachers. Principals rely on traditional evaluation systems, which often fail to distinguish teachers who are more skilled at facilitating student learning (Peterson, 2000). Researchers have developed methods of distinguishing teachers on other bases, such as identifying the “value added” by the teacher to students’ growth in a school year (Sanders, 2000), but some argue that this statistical approach may not be philosophically appropriate or transparent enough to gain broad support (Darling-Hammond, 1997). States use their own licensure and certification systems, as well as National Board Certification. However, these distinctions tend to identify adequate teachers and a small number of excellent teachers, respectively, but fail to provide yearly information on teacher performance with each classroom of students.

Before the advent of state standards-based reform and NCLB, these methods of attempting to identify high-quality teachers may have sufficed. Even the criteria to satisfy the highly qualified teacher provision of NCLB, teacher certification and a major or minor in the subject taught, may not be enough to adequately identify the teachers who are getting results in the classroom (Aspen Institute, 2007). Rather, this paper argues that the current policy environment’s focus on education outcomes requires a means of identifying quality teachers on the basis of the mechanism by which they influence student outcomes: classroom instruction. Without a direct link between what the teacher is doing in the classroom in a given school year and the assessment that his or her students take during that year, student assessment data gives a less than complete picture of the quality of the teacher in that classroom.

One method that holds some promise in identifying quality teachers in the current policy environment is a standards-based teacher evaluation. Standards-based teacher evaluation is a much more comprehensive process than the traditional principal evaluation, involving multiple observations of classroom practice in which teachers are rated on specific areas of instruction (Danielson and McGreal, 2000). Ongoing CPRE (Consortium for Policy Research in Education) research in Washoe County, (Reno) Nevada has been tracking the association between their standards-based teacher evaluation system and student achievement. Using two-level hierarchical linear models where students are nested in classrooms, researchers have identified a positive, statistically significant relationship: the higher the teacher’s evaluation score, the higher the student’s level of achievement (Milanowski & Kimball, 2005; Milanowski, Kimball & Odden, 2005; Milanowski, 2004).

This study builds on the two-level models used in previous CPRE research, adding a third level with controls for school-level characteristics that may affect student learning such as school size, per-pupil expenditure, and school-level poverty. In doing so, the research questions it investigates are whether, in the context of this more fully specified model, there is evidence that standards-based teacher evaluations are a valid and stable method of identifying high-quality teachers. This study also addresses questions about other teacher-level characteristics as well as some school-level characteristics that may affect the presence and distribution of teacher quality. These include, at the second level, whether more traditional measures of teacher quality (teacher experience and teacher education) explain much of the variation in student achievement. At the third level, this study analyzes the degree of which school size and per-pupil

expenditures may affect student outcomes, as well the percent of students who participate in free or reduced price lunch and the average achievement level at the school. All of these questions are investigated for three separate years of student data linked to teachers linked to schools, allowing for more than a snapshot view of these relationships.

This paper is divided into four sections following this introduction. The first section reviews the literature pertinent to this study. Section two describes the methodology of the study, including the research site, data sample, and statistical models used to investigate the research questions listed in the previous paragraph. Section three gives the results of the analyses, and finally, section four discusses the results and places this study in a broader context.

Section One: Literature Review

The research questions addressed in this paper have in part been addressed in previous studies of teacher effects; those studies are summarized here. These studies have largely measured teacher effects using proxy measures such as teacher experience, teacher education, or teacher subject matter knowledge (just to name a few). The next subsection discusses the evolution of better measures that can be used in teacher effect studies, including this one. However, because this analysis involves three levels – students, teachers, schools – a subsection on school effects is also included, citing previous studies which influenced this work. Finally, the last subsection highlights studies using multilevel models, the tool used to analyze the research questions in this paper.

Teacher effects. Conventional wisdom suggests that classroom teachers have a major influence on how much students learn. However, teacher effect studies have sought to formally parse the proportion of student achievement that can be attributed to the classroom teacher. In the process, some studies have shown that after accounting for student background characteristics, the largest portion of the remaining unexplained variance is due to the characteristics of the classroom teacher (Darling-Hammond, 1996, 2000; Sanders, 2000; Sanders & Rivers, 1996). Other studies confirm the importance of teacher effects, and also help determine the relative size of these effects, including an analysis of Prospects data by Rowan, Correnti & Miller (2002) and a reanalysis of Tennessee STAR data by Nye, Konstantopoulus & Hedges (2004). In the Prospects study, the authors analyzed the different methods by which existing studies calculated teacher effects, and gave ranges for the effects by study type. In the STAR study, the authors reanalyzed the Tennessee STAR data to see which portion was attributable to teacher effects.

These studies have drawn attention to the fact that teacher effects are significant, and have helped put the size of teacher effects into perspective, but they do not help identify which characteristics contribute to teachers' effectiveness. The latter is of vital interest to policymakers, who need information on what makes a teacher effective in order to create policies that help engender those characteristics.

A substantial amount of research has been conducted on different attributes of teachers and their relative influence on teacher effectiveness. In 2002, Wayne and Youngs reviewed the literature related to teacher effects and student achievement, which included 21 studies. Their analysis yielded four categories of teacher characteristics for

which conclusive evidence could be determined: college ratings, test scores, degrees and coursework and certification status. They found that the first two categories, undergraduate institution teachers attended and their test scores², were linked to higher student achievement across subjects and grade levels, and last two, degrees and coursework and certification status, could only be definitively linked to higher achievement in mathematics and only at the high school level.

Reviewing the literature more broadly, numerous studies can be found focused on the categories identified by Wayne and Youngs (including teachers' level of education, certification status, content knowledge, and academic ability), as well as those focused on teachers' level of experience, with mixed findings (e.g. Darling-Hammond, 2000; Hanushek, 1992; Rice, 2003).

Better measures. A number of gaps remain following the studies cited above. First, with the exception of strong evidence that having at least a year of teaching experience has a positive relationship to student achievement, the amount of variation in teacher effectiveness explained by these attributes is often debatable and often quite small even when present. Second, and perhaps even more troubling, all of these studies lack a measure of instructional practice.

In an era of accountability for student results, teacher evaluation is one way to measure teacher's instructional practice, the mechanism by which teachers influence student achievement. However, traditional teacher evaluation systems have not provided the kind of information needed in today's standards-based environment (Peterson, 2000). They tend to have no standards, are scored as pass or fail, use the same standards for

²Two of the tests used in the reviewed studies were the ACT and the Texas Examination of Current Administrators and Teachers (TEACAT).

novice to expert teachers, and are often not linked to learning gains. In their 1987 study of traditional teacher evaluations, Medley and Coker found that principals' ratings of teachers tended to have low correlations with student outcomes, with a mean of .2.

In response to the standards-based reform movement in the 1990s, a number of more comprehensive, standards-based teacher evaluations emerged. Standards-based teacher evaluation systems are comprised of domains, with multiple data sources for each domain, as well as a set of scoring rubrics that help assessors delineate particular teaching behaviors and levels of practice. The rubrics provide evaluators and teachers with a guide to monitor and evaluate teaching performance intended to benefit all students and to a set of performance standards (Kimball, 2002). This type of evaluation system includes multiple sources of evidence including classroom observations and student work. (For more see McGreal, 2000). Odden (2003; 2004) has argued that when considered in total, standards-based teacher evaluations can be used to construct an overall measure of teaching quality.

Emerging research using teacher evaluation scores as a measure of instructional practice have linked these scores to student learning gains (Gallagher, 2004; Kimball, White, Milanowski & Borman, 2004; Milanowski and Kimball, 2005). This study builds on the prior research by offering an analysis of teacher effects that includes this measure of instructional practice but also takes the analysis further by including a third level for the school and studying the stability of the teacher evaluation scores over time.

