Examining the Relationship Between Teacher Evaluation and Student Assessment Results in Washoe County

Article *in* Peabody Journal of Education · October 2004

DOI: 10.1207/s15327930pje7904_4

CITATIONS
66

READS
319

4 authors, including:

Anthony Milanowski
Westat
46 PUBLICATIONS **779** CITATIONS
SEE PROFILE

Geoffrey D. Borman
University of Wisconsin–Madison
67 PUBLICATIONS **2,503** CITATIONS
SEE PROFILE

PEABODY JOURNAL OF EDUCATION, 79(4), 54–78

Examining the Relationship Between Teacher Evaluation and Student Assessment Results in Washoe County

Steven M. Kimball, Brad White, and Anthony T. Milanowski
Consortium for Policy Research in Education
University of Wisconsin–Madison

Geoffrey Borman
Department of Educational Administration and Consortium for Policy Research in Education
University of Wisconsin–Madison

In this article, we describe findings from an analysis of the relationship between scores on a standards-based teacher evaluation system modeled on the Framework for Teaching (Danielson, 1996) and student achievement measures in a large Western school district. We apply multilevel statistical modeling to study the relationship between the evaluation scores and state and district tests of reading, mathematics, and a composite measure of reading and mathematics. Using a value-added framework, the teacher evaluation scores were included at the 2nd level, or teacher level, of the model when other student and teacher-level characteristics were controlled. This study provided some initial evidence of a positive association between teacher performance, as measured by the evaluation system, and student achievement. The coefficients representing the effects of teacher performance on student achievement were positive and were statistically significant in 4 of 9 grade–test combinations studied.

Previous versions of this article were presented at the March 2003 meeting of the American Educational Finance Association in Orlando, FL and the Annual Meeting of the American Educational Research Association, Chicago, April 2003. We are grateful for the collaboration of Superintendent James Hager, district staff, teachers, and school administrators of the Washoe County School District in this research.

The research reported in this article was supported in part 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 OERI–R3086A60003). 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.

Requests for reprints should be sent to Steven M. Kimball, University of Wisconsin–Madison, Consortium for Policy Research in Education, 1025 West Johnson Street, Madison, WI 53706. E-mail: skimball@education.wisc.edu

The quest for valid and consistent measures of teacher behaviors that are related to student learning has long been pursued in educational research and school improvement endeavors. The identification of such measures could help inform a variety of important educational purposes including instructional improvement and accountability, professional development, finance and personnel resource allocations, and teacher compensation reform (Odden, Borman, & Fermanich, 2004/this issue). In the search for adequate measures of teacher or classroom effects on student achievement, teacher performance assessment results could be considered as one possible alternative if the evaluation scores can be shown to be valid measures of teaching practice and to have the expected positive relationship to student achievement.

New, standards-based teacher evaluation practices have recently emerged to respond to historical deficiencies in evaluation practices and improve instruction and accountability (Danielson & McGreal, 2000; Davis, Pool, & Mits-Cash, 2000; Kimball, 2002; Milanowski & Heneman, 2001). As discussed by Milanowski (2004/this issue), not only do these evaluation reforms provide a promising new direction in school personnel evaluation practices, but also the results from evaluations may represent a useful source of information on classroom or teacher effects on student achievement.

The standards-based teacher evaluation system implemented in Nevada's Washoe County School District (WCSD) provides an interesting contrast to the system in Cincinnati Public Schools (Milanowski, 2004/this issue). Although both districts based their evaluation systems on the standards and evidence sources recommended in the Framework for Teaching (Danielson, 1996), Washoe County represents a more typical adaptation of the standards and a more common type of teacher evaluation process. Unlike Cincinnati, Washoe made few changes to the evaluation standards and

did not design its evaluation system for use in high-stakes decisions such as salary determinations. Instead, the evaluation system was designed for low-stakes purposes. It was intended to provide a comprehensive and research-based conception of teaching quality that would guide evaluation discussions, promote formative feedback and teacher reflection on instruction, and substantiate summative evaluation judgments including contract renewal and tenure decisions.

If the teaching standards included in the Framework-based evaluation systems, such as those implemented in Washoe County, represent quality teaching, then one might expect that assessments of teaching behaviors using the standards will reflect measures of student achievement (Milanowski, 2004/this issue). In this article, we explore this hypothesis by analyzing the relationship between teacher behavior, as measured through the evaluation system, and the amount of student achievement attributable to teachers. A positive relationship would provide evidence for the criterion-related validity of the evaluation system and provide a reason to pursue the use of evaluation scores as measures of teacher practice in broader research contexts. Following a brief review of system implementation in Washoe County, the analysis turns to the key question underlying standards-based teacher evaluation reform innovations: Do teachers who score well on such evaluation systems also help produce higher levels of student learning?

Implementation Context

The WCSD encompasses the cities of Reno and Sparks, Nevada, and their outlying areas. The district is the second largest in the state, with over 58,000 students and 84 schools. Thirty-eight percent of the students are non-White, and two thirds of the minority student population is Hispanic. There are over 3,700 certified staff and about 270 administrators working in the district (Washoe County School District, 2003).

In 1997, the WCSD and Washoe County Teachers' Association came to an agreement that teacher evaluation practices in the district were deficient and that they needed a system that would facilitate teacher growth and strengthen instructional accountability. District and association officials recommended a new system that would empower teachers and allow them input into the evaluation process, reduce top-down communication, and attempt to standardize evaluation quality across evaluators (Sawyer, 2001). In addition, the officials wanted a system that would differentiate teacher performance using rubrics that delineated weak to strong instructional behaviors. Under the prior system, examples of exemplary performance were described, but the evaluations resulted only in ratings of satisfactory or unsatisfactory. Because the system lacked multiple performance rubrics, both marginally performing and strongly performing teachers could receive the same satisfactory ratings. The district desired a system based on a progressive set of teaching expectations to monitor and guide a teacher's performance.