School effects. With the addition of the third level to control for school-level characteristics, it is also necessary to briefly highlight school effect studies that influenced this work. Raudenbush and Willms (1995) specify two types of school effects,

type A and type B, both of which estimate the difference between the performance of a child in a particular school and the performance that might have been expected if that child had been in another setting. The effects differ on how the alternative setting is specified, with Type A being a school that is average in every way and Type B being a school with identical context to the one the child attends but an average level of effectiveness in practice. This study is not primarily a school effects study, but it does include some analysis of Type A school effects. In this paper, the school effects examined are per-pupil spending, percent free or reduced price lunch, average student achievement, and school size. These variables were identified by the researcher using theories found in the school-effect studies; the rationale behind including each of these variables is provided below.

Per-pupil spending. Empirically it seems that the resources in a school must impact student learning in some way. However, a review of the literature found mixed results on this question – not surprising given that the methodologies – level of analysis, included variables – were different as well. Upon reviewing the studies analyzing the connection between expenditures and student achievement, Hanushek (1989, 1994, 1997) concluded that there was not enough evidence to say anything definitive about the relationship between these two variables. However, in their review of the same literature, Hedges, Laine and Greenwald (1994; Greenwald, Hedges & Laine, 1996) argued that more of these studies than not showed a positive correlation between level of resources and the level of student learning. While this particular study cannot resolve this debate, it could provide evidence of a relationship between resources for instruction and student

learning (or lack thereof) from a more comprehensive model than many of the previous analyses used.

School size. Current policy buzz about small schools suggests that students may learn more in more intimate settings. Indeed, some studies have shown a negative relationship between the size of a school and student achievement, suggesting that smaller schools may be more conducive to learning (Andrews, Duncombe and Yinger, 2002). In addition, in a review conducted by Darling-Hammond, Wise and Pease (1983), school size was one of the factors identified as mediating teacher and student performance.

School-level poverty. Research has also shown that the poverty concentration of a school can impact student learning (Jencks & Phillips, 1998, Jencks & Mayer, 1990; Borman & Dowling, 2003)³, and researching this relationship further seems particularly important in this new era of accountability for the distribution of teacher quality (Peske & Haycock, 2006).

School average achievement. Studies show that student learning is affected by the average level of achievement in the school. As Levine and Painter (2000) show in their school effects study, after controlling for family background, a one standard deviation increase in school test scores raises student achievement by .52 standard deviations. They acknowledge that this estimate is biased upward since some of this effect can be attributed to sorting – higher achieving students tend to go to higher achieving schools.

³ These studies cite not only the effect of poverty as measured by free and reduced lunch (frl) participation but also concentration of minority and ELL students. However, because of the potential for multicollinearity between these variables and the limitations of the sample size for including more variables at level three, these variables are not included in the model.

However, their study used a matched sample to reduce the bias in the estimate, suggesting that school-level achievement has a significant impact on student learning.

Multilevel modeling. Because of the nested nature of student achievement – students are nested in classrooms which are nested in schools – it is necessary to use a statistical tool that accounts for this data structure. Hierarchical Linear Modeling, or HLM, was created to meet this need (Raudenbush and Bryk, 1988).

With the advent of HLM software came more interest in generating models of student achievement that included nested data – students in classrooms (with teachers) in schools. One of the central arguments in Odden, Borman and Fermanich (2004) is that many previous studies where student achievement is the outcome variable have neglected to include some measure of teacher performance in the classroom. Further, many of these studies neglect the teacher level altogether, rendering it impossible for the studies to discern within-school variation. The use of a multilevel model where the variance among students, teachers and schools are allowed to be separate, with separate error terms, is likely to yield a more accurate estimate (Raudenbush and Bryk, 1986; Bryk and Raudenbush, 1988).

An example of this is the famous Coleman study, which assigned much of the responsibility for low achievement to family background characteristics (Coleman et al., 1966), and left the impression that schools did not matter much. However, this study may suffer from some of the methodological complications that have been dealt with using multilevel models that reflect the nested nature of students within classrooms within schools (Bryk and Raudenbush, 1988). A reanalysis of Coleman data using hierarchical linear modeling found that student characteristics were overestimated and school

characteristics were underestimated (Borman and Dowling, 2003). This study, attempting to recreate the Coleman analysis using updated statistical software, included a vast array of variables at the school level, from the percent of blacks in the school, to a mean of parental education, to a measure of the school's physical capital, to its geographic location, and many more.

Two studies have been conducted using the Odden, Borman and Fermanich (2004) three-level model. In the first one, Fermanich (2003) analyzed student outcomes of 5th grade teachers in Minneapolis. In his study, teacher effectiveness was determined by empirical Bayes residuals, and the findings were related to the different methods teachers used in the classroom and to which professional development activities they were engaged. In the second one, Kim (2006) used NELS data in a 3-level HLM analysis that examined which factors impact math and science scores. This study was influenced by the Odden, Borman and Fermanich model, but the data limitations were such that the model could not be fully employed. The next section describes the methodology in more detail.

Section Two: Methodology

This study explores the question of whether standards-based teacher evaluations are a valid and stable method of identifying high-quality teachers. This section begins by providing more information about the research site and the collection of data for this analysis. It then details the analyses included in this study, provides the measures used at all three levels and gives the models used to analyze how well standards-based teacher evaluations identify high-quality teachers.

Research Site and Data Collection

The site on which this study is based is the Washoe County (NV) School District, serving more than 60,000 students residing in Reno, Sparks and outlying communities. The US Census Bureau classifies Reno as a mid-sized city, comparable to Sacramento, California, and Indianapolis, Indiana. In 2006, the district had a total of 93 schools. The district enjoyed steady enrollment growth through the 1970s, 1980s, 1990s and early 2000s, leveling off in the last couple of years. The largest racial group in this district is White, at 57 percent of the population, and the second largest group is Hispanic, comprising another 30 percent. Approximately 41 percent of students are served by the free and reduced-price lunch program, 18 percent qualify as limited-English proficient, and 12 percent qualify for special education. The students in this district perform better than the state average in all subjects on both standardized tests administered, including the Iowa Test of Basic Skills (ITBS) and the state-administered criterion reference test, Nevada CRT.

Washoe County Schools employed more than 3,000 teachers in 2006, at least two-thirds of whom have masters' degrees and 102 are National Board Certified Teachers. Even with these positive indicators of teacher quality, the standards-based reforms of the late 1990s and early 2000s continued to put pressure on the district to improve the achievement of all of its students. Motivated by this pressure to improve and united by their dissatisfaction with the old evaluation system, administrative and teacher leaders worked together to create a system that would recognize high-quality instruction while offering mechanisms for feedback where improvement was needed. In the fall of 2000,

the district implemented a new system closely modeled on Danielson’s Framework for Teaching (1996; see also Kimball, 2002 for a description of the program).

As specified in the Danielson Framework, teachers are evaluated using rubric rating scores over four domains: 1) Planning and Preparation; (2) Classroom Environment; (3) Instruction; and (4) Professional Responsibilities. In Washoe, the principal or assistant principal conducted the evaluations, though research has shown that the use of multiple trained, objective evaluators increases the validity of such systems (Heneman, Milanowski, Kimball & Odden, 2006).

Each teacher is evaluated annually, but only probationary, or non-tenured, teachers are evaluated on all four domains. Probationary teachers are also required to be observed nine times during the school year. Tenured teachers then begin a three year major-minor cycle, the period of time over which they will be evaluated on all four domains. However, because of a desire to have annual information about the instruction domain, teachers not formally being evaluated on this domain are subject to a supplementary evaluation on a subset of instruction-related standards. (A copy of the supplemental evaluation form can be found in Appendix A.) According to Borman and Kimball, composite scores can be calculated from this subset of standards that represent psychometrically sound summary measures of teachers’ instructional performance (2005⁴). As described in the measures portion of this chapter, this study uses the composite measures to maximize the number of teachers that could be included in the analyses.

⁴ Borman and Kimball also found, in line with previous research, that the distribution of high quality teachers is not equal across socioeconomic and racial lines in this district. Poor and minority students have a higher proportion of under-qualified teachers.

Soon after the district implemented the system, they entered into a research relationship with CPRE, sharing evaluation and test score data in exchange for an ongoing analysis of their teacher evaluation system (e.g. Kimball, White, Milanowski and Borman, 2004). Further, having learned from other studies that the most appropriate method for analyzing the effect of teacher characteristics on student achievement is in the context of a three-level model, in order to discern school-level effects it is necessary to study a district where sufficient variation occurs at the third level. The data from Washoe satisfy this criterion. The richness of the data available from this district, in terms of the ability to employ a multilevel model and the opportunity to look at more than one year, makes this site an excellent choice for this study.

The data was obtained from two primary sources. Most of the school-level data came from the district-wide accountability report on the Washoe County School District website. Most of the student and teacher level data was obtained from contacts at the district office; as previously mentioned, the reciprocal relationship between CPRE and the Washoe school district provided CPRE researchers with data access in exchange for evaluation.

Sample

This study uses the aforementioned data from elementary⁵ schools in the Washoe County, Nevada, School District. The data come from grades 3 and 5 in three consecutive school years, 2002-03, 2003-04 and 2004-05. (Although the author, in a previous paper, included grades 3-6 in a three-level analysis of 2002-03 and 2003-04, the school district was only able to provide grades 3 and 5 for the 2004-05 school year, and

⁵ Only elementary schools are included in this analysis because data was only available for grades 3 and 5 across all three years.

therefore the decision was made to analyze all three years with two grades combined but also report the results from the two years for which four grades could be combined.)

Three years of data were included in the analysis to minimize the chance that the findings would be anomalous as well as to better address the issue of stability in the standards-based teacher evaluation scores.⁶

In order to be included in this analysis, a student needed to be able to be linked to his or her teacher and that teacher needed to be linked to a school. Table 1 shows the number of students, teachers and schools included in each analysis. The two most common reasons for not being able to include a student in the analysis were a missing pre- or post-test score or a teacher who was not evaluated in the year for which student test scores were available; if a teacher did not have evaluation scores, his or her students could not be included either.

Table 1
Size of Groups in Analyses

	2002-03 Grades 3 & 5 (Grades 3-6)	2003-04 Grades 3 & 5 (Grades 3-6)	2004-2005 Grades 3 & 5
Math: 2 Level			
Students	3223	5297	4690
Teachers	224	294	246
Math: 3 Level			
Students	3113 (8173)	5274 (10724)	4690
Teachers	220 (454)	291 (521)	246
Schools	55 (59)	58 (59)	56
Reading: 2 Level			
Students	3223	5273	4686
Teachers	224	294	246
Reading: 3 Level			
Students	3113 (8173)	5250 (10695)	4686
Teachers	220 (454)	291 (521)	246
Schools	55 (59)	58 (59)	56

⁶ The student and teacher level data for 2002-03 and 03-04 was obtained and cleaned by Tony Milanowski and Steve Kimball – this study added the school level data. The data for 04-05 were cleaned for this study specifically by the author.

Analysis

To address the question of how well standards-based teacher evaluations identify high-quality teachers, the study uses a quantitative analysis in the form of a covariate adjustment model with three nested levels of data, students nested in classrooms, nested in schools. The dependent variable is student-level post-test score. This study uses a covariate adjustment model because the district's use of different tests precluded the use of a growth model. Unlike a growth model, a covariate adjustment model is not actually a gain score or a true measure of growth – it simply uses the student pretest score, an indicator of student achievement status, as a first-level predictor – so the outcome variable specifies the extent to which the student achievement status at the time of the posttest, controlling for student characteristics, classroom/teacher characteristics and school-level characteristics, differs from the expected score. The study also calculates the correlation between the residuals at level two and the evaluation scores to examine the relationship between the evaluation score and student achievement.

Measures

The following discusses the student, teacher/classroom and school measures used in this study, and mentions the strengths and limitations of the data where appropriate.

Student measures. Student demographic data collected and maintained by the district were used to construct a series of dummy variables describing student background. These variables include minority status, gender, participation in special education and participation in free or reduced price lunch. Test score data for reading and mathematics from the 2002-2003, 2003-2004 and 2004-2005 school years were obtained from the district; students' pretest score was used as a control variable in this

analysis and posttest score was used as the outcome variable. To allow combination of student samples across grades in which different tests were administered, student test scores were transformed into z-scores within grade and subject.

Teacher measures. As detailed in the literature review, it makes sense for a study in which student achievement is the outcome variable to include the mechanism by which teachers help students achieve: classroom instruction. For a measure of teacher instructional practice, this study used the teacher's standards-based evaluation score derived from the district's performance-based evaluation system. This score is used as a measure of teacher quality based on the theory that such a measure must be tied directly to the teacher's classroom instruction. Including this variable in the analysis is one of the strengths of this study. More detail about what the composite teacher performance measure includes can be found in Appendix A.

Also included at the teacher level is a dummy variable indicating whether or not the student's teacher has a master's degree, and a dummy variable indicating whether a teacher is experienced (defined here as having more than three years of teaching) or inexperienced (which includes teachers in their first, second and third years of teaching).

School measures. At the school level, school size as indicated by total enrollment is included to test its relationship to student achievement. The model also includes a per-pupil spending figure from the Nevada state report card website. Research has also shown that the overall level of SES of a school can impact student learning. Accordingly, the analysis tested the effect of the percent of students qualifying for free or reduced-price lunch on student achievement. Finally, beyond a student's own prior achievement as a

control, this study included a measure of school average achievement, since this has also been found to affect student learning.

Models

This analysis was conducted using a 3-level HLM model developed by Bryk and Raudenbush (1988). This software enables researchers to more easily parse the variance that occurs within classrooms, mostly tied to student characteristics, the variance that lies between classrooms, mostly tied to teacher characteristics, and the variance that lies between schools, mostly tied to school characteristics or characteristics of the students and/or teachers as a whole. Some of the variation at each of the levels, as can be observed in the models below, is due to randomness, or error, but this program also allows researchers to partition the error by level. In addition, these models only examine fixed effects, meaning that the intercept for each variable was allowed to vary, but the slope was not. All of the independent variables are centered around the grand mean to ease the interpretation of the results.

Level-1 Model

$$Y_z = P_0 + P_1*(FRL) + P_2*(FEMALE) + P_3*(ETHNIC) + P_4*(SPECED) + P_5*(ZPRE) + E$$

Where:

Y_z is the model's standardized estimate of the student's posttest score, holding the other characteristics constant⁷.

FRL indicates whether or not the student participates in the free or reduced price lunch program (1) or not (0).

FEMALE indicates whether the student is male (0) or female (1).

⁷ Since the outcome variable is standardized, there is no need to put in a dummy for grade-level at the student level. The average student's score has a mean of 0 within each grade.

MINORITY indicates whether the student is a member of a minority group (1) or not (0).

SPECED indicates whether the student qualifies for special education services (1) or not (0).

ZPRE is the standardized pre-test score for the student for whom the model is predicting the post-test score (in either reading or math).

Level-2 Model

$$P0 = \beta00 + \beta01*(MA) + \beta02*(FIRSECTH) + \beta03*(PCOMAVG) + R0$$

Where:

MA indicates whether or not the teacher has a master's degree (1) or not (0).

FIRSECTH is a dummy variable indicating whether a teacher is in the first three years of his or her teaching career (1) or has four or more years of experience (0).

PCOMAVG is a variable indicating the teacher's performance evaluation score on the district's standards-based teacher evaluation system.

Level-3 Model

$$\beta00 = G000 + G001*(PERFRL) + G002*(PERPUP) + G003*(SCHSIZE) + G004*(AVGACH) + U00$$

Where:

PERFRL is a variable indicating the percent of students participating in the free or reduced-price lunch program in a given school.