A task force made up of members of the teachers' association, principals' association, district staff, and the school board worked collaboratively to restructure the system for evaluating teachers. The task force researched a number of evaluation design alternatives and chose the standards and procedures included in the Framework for Teaching (Danielson, 1996). This evaluation system is intended to measure four domains of practice: planning and preparation (Domain 1), classroom environment (Domain 2), instruction (Domain 3), and professional responsibilities (Domain 4). Each domain has a number of teaching "components," and every component has several "elements." As adapted by the district, each element includes separate behavioral descriptions on a four-level rubric: unsatisfactory, target for growth (Level 1), proficient (Level 2), and area of strength (Level 3). There are 23 components of professional practice and 68 elements in the Washoe County system. The following is an example of one element (knowledge of prerequisite relationships) from the Demonstrating Knowledge of Content and Pedagogy component of the Planning and Preparation domain:

- *Unsatisfactory:* Teacher displays little understanding of prerequisite knowledge important for student learning of the content.
- *Target for growth* (Level 1): Teacher indicates some awareness of prerequisite learning, although such knowledge may be incomplete or inaccurate.
- *Proficient* (Level 2): Teacher's plans and practices reflect understanding of prerequisite relationships among topics and concepts.
- *Area of strength* (Level 3): Teacher actively builds on knowledge of prerequisite relationships when describing instruction or seeking causes for student misunderstanding.

The new, standards-based evaluation system was launched in 2000 after 1 year of planning and 2 years of field testing. The evaluation system calls for multiple sources of evidence to demonstrate teacher performance relative to the standards. Principals or assistant principals serve as teacher evaluators, and they have some discretion in the specific sources of evidence to gather and how the evidence is applied to the standards to make evaluation decisions. Evidence may include a teacher self-assessment, a

preobservation data sheet (lesson plan), classroom and nonclassroom observations with preobservation and postobservation conferences, instructional artifacts (e.g., assignments and student work), a reflection form, a 3-week unit plan, and logs of professional activities and parent contacts. Departing from recommendations by Danielson (1996), there is no instructional portfolio requirement. The combined sources of evidence are intended to provide the basis for evaluators' formative and summative evaluation decisions and related performance feedback.

In the system, teachers advance through three evaluation stages: probationary, postprobationary major, and postprobationary minor. All teachers undergo one of these three stages each year. Teachers who are novice teachers or new to the district are considered probationary and are evaluated on all four of the performance domains in which they must meet at least Level 1 (target for growth) scores on all 68 elements. Probationary teachers are observed at least nine times over three periods of the year, and a written evaluation is provided at the end of each period; based on their performance, they may be required to undergo a second probationary year, be advanced to postprobationary status, or be dismissed from teaching in the district.

Teachers in postprobationary status undergo a "major evaluation" on two performance domains. They are formally observed three times over the course of the year and receive one written evaluation at the end of the year. Once teachers are successfully evaluated under the major evaluation, they move to two "minor evaluation" years. Teachers on the postprobationary minor cycle are evaluated on one domain and are formally observed at least once during the year. Each year of the 2-year minor evaluation results in one written, year-end evaluation. As a result of the evaluation cycle, most teachers are not evaluated on the same domains each year but instead will have all four domains evaluated over a 3-year cycle. For each evaluation, any standard rated unsatisfactory results in an overall unsatisfactory score and the teacher participates in a structured intervention process that results in the teacher moving back into the regular major–minor cycle or the initiation of dismissal proceedings.

Method

Prior research on the impact of the teacher evaluation system in Washoe County has suggested that teachers and administrators agree with and accept the new performance standards and evaluation process (Kimball, 2001). Here, we expand the research scope to explore the relationship

between evaluation scores and student achievement. To the extent that the teacher evaluation scores have a positive relationship with student achievement measures, there will be some evidence that the evaluation system can identify teachers who, by one measure of teaching effectiveness, are producing better results. The analysis will provide evidence of the ability of the evaluation system to distinguish between teachers whose classes show different levels of average student achievement. A substantial positive relationship between evaluation scores and student achievement would suggest that helping teachers improve their practice in accordance with the teaching standards has the potential to contribute to improvements in student learning.

Given the multiple levels of data in this analysis, with students nested in classrooms, we applied multilevel statistical modeling (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999) to estimate the random effects of classrooms on mathematics and reading achievement across three grades. Using hierarchical linear modeling (HLM) software (Raudenbush, Bryk, Cheong, & Congdon, 2001), several two-level models were estimated. The basic strategy at Level 1 was to predict students' posttest scores in reading and mathematics from their demographics and pretest scores. At Level 2, predictors included the teacher's evaluation score and other potentially relevant teacher characteristics. In this way, we estimated the effect of attending a classroom taught by a teacher with a higher or lower evaluation score after statistically controlling student background and other conventional measures of teacher quality.

Measures

Student achievement. We collected several measures of student achievement, including 2000–01 and 2001–02 results from district, state, and national norm-referenced tests for third-, fourth-, and fifth-grade students in mathematics and reading. The tests used for each grade level analysis are presented in Table 1.

Table 1
Student Achievement Measures by Grade

Grade	*Pretest*	*Posttest*
3	District CRT (spring 2001)	State CRT (spring 2002)
4	Terra Nova (fall 2001)	Terra Nova (spring 2002)
5	Terra Nova (spring 2001)	State CRT (spring 2002)

Note. CRT = criterion-referenced test.