PERPUP is a variable that indicates how much the district spends per pupil at the school where the teacher teaches and the student attends.

SCHSIZE is a variable indicating the number of students enrolled at the school in question.

AVGACH is a variable indicating the average student achievement level at the school.

This three-level model is primarily focused on the impact of the teacher on student achievement, holding constant student- and school-level variables that may also

influence student learning. The results of the analyses using these models are presented in the next section.

Section Three: Results

Returning to the research questions, the goals of this study were to further determine the validity and stability of standards-based teacher evaluation scores and evaluate whether these scores identify high-quality teachers. Previous studies found evidence of a valid relationship between these scores and student achievement. One of the goals of this study was to test the utility of using a three-level model to predict the relationship rather than the two-level used in previous studies. An initial test to determine whether a three-level model was needed involved running an empty three-level model for reading and math and checking to see how much variation is at the school level. Tables 2, 3 and 4 provide the variance decomposition for all of the models. Note that the amount of variance to be explained at the school level across all years at subjects was a minimum of 11 percent and a maximum of nearly 18 percent, confirming the need for a 3-level model.

Table 2
Variance Decomposition for 2002-2003 Models

	Level 1	Level 2	Level 3
Math Grades 3 & 5			
Empty Model	.801	.053	.160
3-Level Model	.527	.036	.009
Math Grades 3-6			
Empty Model	.735	.079	.177
3-Level Model	.436	.073	.010
Reading Grades 3 & 5			
Empty model	.868	.064	.125
3-Level Model	.543	.037	.007
Reading Grades 3-6			
Empty Model	.816	.041	.151
3-Level Model	.443	.028	.004

Table 3
Variance Decomposition for 2003-2004 Models

	Level 1	Level 2	Level 3
Math Grades 3 & 5			
Empty Model	.808	.046	.130
Full 3-Level Model	.463	.040	.023
Math Grades 3-6			
Empty Model	.766	.055	.147
Full 3-Level Model	.404	.053	.016
Reading Grades 3 & 5			
Empty model	.788	.023	.164
Full 3-Level Model	.382	.023	.007
Reading Grades 3-6			
Empty Model	.781	.021	.163
Full 3-Level Model	.354	.022	.005

Table 4
Variance Decomposition for 2004-2005 Models

	Level 1	Level 2	Level 3
Math Grades 3 & 5			
Empty Model	.857	.033	.109
Full 3-Level Model	.462	.030	.030
Reading Grades 3 & 5			
Empty model	.829	.016	.134
Full 3-Level Model	.379	.015	.010

These tables show that all the models explain approximately half of the variation at the student level, a small amount of the variation at the teacher level (where the least amount of variation exists), and almost all of the variation at the school level. At the student level, much of the unexplained variance is due to student characteristics not captured in this model, such as student motivation and family income (whether the student received free or reduced price lunch has limitations as a poverty indicator). Some of the remaining unexplained variance is error (this is true at all three levels). At the teacher level, the sample used in this study did not include a lot of variation. Of the unexplained variance that occurred at the teacher level that was not due to error, a

substantial portion could be related to characteristics of the teacher or classroom not captured here; the primary variable that could not be included in this study but that likely explains a portion of this variation is class size. Finally, at the school level, between just half a percent and two percent of the variation remains unexplained. The portion of the unexplained variation not due to error is probably related to a school-level characteristic not included in the model such as principal leadership or professional community.

Correlations between teacher evaluation scores and Bayes residual achievement scores. The variance in the empty models shown in Tables 2-4 confirmed the need for a third level of analysis. The next step was to calculate the correlations between the teacher evaluation score and student achievement as measured by the empirical Bayes residual from the HLM models (when the second level includes no predictors). Table 5 provides these correlations for math and Table 6 gives the correlations for reading. In both of these tables, the first row reports the correlations from previous studies using two-level models. Because these correlations include student data from grades 3-6, the next row reports the correlations from the three-level models that combine all four grades; these two correlations are most directly comparable. The third row provides the correlations from two-level models that only include grades 3 and 5, since these are the only grades for which data was available across all three years. The last row gives the correlations from the three-level models using grades 3 and 5; these two rows are also comparable.

Table 5
Correlations for Math, 2 and 3 levels, 2002-03, 2003-04, 2004-05

	2002-2003	2003-2004	2004-2005
Two-level model from previous study	.24**	.21**	NA
Three-level model combining gr 3-6	.078	.129*	NA

Two-level model combining gr 3 & 5	.132*	.224*	.136*
Three-level model combining gr 3 & 5	.067	.165**	.117

*Correlation is significant at the 0.05 level (2-tailed)

**Correlation is significant at the 0.01 level (2 tailed)

Table 6
Correlations for Reading, 2 and 3 levels, 2002-03, 2003-04, 2004-05

	2002-2003	2003-2004	2004-2005
Two-level model from previous study	.25**	.19**	NA
Three-level model combining gr 3-6	.155*	.109*	NA
Two-level model combining gr 3 & 5	.127	.177**	.074
Three-level model combining gr 3 & 5	.073	.111	.031

*Correlation is significant at the 0.05 level (2-tailed)

**Correlation is significant at the 0.01 level (2 tailed)

The results of these comparisons reveal two important findings. The first is that the correlations between standards-based teacher evaluation scores and student achievement are smaller when the school level is included in the analysis. Returning to the theory that adding a school-level to this analysis was desirable because the relationship being modeled included data nested in three levels – students nested in classrooms nested in schools – the fact that the correlations are smaller with the school level included suggests that the two level correlations are probably absorbing some of the variation that is actually at the school level. The second finding is that the correlations that include more cases, shown in rows 1 and 2 of Tables 5 and 6, are larger than those that include fewer cases, shown in rows 3 and 4. This suggests that to estimate this relationship, it is preferable to have as many cases as possible. (Table 3 illustrates the size

of the groups included in the analyses – in many cases, combining all four grades more than doubled the number of cases.)

Given the theoretical basis for using a three-level model rather than a two-level model, the variance decomposition, and the results of the comparisons of correlations, the rest of the results are presented for the three-level models only. However, the results from both the three-level models combining grades 3 and 5, the only grades available across all three years, and the results of the three-level models combining grades 3-6 (in the years available) are presented. Not surprisingly, the results of the comparison of correlations suggest that the more robust estimates come from the models that include more cases.

3-Level HLM analyses. The coefficients for each of the measures from the three-level models are provided in Table 7 for math and Table 8 for reading. Coefficients that are statistically significant are marked with a single asterisk when significant at the .05 level and with two asterisks when significant at the .01 level. Because this analysis uses z-scores, these coefficients can be considered effect sizes. The following paragraphs discuss the results of the measures by level, also translating the effects into percentile points for ease of interpretation (this assumes the test scores are normally distributed).

Table 7
HLM Coefficients for 3-Level Math Models

Level 1	2002-03 (3&5)	2002-03 (3-6)	2003-04 (3&5)	2003-04 (3-6)	2004-2005 (3 & 5)
Intercept ⁸	-0.043	.0119	.0351	.0185	.04368
Female	.0833*	.0083	-0.0422*	-0.0440**	-0.0073
Special Ed	-.4115**	-0.4336**	-0.3587**	-0.3892**	-0.3664**
Minority	-.2194**	-0.1408**	-0.1290**	-0.1185**	-0.1196**
FRL	-.0796*	-0.0493*	-0.0601*	-0.0892**	-0.1140**
Pretest Score	.5020**	.5199**	.5727**	.5809**	.5952**

⁸ Since the outcome variable is standardized, the intercept should be zero; the intercepts shown in Tables 7 and 8 are nearly zero because of a few students being lost from standardizing to analysis.