The pretest for the third-grade students was a second-grade district criterion-referenced test (CRT) from spring 2001, and the posttest was a spring 2002 state-administered CRT. The fourth-grade student pretest was the Comprehensive Test of Basic Skills (5th edition; CTBS/5), Terra Nova norm-referenced test administered in October 2001, and the posttest was the Terra Nova exam administered in April 2002. The pretest for fifth-grade students was the Terra Nova administered in spring 2001, when these students were in fourth grade. The posttest for these students was the fifth-grade state CRT administered in spring 2002.

Each test was aligned to the state academic content standards. At the elementary level, the district administers criterion-referenced exams it has developed for 1st, 2nd, 4th, and 6th grades.[1] The state administers criterion-referenced exams developed by Harcourt Brace for 3rd- and 5th-grade students. Test items include both constructed and written responses. Individual scaled-score results on the 5th-grade state CRT were used for the 5th-grade analysis. The Terra Nova exams was used as part of the Nevada Proficiency Examination Program that requires norm-referenced testing to be administered in Grades 4, 8, and 10 in reading, language, mathematics, and science (La Marca, 1999). Students were assessed on selected-response items, and results were reported in composite and scaled scores. Scaled-score results were used for our analysis. The Terra Nova was administered statewide in October, and the district also administered the Terra Nova in the spring to 4th-grade students to allow for an assessment of achievement growth over 1 academic year. The Terra Nova has since been discontinued and replaced with the Iowa Test of Basic Skills.

A substantial proportion of the students enrolled in each grade in the 2001–02 school year could not be included in the analysis because either or both the pretest and posttest scores were not available. Other students could not be included because their teachers were not evaluated on the evaluation standards used for this analysis during the 2001–02 school year, because other data were missing, or because they could not be matched with a single classroom teacher. As a result of these factors, 43% of the students tested in third grade were available for the analysis. About 45% of the tested fourth- and fifth-grade students were included in the analysis. Appendix A shows the total number of students by grade, the numbers for whom test scores were available, and the number of students included in the final sample.

[1]Due to incomplete data or large amounts of missing records, results from the first- and sixth-grade exams could not be utilized in this study, and the value-added analysis was not conducted for students in second and sixth grade. If available, these exam results will be used in future analyses.

Other student (Level 1) variables. The district also provided data on other student characteristics including ethnicity, gender, free/reduced-price lunch status, and special education status. This information was used to construct a set of dummy variables for gender (female = 1), minority status (non-White = 1), free/reduced-price lunch (recipient of free/reduced-price lunch = 1), and special education status (recipient of special education services = 1). These variables, along with the pretest scores, were included as controls at Level 1 of the HLM described next. Appendix B provides descriptive statistics from the sample of students and teachers used in the analyses.

Teacher performance. Teacher evaluation scores from the 2001 to 2002 school year were obtained for each teacher who could be matched to students with pretest and posttest scores. As previously mentioned, the evaluation system design results in some teachers being evaluated on one or more teaching domains depending on their status in the evaluation cycle. If a teacher is not evaluated on the instruction domain, they are evaluated using a supplemental evaluation form (which we refer to as the *performance composite*). The supplemental evaluation form consists of selected components and elements from Domain 1 (planning and preparation) and Domain 3 (instruction), representing 7 out of 23 evaluation components. As with the rubrics for the 68 elements, the composite standards were evaluated using four performance designations (i.e., unsatisfactory, target for growth, proficient, and area of strength). The composite standards follow:

- The teaching displays solid content knowledge and uses a repertoire of current pedagogical practices for the discipline being taught (reference: Components 1a, 1c, 3e).
- 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 (reference: Components 1f, 1e, 3f).
- 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 (reference: Component 3e).
- The teaching engages students cognitively in activities and assignments, groups are productive, and strategies are congruent to instructional objectives (reference: Component 3c).

We chose to focus the analysis of teacher evaluation scores on the performance composite measure because although there are a limited number of standards included in the performance composite, the measure represents key elements from two domains relating to instructional practice. In addition, more teachers received evaluation scores for the performance composite than for any one domain. Probationary teachers are evaluated on all domains, and they do not receive the composite scores; however, we were able to compute a composite score by combining the same elements that make up the performance composite, allowing us to include probationary teachers in the study. Similarly, we were able to include those teachers on the major cycle who were evaluated on Domains 1 and 3 but did not have an evaluation on the composite evaluation form because the instruction domain was covered.

The scores on the four performance composites were averaged to derive a single indicator for teacher quality as defined by the evaluation system. As shown in Appendix C, the intercorrlations between the performance composites ranged from .56 to .86 across the three grades. Cronbach's coefficient alpha for our indicator of teacher quality derived from the four composites was .87 for Grade 3, .89 for Grade 4, and .91 for Grade 5, indicating high internal consistency for the measure.

Unfortunately, there were still some teachers with evaluation scores that could not be included in the analysis because they were evaluated on different domains and elements and therefore did not have comparable scores. About 50% to 70% of the teachers for which we had matches to student achievement data and evaluation scores were included at each grade. Included in the final analysis were 123 third-grade teachers, 87 fourth-grade teachers, and 118 fifth-grade teachers.

Other teacher (Level 2) variables. In addition to the teacher evaluation scores, in some of the analyses, we included a combined measure of teachers' education and experience derived from the district salary schedule. Rather than including separate measures of experience and education/training, this combined variable was used in part due to the correlation between our measures of experience and education. Further, using this combined measure allowed for an exploration of whether a relationship existed between teachers' placement on the salary schedule and student achievement.