Level 2					
Evaluation Score	.0551	.0697	.1397*	.1396**	.1057*
Master's	.0796*	-0.0105	-0.0069	.0068	.0294
<= 3 yrs experience	.0203	-0.0017	.0280	.1151*	.159891
Level 3					
Per-pupil spending	0.0001	.0000	-0.0000	-0.0000	.0000
School Size	-0.0003	-0.0003	-0.0002	-0.0016	-0.0001
Percent FRL	-.8671*	-.4213*	-0.2681	-0.0908	.2121*
Avg Pretest Score	-.1630	.1675*	.0161	.1552	.0969

*Significant at the .05 level

**Significant at the .01 level

Table 8
HLM Coefficients for 3-Level Reading Models

Level 1	2002-03 (3&5)	2002-03 (3-6)	2003-04 (3&5)	2003-04 (3-6)	2004-2005 (3 & 5)
Intercept	.0140	.0170	.0439	.0370*	.0440
Female	-0.0938**	-0.0193	.0480*	.0161	0.0575*
Special Ed	-0.243**	-0.3763**	-0.3117	-0.4030**	-0.2374**
Minority	-0.0351	-0.1097**	-0.1280**	-0.1348**	-0.1263**
FRL	-0.0985*	-0.0795**	-0.0820**	-0.0865**	-0.1257**
Pretest Score	.5983**	.6000**	.6295**	.6300**	.6588**
Level 2					
Evaluation Score	.0623	.1092**	.0676*	.0729*	.0255
Master's	.0040	-0.0138	.0043	.0016	.0325
<= 3 yrs experience	.0262	0.0184	-.0052	.0311	.1106
Level 3					
Per-pupil spending	.0001*	.0001*	.0000	.0000	.0000
School Size	-0.0002	-0.0002	-0.0002	-0.0001	-0.0002
Percent FRL	-0.5948*	-0.3902*	-0.3615*	-0.1770*	0.0998
Avg Pretest Score	-0.1150	0.0378	.0206	.0904	.1075

*Significant at the .05 level

**Significant at the .01 level

Level 1 Measures

Most of the level 1 measures were consistent over time and as expected based on theory and prior research. For example, the average posttest score of a student enrolled in special education is considerably lower than a regular education student, and this result is statistically significant. In these analyses, this effect ranges from $-.24$ to $-.43$ of a standard deviation, or 13 to 16.5 percentile points lower for a special education student. Student performance on the previous year's test is consistently a large, statistically significant, positive indicator of his or her performance on the posttest, between $.50$ and $.66$ of a standard deviation, or 19 to 24 percentile points.

Both minority status and participation in the free and reduced-price lunch program are consistently negative and statistically significant, with the exception of reading in 2002-03, when minority status is not a statistically significant indicator of student achievement on the posttest. The effect size for minority status ranged from $-.03$ to $-.22$ of a standard deviation, or between 1 and 9 percentile points lower than for white students. The size of the coefficient for minority status in 2002-03 ($-.03$) is peculiar; it is considerably smaller than in the other analyses. The effect size for participation in free or reduced-price lunch ranged from $-.05$ to $-.13$, or between 2 and 5 percentile points lower than a student not participating in free or reduced lunch. This is considerably smaller than expected; however, in 2003-04 and 2004-05 (not in 02-03), these two variables were highly correlated ($.5$). There was also a correlation between free and reduced lunch and prior test score, though it was not quite as strong, ranging from $.2$ to $.35$.

Another less consistent predictor is the gender of the student. Based on the theory that teachers sometimes assume that boys are better at math and girls are better at reading, this study predicted that the coefficient might be positive for reading and

negative for math. The results do not follow any such predictable pattern. In both reading and math, the female coefficient was positive and statistically significant in some years, negative and statistically significant in others, and in still others not statistically significant. In every case it had a small effect, ranging from $-.09$ to $.08$ of a standard deviation, which, on the normal curve, translates to between a loss of 3.5 percentile points to a gain of 3 points. However, as previously stated, minority status is likely a significant predictor with practical significance as well, but not showing up that way because of collinearity issues. No such explanation is available for gender.

Level 2 Measures

At Level 2, the more traditional proxies for teacher quality, teacher education (coded as having a master's degree or not) and teacher experience (coded as being in the first three years of teaching or beyond), were each statistically significant in only one year of analysis and only in mathematics. In all other instances, these predictors were not statistically significant and the one that one would expect to be positive, master's degree, was sometimes negative, and the one that one would expect to be negative, being an inexperienced teacher, was sometimes positive. In 2002-03, students in third and fifth grade classrooms with teachers who had masters' degrees had math test scores that were $.0796$ standard deviations (or 3 percentile points) higher than those in classrooms with teachers who did not have masters' degrees. Although this is not a large effect, it is close to the magnitude of the teacher evaluation score variable in some years and subjects (see Tables 7 and 8). In 2003-2004, students in classrooms (grades 3-6) with teachers who had been teaching four or more years had math test scores that were $.1151$ standard deviations higher than the test scores of students with less experienced teachers. (In terms of

percentile points on a normal curve, this effect would mean a test score that was 4.5 points higher.) However, these were not consistent effects –neither of these proxies predict student achievement particularly well.

As predicted, the teacher evaluation score appears to be a better measure of teacher quality (where high quality teachers are defined as those whose students meet or exceed expectations on standardized tests). The teacher evaluation score variable was positive in every year of analysis for both subjects and statistically significant in almost all analyses – the exceptions were that there was no statistically significant relationship detected for math in 2002-03 or for reading in 2004-05. The size of this effect varied from .0551 to .1397 standard deviations, or between 2 and 5.5 percentile points. This provides evidence of its validity as an identifier of high-quality teachers (as defined here).

Level 3 Measures

In most of the analyses, school-level poverty was the largest predictor of student achievement. This finding is in line with previous studies, which showed that school context factors play a role in determining student outcomes. With the exception of the analyses for reading and math in 2004-2005, when the effects of this variable were positive (.10 and .21, respectively), this variable was consistently negative, ranging from -.09 to -.87. The -.87 result is from the smaller sample of math students in 02-03; the result from the analysis of the larger sample was -.42, so this could be considered an outlier, especially since the next highest coefficient is -.59. Using this process to eliminate the coefficients that don't seem sensible, the effect for school-level poverty ranges translates to a reduction in test scores by between 3.5 and 22 percentile points.

This finding is in line with many previous studies – the higher the poverty level in a school, the lower the student achievement.

The next largest predictor (and in a couple of instances, *the* largest – see Tables 7 and 8) was average achievement at the school level. In most cases school-level achievement was a positive predictor – the higher the overall level of achievement in a school, the higher the individual student’s score is likely to be. The positive coefficients ranged between .02 and .16 of a standard deviation, which translates to between 1 and 6 percentile points. The exceptions to this were the analyses of the smaller sample of students for 2002-03, when this predictor was negative, though the sign changed to positive when more cases were included.

The detected effect of per-pupil spending was very small (essentially zero). In 2002-2003, per pupil spending was statistically significant in both the analysis combining two grades and the analysis combining all four grades, but still too small to be meaningful. In previous iterations of this study, the researcher explored whether breaking per-pupil spending into smaller, more descriptive categories might prove more useful in identifying the relationship between spending and student achievement. A variable was created that included only expenditures for instruction and instructional support to test this theory, but these more finite measures did not improve the prediction of this relationship in any meaningful way (See Archibald, 2006 for more detail). Using similar logic, this study also tested whether the average level of experience and education of teachers in a school might have an effect on student achievement since these factors determine the average salary and the bulk of school budgets go to teacher salaries.

However, these variables were not significant (either statistically or practically) predictors of student achievement either.

The coefficient for school size was consistently negative, which is consistent with prior research. The effect is quite small, ranging from -.0001 to -.0016, or essentially zero.

Stability

One way to determine whether standards-based teacher evaluation scores are stable over time is to examine the estimates across years. Looking back to Tables 7 and 8 and using the coefficient from the analysis with the larger sample size when possible to compare across years, in math, the coefficients for teacher evaluation scores in 02-03, 03-04 and 04-05 were .07, .14 and .11, respectively. In reading, the coefficients across the same years were .11, .07 and .03. These results do not appear particularly stable.

Another measure of the stability of the teacher evaluation score is to examine how closely these scores from a prior year can predict student achievement in a subsequent year. Milanowski and Kimball (2005) reported these correlations for math and reading twice, using 2001-02 evaluation scores and 2002-03 empirical Bayes residuals test scores and using 2002-03 evaluation scores and 2003-04 empirical Bayes residuals. This study extended this analysis by calculating the correlation between standards-based teacher evaluation scores in 2003-2004 with empirical bayes residuals from 2004-2005. Table 9 displays these results.

Table 9
Predicting Student Achievement using Prior Year Teacher Evaluation Scores

	Grades	Reading	Math
Student Achievement in 02-03 & eval scores in 01-02	4-6	.20*	.27*

Student Achievement in 03-04 & eval scores in 02-03	3-6	.15**	.11**
Student Achievement in 04-05 & eval scores in 03-04	3 & 5	.071	.117

*Correlation is significant at the .05 level

**Correlation is significant at the .01 level

This table shows that the correlation is lower and less significant in the analysis using 04-05 test score data and 03-04 teacher evaluation scores. Particularly with reading in the 04-05 school year, the results from the HLM displayed in Table 8 showed less of a relationship between student achievement and teacher evaluation score than in any of the other years in reading or math. This may provide some explanation of why the correlation is so small for reading in the last row of Table 9. The correlation for math was about the same size as in the previous year but the p-value was .074, and therefore not statistically significant at either the .05 or .01 level. Another possible reason that the correlations are smaller in the last year could be because the sample size was smaller. However, the fact that there is one fewer grade included in the correlation of student achievement in 02-03 to evaluation scores in 01-02 but the correlations are larger than for the following year suggests that this may not be the case. The next section discusses this and other interesting findings and concludes the paper.

Section Four: Discussion and Conclusion

Although there are many ways to judge the success of a teacher or a school, today's standards-based accountability systems focus most on student performance on standardized tests. This paper does not argue the relative merits of such policies, but rather, it uses them as a pragmatic starting point for an analysis designed to give insight into possible policy levers that may impact student achievement on standardized tests.

When using standardized tests as the outcome variable, the ideal is to have vertically scaled tests that can be kept in their original metric and compared across years. This was not possible with this analysis because the tests administered in this district differed from one grade to the next. However, the tests are compatible enough to allow analysis across grades using Z-scores. Although this is not ideal, it is the best possible analysis of this data and it does provide evidence of how different predictors influence a student's deviation from the mean. Another issue in this study is the small number of classrooms in each school that are included in the analyses; this means that this study cannot definitively comment on within-school variation.

As Tables 2-4 showed, the variance decomposition confirmed the need for a 3-level analysis. In addition, the fact that the correlations from the three-level models are smaller than the correlations from the two-level models implies that there are some school level characteristics which affect how teachers influence student achievement, which supports the theory that a three-level model was needed to analyze this relationship.

The fact that there were two occasions, math in 02-03 and reading in 04-05, where no significant relationship between teacher evaluation scores and student achievement was detected could be a result of having too little variance to explain in the first place rather than an indication that there is no such relationship. As shown in Table 4, the variance at the classroom level in the empty model for reading was considerably lower (.016) than in any other year of analysis for either subject. However, this does not help explain why no correlation was found in the 2002-03 analysis; the variation in these analyses was .053 when two grades were combined and .074 when all four grades were

included in the analysis. This is one of the reasons it is important to view these relationships longitudinally; not being able to detect a relationship in two of six instances (counting a school year and subject as one instance) does not necessarily lead one to the conclusion that a relationship does not exist.

Similarly, the results of this study's look at the stability of standards-based teacher evaluation scores over time might have less to do with whether these scores are stable and more to do with the different groups available for analysis in different years – these are not necessarily the same teachers across all three years.

It is hard to say why the link between evaluation scores and student achievement was higher in the first year and then went down in subsequent years. It could be because of testing changes introducing more error into the analysis. Also, as Appendix B shows, the amount of variation in standards-based teacher evaluation scores was lower in the third year of this analysis than in the two prior years – the mean score was higher and the standard deviation was lower. So it may again be a case of there being less variation to detect and therefore the relationship appears less significant than in prior years when there was more variation to explain. Also, the higher mean evaluation score may mean that principals were more generous (and therefore, perhaps, less accurate) in their ratings in the last year included in this analysis.

There were some findings of particular note at the individual and school level of analysis. Recall that this study was guided by Odden, Borman and Fermanich's (2004) theory, which purported that including more variables in an analysis of student achievement that were related to the education students were receiving might reduce the magnitude of the estimated effect that student background characteristics have on student

performance. The relatively small size of the coefficients for both minority and free- or reduced-price lunch found in this study may be an indication that their theory was correct – that the estimated impact of demographic, student background variables (with an indirect relationship to student learning) are reduced in a model with real education variables such as prior achievement and a proxy for the quality of instruction.

At level three, the effect of per-pupil spending was tiny, and as indicated in the results section, other versions of this variable did not increase its effect on student achievement in this model. However, this should not be interpreted as an indication that resources have no relationship to student achievement, but rather, that beyond a possible floor effect (which this study could not test because all of the data come from one district), it is, as Cohen, Raudenbush and Ball (2003) argue, *how* resources are used rather than that more resources necessarily means more student learning. In addition, none of the per-pupil variables used in this analysis included categorical or grant funding.

Perhaps the most important finding from this three-level analysis is that a teacher's standards-based teacher evaluation score remains a positive, and for the most part, statistically significant predictor of student achievement. The magnitude of this effect decreases when some of the variation previously assigned to the teacher level is allowed to move to the school level, but the effect does not disappear. This provides more evidence of the validity of the teacher evaluation score as a measure of teacher effectiveness. The size of the effect is small, it is true, but teacher quality is something that can be more influenced by policy than some factors, such as the family income of the student, and this study reiterates the finding from other similar studies that many students would do at least slightly better if they had a high-quality teacher.

References

- Andrews, M., Duncombe, W., & Yinger, J. (2002). Revisiting Economies of Size in American Education: Are We Any Closer to a Consensus? *Economics of Education Review* 21 (3), 245-262.
- Archibald, S. (2006). Narrowing in on Educational Resources That Do Affect Student Achievement. *Peabody Journal of Education* 81 (4), 23-42.
- Ashton, P. & Crocker, L. (1987). Systematic study of planned variations: The essential focus of teacher education reform. *Journal of Teacher Education*, 2-8.
- Aspen Institute. (2007). Beyond NCLB: Fulfilling our promise to our nation's children. Report of the Commission on No Child Left Behind, Secretary Tommy G. Thompson and Governor Roy E. Barnes, Co-Chairs. Washington, DC: Aspen Institute.
- Borman, G.D., & Dowling, N.M. (2003). Schools and inequality: A multilevel analysis of Coleman's Equality of Educational Opportunity data. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Borman, G. D., & Kimball, S. M. (2005). Teacher quality and educational quality: Do teachers with higher standards-based evaluation ratings close student achievement gaps? *The Elementary School Journal*, 106(1), 3-20.
- Bryk, A., & Raudenbush, S. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education*, 97, 65-108.
- Cohen, D.K., Raudenbush, S. W., & Ball, D. L. (2003). "Resources, Instruction, and Research." *Educational Evaluation and Policy Analysis*, 25 (2): 119-142.
- Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfield, R., & York, R. (1966). *Equality of Educational Opportunity*. Washington, D.C.: Government Printing Office.
- Condrón, D. J. & Roscigno, V. J. (2003). Disparities within: Unequal Spending and Achievement in an Urban School District. *Sociology of Education*, 76, (1), 18-36.
- Danielson, C. (1996). *Enhancing Professional Practice: A Framework for Teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Danielson, C., & McGreal, T.L. (2000). *Teacher evaluation to enhance professional practice*. Alexandria, VA: Association for Supervision and Curriculum Development.