Finally, because there were 15 elementary schools that operated on a year-round calendar at the time of the study, we controlled for potential differences in results for students attending schools with these different academic schedules. This was particularly important in Grade 4 because there were differences across schools in the number of instructional days between

the fall and spring administrations of the Terra Nova. Arguably, this variable would be more appropriately included as a third- (school-) level variable. Unfortunately, we did not have enough schools and teachers within schools to estimate a meaningful three-level model. Therefore, a dummy variable (year-round = 1) was included at the teacher level to represent whether a teacher's class was on the traditional or year-round schedule.

Analysis

As a first step in the analysis, an unconditional, or "empty," model was used to get an estimate of the total variance in test scores at the student and teacher levels. Then, a random intercept model was estimated with explanatory variables (pretest and student demographics) at Level 1 and no Level 2 predictors. This showed how much Level 2 variance was available to be explained by teacher evaluation scores and other teacher characteristics after controlling for the Level 1 predictors. Next, the Level 2, or teacher-level, prediction models were estimated for the random intercepts. These represented the primary analyses used to address the question of the relationship between evaluation scores and classroom mean achievement. A random slope model was then estimated in which the slopes of the relationship of pretest to posttest were allowed to vary by teacher. Finally, teacher evaluation scores were included as Level 2 predictors of the random slopes of the pretest–posttest relationship.

For the primary analysis, the model at Level 1 was

$$\text{Posttest} = \beta_0 + \beta_1\text{Pretest} + \beta_2\text{Female} + \beta_3\text{Non-White} + \beta_4\text{Special Education} + \beta_5\text{FRL} + R,$$

which represents achievement on the posttest regressed on the pretest score, gender, minority status, special education status, free- and reduced-lunch status, and the Level 1 residual variance (R). All Level 1 predictors were grand-mean centered. The Level 2 model was

$$\beta_0 = \gamma_{00} + \gamma_{01}\text{Evaluation Score} + \gamma_{02}\text{Education/Experience} + \gamma_{03}\text{Year-Round Schedule} + U_0$$
$$\beta_1 = \gamma_{10}$$
$$\beta_2 = \gamma_{20}$$
$$\beta_3 = \gamma_{30}$$
$$\beta_4 = \gamma_{40}$$
$$\beta_5 = \gamma_{50},$$

where the classroom mean achievement is regressed on the teacher evaluation score, education and experience of the teacher, whether the classroom is on a year-round schedule, and the classroom residual variance (U_0). Level 2 predictors were not grand-mean centered, and the slopes for all (β_1-β_5) Level 1 variables were treated as fixed for this analysis.

Because most students were instructed in both mathematics and reading by the same teacher and tested in both subjects in spring 2002, these scores were averaged to create a composite test score for each student. These composite scores were also used to study the relationship between student achievement and the teacher evaluation scores. Separate analyses of mathematics and reading achievement were also conducted. The evaluation system is not subject-matter specific; therefore, combining the two exams allowed an exploration of whether the system was picking up more general pedagogical skill that would be reflected in achievement on the combined tests.

Results

Table 2 reports the proportion of current year test score variance at the teacher level and the reliabilities of the random intercepts at the teacher level for the model with controls for prior year achievement and other student characteristics at Level 1. The table shows that approximately 17% to 27% of the variance in student achievement for each assessment at each grade was at the teacher level without controlling for prior test scores and student characteristics and between 5% to 15% after controlling for these

Table 2

Percentage of Test Score Variance at Teacher Level From Empty Model Compared With Level 1 Covariates and Reliability of Random Intercepts

		% Variance at Level 2		
Grade	*Test*	*Empty Model*	*Model With Level 1 Covariates*	*Reliability of Random Intercepts From Model With Level 1 Covariates*
3	Reading	16.87	5.44	.513
3	Math	18.84	9.64	.656
3	Combined	20.71	8.49	.625
4	Reading	21.35	7.90	.644
4	Math	23.74	14.70	.799
4	Combined	26.48	13.67	.765
5	Reading	21.25	7.89	.648
5	Math	23.73	13.10	.759
5	Combined	26.67	12.17	.744

factors. These results suggest that there is sufficient reliable variation in student achievement at the teacher level to be related to teacher evaluation scores. The chi-square tests for random effects of U_0 were significant for each test in each grade, indicating that average achievement does differ between teachers' classes after controlling for the Level 1 variables.

Teacher Evaluation Effect

The next analysis was aimed at estimating the relationship between the teacher evaluation scores from the performance composite and student achievement while controlling for prior student achievement and the other student characteristics at Level 1 and only the teacher evaluation composite score included at Level 2. This represents the simplest assessment of criterion-related validity because it estimates the relationship between evaluation scores and student achievement without any other potential teacher characteristics or other influences being controlled at the teacher level.[2] Table 3 presents the results from the random intercept model including the student pretest scores and other characteristics controlled at Level 1 and only the teacher evaluation scores at Level 2.

As shown in Table 3, the results are mixed as to whether the teacher evaluation score is a statistically significant predictor of student achievement after controlling for prior achievement and various student characteristics. The evaluation score was a statistically significant and positive predictor of student achievement in four of the nine models, each at or less than the .01 alpha level. These results suggest that for every 1-point increase in teacher evaluation scores, student performance on the fourth-grade reading assessment increased 5.41 points. The teacher evaluation score was not a significant predictor of math achievement or the combined results at the fourth grade. With a 1-point increase in fifth-grade teacher's evaluation score, student performance increased 12.66 points in reading, 20.08 in math, and 16.29 points on the math and reading combination. Although the coefficients were positive for the third-grade results, they were not statistically significant.

Correlation of Bayes residuals and evaluation scores. Another way to examine the criterion-related validity of the evaluation system is by correlating the empirical Bayes intercept residuals, which represents the average student performance attributed to each teacher, with the teacher evaluation scores. We calculated the Bayes residuals with all Level 1 variables in-

[2]There was little difference in the results for the teacher evaluation coefficients when the analyses were run with only the student pretest scores controlled at Level 1.