- Darling-Hammond, L. (1996). Restructuring schools for high performance. In S. H. Fuhrman & J. A. O'Day (Eds.), *Rewards and reform: Creating educational incentives that work* (pp. 144-192). San Francisco: Jossey-Bass.
- Darling-Hammond, L. (1997). Toward What End? The Evaluation of Student Learning for the Improvement of Teaching. In J. Millman (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* Thousand Oaks, CA: Corwin Press.
- Darling-Hammond, L. (2000). Teacher Quality and Student Achievement: A Review of State Policy Evidence. *Educational Policy Analysis Archives*, 8 (1); <http://epaa.asu.edu/epaa/v8n1/>
- Darling-Hammond, L., Wise, A. & Pease, S. (1983). Teacher Evaluation in the Organizational Context: A Review of the Literature. *Review of Educational Research*, 53 (3), 285-328.
- Ferguson, R. F. (1991, summer). Paying for Public Education: New Evidence on How and Why Money Matters. *Harvard Journal on Legislation*, 28 (2), 465-498.
- Fermanich, M. (2003). School Resources and Student Achievement: The Effect of School-Level Resources on Instructional Practices and Student Outcomes in Minneapolis Public Schools. Unpublished dissertation, University of Wisconsin-Madison.
- Gallagher, H. A. (2004). Vaughn Elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement. *Peabody Journal of Education*, 79(4), 79-107.
- Greenwald, R., Hedges, L., & Laine, R. (1996). The Effect of School Resources on Student Achievement. *Review of Education Research*, 66(3), 361-396.
- Hanushek, E. (1989, May). The Impact of Differential Expenditures on School Performance. *Educational Researcher*, 18(4), 45-51.
- Hanushek, E. (1992). The Trade-off between child quantity and quality. *Journal of Political Economy*, 100, 84-117.
- Hanushek, E. (1994). Money Might Matter Somewhere: A Response to Hedges, Laine and Greenwald. *Educational Researcher* 23 (4), 5-8.
- Hanushek, E. A. (1997). Assessing the Effects of School Resources on Student Performance: An Update. *Educational Evaluation and Policy Analysis*, 19(2), 141-164.

- Hedges, L. V., Laine, R.D., & Greenwald, R. (1994). Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes. *Educational Researcher*, 23(3), 5-14.
- Heneman III, H. G., Milanowski, A., Kimball, S. M., & Odden, A. (2006). Standards-based Teacher Evaluation as a Foundation for Knowledge-and Skill-Based Pay (RB-45). Philadelphia, PA: University of Pennsylvania, Graduate School of Education, Consortium for Policy Research in Education.
- Holtzapple, E. (2003). Criterion-Related Validity Evidence for a Standards-Based Teacher Evaluation System, *Journal of Personnel Evaluation in Education*, 17 (3), 207-219.
- Jencks, C., & Phillips, M. (1998). *The Black-White Test Score Gap*. Washington, DC: Brookings Institution Press.
- Jencks, C. S., & Mayer, S.E. (1990). The social consequences of growing up in a poor neighborhood. In L. E. Lynn & M. McGeary (Eds.), *Inner-city poverty in the United States* (Vol. 111-186). Washington, D.C.: National Academy of Sciences.
- Kim, J. M. (2006). School, Classroom/Teacher, and Student Effects on Student's Mathematics Achievement. Unpublished dissertation, University of Wisconsin-Madison.
- Kimball, S. (2002). *Washoe County Teacher Performance Evaluation System: A Case Study*. Consortium for Policy Research in Education: University of Wisconsin-Madison.
- Kimball, S.M., White, B., Milanowski, A.T., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education*, 79(4), 54-78.
- Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools. *Educational Evaluation and Policy Analysis*, 24(61), 37-62.
- Levine, D. & Painter, G. (2000). Are Measured School Effects Just Sorting? Identifying Causality in the National Education Longitudinal Survey. Institute of Industrial Relations Working Paper Series: <http://repositories.cdlib.org/iir/iirwps/iirwps-078-00>.
- Medley, D. M., & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *Journal of Educational Research*, 80(4), 242-247.
- Milanowski, A.T. (2004a). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education*, 79(4), 33-53.

- Milanowski, A. T. (2004b). Relationships among dimension scores of standards-based teacher evaluation systems, and the stability of evaluation score-student achievement relationships over time (*CPRE-UW Working Paper Series TC-04-02*). Madison: University of Wisconsin-Madison, Wisconsin Center for Education Research, Consortium for Policy Research in Education. Available online: <http://www.wcer.wisc.edu/cpre/papers/AERA04Measurement.pdf>.
- Milanowski, A., & Kimball, S. (2005). The relationship between teacher expertise and student achievement: A synthesis of three years of data. Paper presented at the American Educational Research Association, Montreal, Quebec, Canada.
- Milanowski, A.T., Kimball, S.M., & Odden, A. (2005). Teacher accountability measures and links to learning. In Stiefel, L., Schwartz, A. E., Rubenstein, R. & Zabel, J.(Eds.), 2005 American Educational Finance Association Yearbook, *Measuring School Performance and Efficiency: Implications for Practice and Research*. (137-161). (Larchmont, NY: Eye on Education).
- Nevada Report Card: <http://www.nevadareportcard.com/>.
- Nye, B., Konstantopoulos, S. & Hedges, L. V. (2004). How Large Are Teacher Effects? *Educational Evaluation and Policy Analysis* 26(3), 237-257.
- Odden, A. O. (2003). An Early Assessment of Comprehensive Teacher Compensation Change Plans. In D. H. Monk & M. L. Plecki (Eds.), *School Finance and Teacher Quality: Exploring the Connections. 2003 Annual Yearbook of the American Education Finance Association* (pp. 209-228). Philadelphia: Eye on Education.
- Odden, Allan R. (2004). Lessons Learned About Standards-Based Teacher Evaluation Systems. *Peabody Journal of Education*, 79(4), 126-137.
- Odden, A. R., Borman, G. & Fermanich, M. (2004). Assessing Teacher, Classroom, and School Effects, Including Fiscal Effects. *Peabody Journal of Education*, 79(4), 4-32.
- Peterson, K. D. (2000). *Teacher Evaluation: A Comprehensive Guide to New Directions and Practices*, 2nd edition. Thousand Oaks, CA: Corwin Press.
- Peske, H. G. & Haycock, K. (2006). *Teaching Inequality: How Poor and Minority Students are Shortchanged on Teacher Quality*. Washington, D.C.: The Education Trust.
- Raudenbush, S. W., & Bryk, A. S. (1986). A Hierarchical Model for Studying School Effects. *Sociology of Education*, 59 (1), 1-17.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: applications and data analysis methods, second edition*. Thousand Oaks, CA: Sage.

- Raudenbush, S. W., & Willms, J. D. (1995). The Estimation of School Effects. *Journal of Educational and Behavioral Statistics*, 20 (4), 307-335.
- Rice, J.K. (2003). *Teacher Quality: Understanding the Effectiveness of Teacher Attributes*. Washington, D.C.: Economic Policy Institute.
- Rowan, B., Correnti, R., & Miller, R.J. (2002). What Large-Scale, Survey Research Tells Us About Teacher Effects on Student Achievement: Insights from the *Prospects* Study of Elementary Schools. *Teachers College Record*, 104(8), 1525-67.
- Roza, M. and Hill, P. (2004). How Within District Spending Helps Some Schools Fail. *Brookings Papers on Education Policy*, 219-226.
- Sanders, W. L. (2000). *Value-added assessment from student achievement data*. Cary, NC, Create National Evaluation Institute.
- Sanders, W. L., & Rivers, J.C. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Knoxville: University of Tennessee Value-Added Research and Assessment Center.
- Washoe County School District. (2006). District-wide Accountability Report. Retrieved March 2006 from (<http://www.washoe.k12.nv.us/district/accountability/>).
- Wayne, A., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, 73(1), 89-122.