Table 3

Teacher-Level Results of Random Intercept Model Including Teacher Evaluation Scores

	Third Grade			Fourth Grade			Fifth Grade		
Teacher-Level Variable	*Coefficient*	*SE*	*p*	*Coefficient*	*SE*	*p*	*Coefficient*	*SE*	*p*
Reading evaluation score	5.10	4.78	.287	5.41	2.09	.010	12.66	3.91	.002
Math evaluation score	6.71	5.88	.254	1.20	2.32	.603	20.08	4.48	.000
Combined math and reading evaluation score	5.28	4.98	.289	3.18	1.99	.111	16.29	3.86	.000

cluded and with only the pretest scores included at Level 1. Excluding the other variables at Level 1 might result in more variance available at Level 2 that could be related to performance on the teacher evaluation system. The correlations of the residuals with the evaluation scores including the Level 1 variables were .101 for third-grade reading and math, .279 for fourth-grade reading, .068 for fourth-grade math, .281 for fifth-grade reading, and .374 for fifth-grade math. Consistent with the coefficients from the HLM analysis, there were statistically significant results for the fourth-grade reading and both fifth-grade tests but not for the third-grade results. Dropping the other Level 1 variables did not substantially change the results.

Other Level 2 Variables

In the next analysis, we added variables for the combination of education and experience and a dummy variable indicating which classrooms were on year-round schedules. Table 4 shows the results at Level 2. The coefficients for the student-level variables are also displayed in Appendix D for interested readers.

Education/experience. Like the teacher evaluation scores, teacher education and experience also did not consistently display a statistically significant relationship to student achievement. There were positive but weak effects on student achievement in the fifth-grade math, reading, and the combined test, but the results on the third- and fourth-grade tests were not statistically significant. The results at the fifth grade suggest that a 1-unit increase in the education and experience measure (equivalent to $1,000 on the salary schedule) was related to approximately a ½-point increase on each of the two exams and combined measure. These results are similar to findings in the teacher quality literature that suggest mixed or weak effects of teacher experience and education on student achievement (Wayne & Youngs, 2003).

Year-round effect. The final variable included was whether or not students were taught in classes that operated on a year-round or traditional school calendar. Interestingly, the year-round school variable was a statistically significant and negative predictor in all but three of the grade–subject combinations. In fourth grade, this result may have been due to students in year-round schools having potentially longer vacation breaks between the pretests and posttests than students in schools following the traditional calendar. However, this explanation does not apply to Grades 3 and 5, so further investigation would be necessary to account for this finding.

Table 4

Teacher-Level Results From Final Random Intercept Model

	Third Grade			Fourth Grade			Fifth Grade		
Teacher-Level Variables	*Coefficient*	*SE*	*p*	*Coefficient*	*SE*	*p*	*Coefficient*	*SE*	*p*
Reading model									
Evaluation score	4.74	5.51	.390	5.61	2.00	.005	10.35	3.93	.009
Education and experience	0.04	0.24	.858	0.05	0.11	.673	0.44	0.17	.010
Year-round schedule	–4.50	3.78	.235	–5.64	1.90	.003	–5.75	3.90	.140
Math model									
Evaluation score	5.41	6.48	.404	1.73	2.14	.419	17.25	4.47	.000
Education and experience	0.17	0.31	.580	–0.05	0.13	.706	0.48	0.19	.015
Year-round schedule	–11.32	4.72	.017	–7.49	2.02	.000	–12.88	4.39	.004
Combined reading and math model									
Evaluation score	4.51	5.6	.421	3.51	1.85	.058	13.70	3.81	.001
Education and experience	0.10	0.25	.685	0.03	0.11	.800	0.46	0.16	.004
Year-round schedule	–6.67	3.90	.087	–6.86	1.71	.000	–9.65	3.81	.012

Slope Variation

The next analysis modified the model to allow the slope of the Level 1 relationship of pretest to posttest to vary among teachers. Although very small in absolute terms, each of the fourth-grade results and two of the third-grade results (reading and the combined test results) showed statistically significant variance in these slopes. None of the fifth-grade results showed statistically significant slope variance. The result in fourth grade could be due to relative consistency among the fourth-grade classes because students generally had the same teacher for pretest and posttest and because the exams were similar (both were different forms of the Terra Nova). The explanation would not apply to the third-grade results in which the pretests and posttests occurred in different school years. Adding the teacher evaluation measure to predict these random slopes was not statistically significant in the third-grade tests but did lead to statistically significant results for the fourth-grade math and combined tests. For the fourth-grade math test, the Level 2 random intercept coefficient for the teacher evaluation scores increased by about 0.66 (from 1.73 to 2.39). For the other exams, the coefficients for the teacher evaluation scores remained essentially the same.

Evaluation Score Versus Education/Experience

Finally, to get some idea of the relative strength of the teacher evaluation score in comparison with education and experience as predictors of student achievement, to determine the reduction in variance at Level 2 (teacher level), each predictor was added separately to the random intercept model with Level 1 controls. As shown in Table 5, for six of the nine

Table 5

Percentage of Level 2 Variance Explained by Evaluation Score Versus Education and Experience Combination

Grade	*Test*	*Variance Component at Level 2, Model With Covariates at Level 1*	*% Reduction in Variance Component From Adding Evaluation Score*	*% Reduction in Variance Component From Adding Education and Experience*
3	Reading	186.59	0	0
3	Math	425.10	0.27	0
3	Combined	261.29	0	0
4	Reading	45.74	8.4	0
4	Math	82.79	0	0
4	Combined	57.14	1.4	0
5	Reading	228.33	10.34	7.9
5	Math	408.66	16.6	8.2
5	Combined	272.28	16.5	9.9

grade/tests, the teacher evaluation scores explained more Level 2 (between-teacher variance) than the experience and education composite measure. The proportion of Level 2 variance explained by the evaluation scores ranged from 0% to 16.6%, whereas the proportion explained by education and experience ranged from 0% to 9.9%.

Discussion and Conclusions

The results of our study were mixed with respect to the question of whether teachers' scores on the Washoe County teacher evaluation system were related to the average achievement of those teachers' students, providing only tentative evidence for the criterion-related validity of the evaluation system and use of evaluation scores as measures of classroom effects for other research or educational intervention purposes. The estimated relationship of the teacher evaluation scores to student achievement was positive for each grade and subject and for the reading and math composite, but the coefficients were not statistically significant in all cases. For fourth-grade reading and for each test at Grade 5, the coefficient for the evaluation score was positive and statistically significant. For the other grades and exams, the corresponding coefficients were not significant. However, compared to education and experience as reflected in the district salary schedule, the teacher evaluation scores do help explain more variation in teacher effects.

There are several possible reasons why stronger effects were not found. In addition to the potential for measurement error and other confounding factors (e.g., potential lack of alignment between what is taught and student exams) that may be found in most value-added studies (see Milanowski, 2004/this issue), the evaluation performance composite used in this study represented only 7 of the 23 evaluation components from the teacher evaluation system. Therefore, we could be missing some important information about teacher performance. The limited representation of performance scores may have restricted the range of variation among teachers compared to a more comprehensive performance measure.

The utility of the district exams for value-added analysis was also limited. Results of the second-grade district CRT, which was used as the pretest for the third-grade analysis, had a non-normal distribution. With mean test scores of 79 (out of 100) for second-grade reading and 89 for math, there was clearly a restricted range of test scores that could be related to achievement in the third grade. Transformations of the pretest results did not improve the overall results.

Another reason for the mixed results could be related to the context of teacher evaluation in the district. Although evaluation decisions could lead to nonrenewal or formal dismissal proceedings, such actions are rare. Given the relatively low stakes nature of the evaluation system, it is possible that evaluators were less focused on differentiating teacher performance than they were on improving staff morale through positive feedback and helping teachers identify areas of growth. The result could be lower reliability of evaluation ratings. Indeed, prior research has suggested that in addition to evaluator emphasis on teacher praise and growth, the evidence gathered tends to vary between evaluators (Kimball, 2001). In addition, there has not been a heavy emphasis on evaluation training and oversight for consistency and accuracy. In contrast, the systems in Cincinnati (Milanowski, 2004/this issue) and Vaughn Elementary School (Gallagher, 2004/this issue) were designed for higher stakes, contained multiple raters, and employed more extensive training for rater consistency.

It is also possible that the standards are not specific enough to comprehensively assess teacher performance on key aspects of instruction. It would be interesting to study whether results would have been different if the evaluation system focused more on instructional content and content-specific pedagogy and emphasized uniform sources of teaching evidence. This evaluation system, like others structured closely on the Framework for Teaching (Danielson, 1996), is generic with respect to instructional content. That is, the same evaluation instrument is used for teachers regardless of the content they teach or the grade level of their students. In contrast to the system studied by Gallagher (2004/this issue), the Washoe County system used no content-specific evaluation standards. Although the standards in the Washoe system encourage evaluators to look at the specific content teachers teach and how they teach it, the interpretation is largely up to the individual evaluator. It is possible that tailoring the system to include more content-specific elements may better capture important aspects of instruction. In addition, applying more uniform sources of teaching evidence, such as the structured instructional portfolio approach and use of classroom videotapes recommended by Odden (2003), may provide better and more consistent evidence of instructional content and content-specific pedagogy. With these changes, we would expect improved measures of validity associated with the evaluation system.

Finally, an interesting result was the lower level of average achievement of students in classes that operated with a year-round academic calendar. Although a detailed exploration of these results was beyond the scope of this study, results of a meta-analysis on the effects of modified school calendars on student achievement conducted by Cooper, Valentine, Charlton, and Melson (2003) suggest that although mixed on average, students from

low socioeconomic backgrounds attending schools on modified schedules performed better than their peers in traditional schedule classrooms. An initial examination of our results was inconclusive on average achievement differences for students from low economic backgrounds in the year-round schools. A more comprehensive investigation of these findings would be worth pursuing in future studies.

Limitations and Future Research

This study has several limitations that we will seek to address in future analyses. First, the results are based on only 1 year's worth of data, which yielded tentative conclusions about the value added from a teacher's individual performance. Second, as mentioned previously, the study applied a composite performance measure that may not have captured the full scope of variation in teacher performance as assessed in the evaluation system. Third, a substantial number of students were lost from the analysis due to missing assessment data.

To address these limitations, we intend to collect teacher evaluation and student achievement data for 2 more years. This extension will allow for the replication of results and should provide a broader base for conclusions about the criterion-related validity of the Washoe system. In addition, we will be working with the district to better track test data and maximize the number of teachers and students in the analyses. We also intend to do more analyses with the subset of teachers for whom scores on more of the performance elements used in the evaluation system are available. This will provide some evidence as to whether the performance composites we used were adequate in representing the aspects of teacher performance that are related to student achievement.

The study did not thoroughly examine slope differences due to considerations of time, space, and a desire to retain a parsimonious model. It might be useful in further studies to explore what may be contributing to different slopes in the third- and fourth-grade data and whether slope differences exist in other grades. Different specifications of the models at Level 1 will be compared to those applied in this study to address slope differences and the presence of floor and ceiling effects.