Appendix A

Supplementary Teacher Performance Evaluation Form

The composite teacher performance measure is based on the following⁹:

- The teaching displays solid content knowledge and uses a repertoire of current pedagogical practices for the discipline being taught.
- The teaching is designed coherently, using a logical sequence, matching materials and resources appropriately, and using a well-defined structure for connecting the individual activities to the entire unit. Instruction links student assessment data to instructional planning and implementation.
- The teaching provides for adjustments in planned lessons to match the students' needs more specifically. The teacher is persistent in using alternative approaches and strategies for students who are not initially successful.
- The teaching engages students cognitively in activities and assignments, groups are productive, and strategies are congruent to instructional objectives.

⁹ In their 2005 study using some of the same data, Borman and Kimball calculated the item inter-correlations for these composite scores ranged from .69 to .75 and the coefficient alpha reliability was .91. Teacher evaluation results, as measured by the overall composite, averaged about 2.63 on the 0-to-3 point scale.

Appendix B

Descriptive Statistics

2002-03 Grades 3 & 5: Math and Reading

Level 1	N	Mean	SD	Min	Max
FRL	3113	0.18	0.38	0.00	1.00
FEMALE	3113	0.50	0.50	0.00	1.00
MINORITY	3113	0.41	0.49	0.00	1.00
SPECED	3113	0.10	0.30	0.00	1.00
ZMATHPST	3113	0.02	1.02	-3.33	2.93
ZMATHPRE	3113	0.02	1.01	-5.64	3.18
ZREADPST	3113	0.02	1.03	-3.38	3.14
ZREADPRE	3113	0.01	1.02	-3.98	2.83

Level 2

STEP*	220	11.11	6.23	1.00	20.00
PCOMAVG	220	2.67	0.42	.97	3.00
MA	220	0.56	0.50	0.00	1.00
FIRSECTH	220	0.16	0.37	0.00	1.00

Level 3

PERPUPIL	55	5872.24	1069.48	4534	11133.00
SCHLSIZE	55	552.49	140.64	173	874.00
PERFRL	55	0.40	0.29	0.01	0.95
ZMATHPRE	55	-0.03	0.39	-0.76	0.90
ZREADPRE	55	-0.05	0.44	-1.09	0.77

*Although STEP is only included in the models as FIRSECTH, it is useful to see its distribution.

2002-03 Grades 3-6

Level 1	N	Mean	SD	Min	Max
FRL	8173	0.22	0.41	0.00	1.00
FEMALE	8173	0.49	0.50	0.00	1.00
MINORITY	8173	0.38	0.49	0.00	1.00
SPECED	8173	0.11	0.31	0.00	1.00
ZMATHPST	8173	0.02	1.01	-3.85	2.93
ZMATHPRE	8173	0.01	1.01	-5.64	3.18
ZREADPST	8173	0.02	1.01	-3.49	3.14
ZREADPRE	8173	0.01	1.01	-3.98	2.84
Level 2					
STEP	454	10.69	6.41	1.00	20.00
PCOMAVG	454	2.64	0.43	0.97	3.00
MA	454	0.53	0.50	0.00	1.00
FIRSECTH	454	0.18	0.38	0.00	1.00
Level 3					
PERPUPIL	59	5886.08	1062.01	4534	11133.00

SCHLSIZE	59	547.76	137.42	173	874.00
PERFRL	59	0.40	0.29	0.01	0.95
ZMATHPRE	59	-0.04	0.39	-0.76	0.90
ZREADPRE	59	-0.06	0.43	-1.09	0.77

2003-04 Grades 3 & 5: Math

Level 1	N	Mean	SD	Min	Max
FRL	5274	0.41	0.49	0.00	1.00
FEMALE	5274	0.49	0.50	0.00	1.00
MINORITY	5274	0.34	0.47	0.00	1.00
SPED01	5274	0.12	0.32	0.00	1.00
ZMATHPST	5274	0.03	0.99	-3.04	2.80
ZMATHPRE	5274	0.01	0.98	-8.16	1.53
Level 2					
STEP*	291	11.95	6.14	1.00	20.00
PCOMAVG	291	2.69	0.43	1.04	3.00
MA	291	0.60	0.49	0.00	1.00
FIRSECTH	291	0.09	0.28	0.00	1.00
Level 3					
PERPUPIL	58	6278.90	1152.44	4785	12221.00
SCHLSIZE	58	558.40	142.47	175	885.00
PERFRL	58	0.42	0.28	0.01	0.87
ZMATHPRE	58	-0.05	0.38	-0.97	0.64

2003-04 Grades 3 & 5: Reading (Where Different from Math)

Level 1	N	Mean	SD	Min	Max
FRL	5250	0.41	0.49	0.00	1.00
FEMALE	5250	0.50	0.50	0.00	1.00
MINORITY	5250	0.34	0.47	0.00	1.00
SPECED	5250	0.11	0.32	0.00	1.00
ZREADPST	5250	0.04	0.98	-2.83	2.91
ZREADPRE	5250	0.01	1.00	-5.16	1.69
Level 3					
ZREADPRE	58	-0.05	0.43	-0.99	0.64

2003-04 Grades 3-6: Math

Level 1	N	Mean	SD	Min	Max
FRL	10724	0.40	0.49	0.00	1.00
FEMALE	10724	0.49	0.50	0.00	1.00
MINORITY	10724	0.35	0.48	0.00	1.00
SPECED	10724	0.12	0.32	0.00	1.00
ZMATHPST	10724	0.04	0.99	-4.41	2.80
ZMATHPRE	5274	0.02	0.99	-8.16	2.85

Level 2

STEP	521	11.34	6.26	1.00	20.00
PCOMAVG	521	2.67	0.43	1.00	3.00
MA	521	0.58	0.49	0.00	1.00
FIRSECTH	521	0.11	0.31	0.00	1.00

Level 3

PERPUPIL	59	6269.71	1144.64	4785	12221.00
SCHLSIZE	59	555.92	142.52	175	885.00
PERFRL	59	0.42	0.28	0.01	0.87
ZMATHPRE	59	-0.04	0.38	-0.97	0.64

2003-04 Grades 3-6: Reading (Where Different from Math)

Level 1	N	Mean	SD	Min	Max
ZREADPST	10695	0.04	0.98	-3.68	2.91
ZREADPRE	10695	0.02	1.00	-5.16	2.92
Level 3					
ZREADPRE	59	-0.04	0.43	-0.99	0.64

2004-05 Grades 3 & 5: Math

Level 1	N	Mean	SD	Min	Max
FRL	4690	0.42	0.49	0.00	1.00
FEMALE	4690	0.50	0.50	0.00	1.00
MINORITY	4690	0.42	0.49	0.00	1.00
SPECED	4690	0.14	0.34	0.00	1.00
ZMATHPST	4690	0.06	1.0	-2.96	2.65
ZMATHPRE	4690	0.03	0.98	-8.05	1.51
Level 2					
PCOMAVG	246	2.73	0.39	1.00	3.00
MA	246	0.62	0.49	0.00	1.00
FIRSECTH	246	0.04	0.19	0.00	1.00
Level 3					
PERPUPIL	56	6701.38	1284.02	5228	12910
SCHLSIZE	56	340.16	90.90	221	853.00
PERFRL	56	0.44	0.29	0.02	0.90
ZMATHPRE	56	-0.03	0.37	-0.85	0.72

2004-05 Grades 3 & 5: Reading (Where Different from Math)

Level 1	N	Mean	SD	Min	Max
ZREADPST	4686	0.05	0.99	-2.77	2.97
ZREADPRE	4686	0.04	0.99	-4.99	1.41
Level 3					
ZREADPRE	56	-0.04	0.43	-1.01	0.73