A study is currently being conducted to determine if the evaluation scores given by some evaluators are more strongly associated with average student achievement than those given by others. A preliminary investigation suggests that such variation does exist. If some evaluators' scores are more strongly associated with student achievement, these evaluators may be, in one sense, better judges of teacher performance. By studying these evaluators, it might be possible to find out whether they used different evi-

dence, had better skills, or used a different decision-making process than those whose evaluation scores were not as strongly related to student achievement. If systematic differences between evaluators emerge, this may become the basis for efforts to improve evaluator training in standards-based teacher evaluation systems as well as providing insight into evaluator decision processes.

References

Cooper, H., Valentine, J. C., Charlton, K., & Melson, A. (2003). The effects of modified school calendars on student achievement and on school and community attitudes. *Review of Educational Research, 73*, 1–52.

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.

Davis, D. R., Pool, J. E., & Mits-Cash, M. (2000). Issues in implementing a new teacher assessment system in a large urban school district: Results of a qualitative field study. *Journal of Personnel Evaluation in Education, 14*, 285–306.

Gallagher, H. A. (2004/this issue). 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.

Kimball, S. M. (2001). *Innovations in teacher evaluation: Case studies of two school districts with teacher evaluation systems based on the framework for teaching*. Ann Arbor, MI: UMI Dissertations Publishing.

Kimball, S. M. (2002). Analysis of feedback, enabling conditions and fairness perceptions of teachers in three school districts with new standards-based evaluation systems. *Journal of Personnel Evaluation in Education, 16*, 241–268.

La Marca, P. M. (1999). *Results of statewide TerraNova testing, fall 1998* (Prepared for Nevada Proficiency Examination program). Carson City, NV: Nevada Department of Education.

Milanowski, A. (2004/this issue). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33–53.

Milanowski, A. T., & Heneman, H. G., III. (2001). Assessment of teacher reactions to a standards-based teacher evaluation system: A pilot study. *Journal of Personnel Evaluation in Education, 15*, 193–212.

Odden, A. R. (2003). An early assessment of comprehensive teacher compensation change plans. In M. L. Plecki & D. H. Monk (Vol. Eds.), *Yearbook of the American Education Finance Association 2003: School finance and teacher quality: Exploring the connections* (pp. 209–228). Larchmont, NY: Eye on Education.

Odden, A., Borman, G., & Fermanich, M. (2004/this issue). Assessing teacher, classroom, and school effects, including fiscal effects. *Peabody Journal of Education, 79*(4), 4–32.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S., Bryk, S., Cheong, U. F., & Congdon, R. (2001). *HLM 5: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International, Inc.

Sawyer, L. (2001). Revamping a teacher evaluation system. *Educational Leadership, 58*(5), 44–47.

Snijders, T., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73*, 89–122.

Washoe County School District. (2003). *Fast facts*. Retrieved June 10, 2003, from www.washoe.k12.nv.us/district/facts

Appendix A

Number of Students, Number Tested, and Number Included in Analysis by Grade and Subject

Student Group	*Third Grade*	*Fourth Grade*	*Fifth Grade*
On active roster	4,731	4,867	4,885
Tested in spring 2002			
Reading	4,317	3,965	4,571
Math	4,382	3,985	4,595
Both pre- and posttest scores available			
Reading	2,637	3,800	3,678
Math	2,698	3,838	3,697
In HLM models			
Reading	1,871	1,783	2,122
Math	1,882	1,803	2,131
Combined reading and math	1,858	1,752	2,073

Note. HLM = hierarchical linear modeling.

Appendix B

Descriptive Statistics

Variable	*N*	*M*	*SD*	*Minimum*	*Maximum*
Grade 3					
Teacher level					
Evaluation score	123	2.67	0.39	1.55	3.00
Education and experience	123	43,488.87	7,858.63	29,331.00	54,799.00
Year-round schedule	123	0.33	0.47	0.00	1.00
Student level					
Reading pretest score	1,871	79.37	16.93	17.50	100.00
Reading posttest score	2,216	298.04	74.33	100.00	500.00
Math pretest score	1,882	88.95	11.06	22.50	100.00
Math posttest score	2,239	293.52	80.07	100.00	500.00
Composite pretest score	1,858	84.27	12.93	27.50	100.00
Composite posttest score	2,194	296.94	71.32	100.00	500.00
Female	2,261	.49	0.50	0.00	1.00
Non-White	2,398	0.38	0.48	0.00	1.00
Special education	2,398	0.13	0.34	0.00	1.00
Free/Reduced-price lunch	2,398	0.35	0.48	0.00	1.00
Grade 4					
Teacher level					
Evaluation score	87	2.62	0.43	1.50	3.00
Education and experience	87	43,739.53	8,054.30	28,014.00	54,799.00
Year-round schedule	87	0.29	0.46	0.00	1.00
Student level					
Reading pretest score	1,783	630.18	36.32	502.00	760.00
Reading posttest score	1,802	647.72	38.57	507.00	760.00
Math pretest score	1,803	607.95	31.62	456.00	740.00
Math posttest score	1,814	630.69	35.10	419.00	740.00
Composite pretest score	1,752	619.41	31.09	507.50	726.50
Composite posttest score	1,754	640.02	33.42	509.00	739.00
Female	1,947	0.51	0.50	0.00	1.00
Non-White	1,947	0.37	0.48	0.00	1.00
Special education	1,947	0.09	0.29	0.00	1.00
Free/Reduced-price lunch	1,947	0.35	0.48	0.00	1.00
Grade 5					
Teacher level					
Evaluation score	118	2.58	.44	1.10	3.00
Education and experience	118	43,628.51	8,369.83	28,014.00	54,799.00
Year-round schedule	118	0.31	0.47	0.00	1.00
Student level					
Reading pretest score	2,122	645.34	39.74	499.00	760.00
Reading posttest score	2,622	292.83	71.22	100.00	500.00
Math pretest score	2,131	628.77	33.49	419.00	740.00
Math posttest score	2,629	298.67	72.48	100.00	500.00
Composite pretest score	2,073	637.59	33.55	496.00	736.00
Composite posttest score	2,590	296.43	65.65	100.00	483.00
Female	2,755	0.47	0.50	0.00	1.00
Non-White	2,755	0.36	0.48	0.00	1.00
Special education	2,755	0.14	0.35	0.00	1.00
Free/Reduced-price lunch	2,755	0.35	0.48	0.00	1.00

Appendix C

Intercorrelations and Coefficient Alphas for Teacher Evaluation PC Scores, By Grade

	PC			
Teacher Evaluations	*1*	*2*	*3*	*4*
Grade 3				
PC 1	—			
PC 2	.708	—		
PC 3	.664	.657	—	
PC 4	.560	.617	.572	—
Grade 4				
PC 1	—			
PC 2	.764	—		
PC 3	.575	.628	—	
PC 4	.763	.692	.666	—
Grade 5				
PC 1	—			
PC 2	.857	—		
PC 3	.725	.765	—	
PC 4	.677	.662	.632	—

Note. PC = performance composite.
[a]a = .87. [b]a = .89. [c]a = .91.

Appendix D

Level 1 and Level 2 Results of Final Random Intercept Model

	Third Grade			*Fourth Grade*			*Fifth Grade*		
Variable	*Coefficient*	*SE*	*p*	*Coefficient*	*SE*	*p*	*Coefficient*	*SE*	*p*
Reading model									
Teacher level									
Evaluation score	4.74	5.51	.390	5.61	2.00	.005**	10.35	3.93	.009**
Education and experience	0.04	0.24	.858	0.05	0.11	.673	0.44	0.17	.010**
Year-round schedule	–4.50	3.78	.235	–5.64	1.90	.003**	–5.75	3.90	.140
Student level									
Pretest score	1.84	0.10	.000**	0.73	0.02	.000**	0.81	0.03	.000**
Female	9.35	2.18	.000**	1.37	0.98	.163	–0.30	1.79	.869
Non-White	–6.10	2.95	.038*	–2.79	1.22	.023*	–12.25	2.35	.000**
Special education	–26.9	4.18	.000**	–2.75	2.34	.240	–46.19	4.47	.000**
Free/Reduced-price lunch	–19.25	2.55	.000**	–3.66	1.32	.006**	–5.74	2.68	.032
Math model									
Teacher level									
Evaluation score	5.41	6.48	.404	1.73	2.14	.419	17.25	4.47	.000**
Education and experience	0.17	0.31	.580	–0.05	0.13	.706	0.48	0.19	.015*
Year-round schedule	–11.32	4.72	.017*	–7.49	2.02	.000**	–12.88	4.39	.004**
Student level									
Pretest score	2.33	0.13	.000**	0.72	0.03	.000**	0.93	0.04	.000**
Female	1.32	2.88	.646	0.43	0.98	.662	–6.10	1.85	.001**
Non-White	–10.15	3.30	.003	–0.06	1.36	.963	–10.66	2.69	.000**
Special education	–45.12	4.76	.000**	–4.69	2.54	.065	–48.40	3.91	.000**
Free/Reduced-price lunch	–19.45	3.11	.000**	–4.60	1.31	.001**	–9.56	2.87	.001**

(continued)

Appendix D *(Continued)*

Variable	Third Grade			Fourth Grade			Fifth Grade		
	Coefficient	*SE*	*p*	*Coefficient*	*SE*	*p*	*Coefficient*	*SE*	*p*
Combined reading and math model									
Teacher level									
Evaluation score	4.51	5.60	.421	3.51	1.85	.058	13.70	3.81	.001**
Education and experience	0.10	0.25	.685	0.03	0.11	.800	0.46	0.16	.004**
Year-round schedule	–6.67	3.90	.087	–6.86	1.71	.000**	–9.65	3.81	.012*
Student level									
Pretest score	2.32	0.13	.000**	0.74	0.02	.000**	0.93	0.034	.000**
Female	5.84	2.19	.008**	0.80	0.77	.297	–3.07	1.55	.047*
Non-White	–6.22	2.65	.019*	–0.25	1.00	.799	–9.83	2.14	.000**
Special education	–29.88	4.40	.000**	0.13	1.81	.942	–43.93	3.64	.000**
Free/Reduced-price lunch	–16.31	2.40	.000**	–3.88	1.15	.001**	–7.55	2.51	.003**

*$p < .05$. **$p < .01$.

PEABODY JOURNAL OF EDUCATION, *79*(4), 79–107

Vaughn Elementary's Innovative Teacher Evaluation System: Are Teacher Evaluation Scores Related to Growth in Student Achievement?

H. Alix Gallagher
Center for Education Policy
SRI International

In this study, I examined the validity of a performance-based, subject-specific teacher evaluation system by analyzing the relationship between teacher evaluation scores and student achievement. From a policy perspective, establishing validity was important because it is embedded in a knowledge-and skills-based pay system, which attached high stakes to evaluation scores.

In the first stage of the study, I used hierarchical linear modeling (HLM) to estimate value-added teacher effects, which were then correlated with teacher evaluation scores in literacy, mathematics, language arts, and a composite measure of student achievement. Additionally, teacher evaluation scores were

The research reported in this article 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 OERI–R308A60003). The opinions expressed are those of the author 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.

Requests for reprints should be sent to H. Alix Gallagher, SRI International, Center for Education Policy, 333 Ravenswood Avenue, BS Third Floor, Menlo Park, CA 94025. E-mail: alix.gallagher@sri.com