



CONSORTIUM FOR POLICY RESEARCH IN EDUCATION  
University of Pennsylvania • Harvard University • Stanford University  
University of Michigan • University of Wisconsin-Madison

## The Relationship Between Teacher Evaluation Scores And Student Achievement: Evidence from Coventry, RI

Brad White  
Consortium for Policy Research In Education  
University of Wisconsin-Madison  
Madison, WI 53706  
(608) 263-4260

April, 2004  
CPRE-UW Working Paper Series  
TC-04-04

This paper was prepared for the Consortium for Policy Research in Education, Wisconsin Center for Education Research, University of Wisconsin-Madison for presentation at the American Educational Research Association annual conference held April 12-16, 2004 in San Diego, California. 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 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.

The Coventry (Rhode Island) School District serves around 6,000 students in nine schools. The student population is 97% white, with 13% of students eligible for free or reduced-price lunch, according to 2003 data. There is a wide range of student achievement in the district, with three schools classified as high performing on the state report card, while three others were classified as moderate performers and two were identified as in need of improvement and making insufficient progress. The district is growing rapidly and recently opened a new middle school to accommodate this new population.

In the 1997-98 school year, the district adopted a knowledge- and skills- based teacher evaluation and compensation system (Odden, Archibald, Milanowski, and Conti, 2001). The evaluation standards and rubrics in this new system are based on those described in Charlotte Danielson's *Framework for Teaching* (Danielson, 1996; Danielson and McGreal, 2000). The Framework, as it is referred to by the districts and teachers who use it, identifies 22 different teaching standards organized into four different teaching domains: Planning for Instruction, Managing the Classroom, Instruction, and Professional Activities. In 2003, Coventry revised these standards to create a new framework consisting of three domains (see Appendix I).

The teacher evaluation system in Coventry is similar other standards-based evaluation systems, such as those adopted in Cincinnati, Ohio (Odden and Kellor, 2000), Washoe County (Reno), Nevada (Kimball, 2002), Vaughn Next Century Learning Center in California (Kellor, Milanowski, Odden, and Gallagher, 2001; Kellor, 2003), and other districts. Unlike these districts, however, Coventry's teachers are assessed on all of the domains each time that they are evaluated, rather than focusing on just one or two domain as in the other aforementioned districts.

In Coventry, evaluations consist of teacher observations, pre- and post- observation conferences, dialogue with the teacher, and a review of a portfolio pertaining to the teacher's performance (Coventry Public Schools and Coventry Teachers Alliance, 2003). Evaluations are low-stakes and do not effect teachers' pay in Coventry. Non-tenured teachers are observed by their principal or department head at least twice annually, while tenured teachers are observed by their principal at least once. Teachers are classified as "Unsatisfactory," "Basic," "Proficient," or "Distinguished" based on their performance as compared to the district's evaluation rubrics. The most frequently observed performance level across all standards becomes the teacher's overall performance classification.

Tenured teachers previously classified as Basic are evaluated every two years, while those previously classified as Proficient are evaluated every three years, and those previously classified as Distinguished are evaluated every four years. Tenured teachers classified as Unsatisfactory are assigned a mentor and evaluated the next year. Non-tenured teachers are evaluated annually for their first three years, but only limited subsets of key instructional standards are used during the first and second year, so as to ease teachers into the system.

### **Methodology**

This study attempts to describe the relationship between a teacher's overall evaluation score and his or her students' achievement, while controlling for prior achievement, in order to determine the criterion-related validity of the evaluation scores. In general, the study follows the methodology and rationale described in Milanowski's (2004) study of this relationship in Cincinnati. The analyses use a value-added approach that involves modeling students' post-test scores as a function of a pre-test score and student demographic data. These models will provide estimates of each teacher's performance relative to predicted expectations, which are then correlated with his or her evaluation score to determine the relationship between the two. Two additional analyses will attempt to determine the importance of including student demographic characteristics in these models, and the relationship between teacher experience and student achievement.

### **Measures**

The data used in this study included student pre- and post-test results and demographic information, and teacher evaluation and experience data. Due to the value-added design of this study, we sought to acquire multiple consecutive years of student assessment data that could be matched to teacher data. However, the district's testing schedule only provided appropriate data records for 2<sup>nd</sup>, 3<sup>rd</sup>, and 6<sup>th</sup> grade students and teachers in 1999-2000 and 2000-2001, and for 4<sup>th</sup> grade students and teachers in 2000-2001 and 2001-2002. There were 3,617 student records and 172 teacher records in this initial sample, though some individual students and teachers are counted as multiple records since we were dealing with multiple years of data.

Teacher evaluation scores. We obtained the overall, final evaluation scores for 78 of the 172 teachers that could be matched to student pre- and post- test results, consisting of eighteen 2<sup>nd</sup> grade teachers (12 from 1999-2000 and 6 from 2000-01), twenty-two 3<sup>rd</sup> grade teachers (11 each from 1999-2000 and 2000-01), twenty-five 4<sup>th</sup> grade teachers (14 from 2000-01 and 11 from 2001-02), and thirteen 6<sup>th</sup> grade teachers (5 from 1999-2000 and 8 from 2000-01). Scores were coded as “1” for Unsatisfactory, “2” for Basic, “3” for Proficient, and “4” for Distinguished. Of these scores, 49 teachers (63%) were rated as Proficient and 25 (32%) teachers were rated as Distinguished, while only 4 teachers (5%) were rated as Basic. Since the sample of Basic teachers was so small, we chose to drop these from the analysis and simply examine whether the system could distinguish between Proficient and Distinguished teachers.

Additional teacher data. We also obtained data on teachers’ years of experience teaching in the Coventry district. Since some teachers may have taught several years outside of the Coventry district, this is not to be construed as a measure of teachers’ total experience. These data were available for 19 of the 25 Distinguished teachers and 38 of the 49 Proficient teachers in our sample. The years of experience of our sample ranged from 1 to 29, with a mean of 11 years. Nine of the teachers rated as Proficient had 3 or fewer years of experience in the district, while none of the Distinguished teachers had fewer than 3 years of experience in the district, and the average Proficient teacher had 10.3 years of experience while the average Distinguished teacher had 12.7 years of experience. In subsequent analyses, we will compare the criterion-related validity of this teacher experience measure with that of the teacher evaluation scores.

Student achievement. The value-added nature of this study required matched student assessment results from consecutive years, in order to provide a pre-test to control for student effects on post-test achievement. The matched assessments used in the Coventry analysis are summarized in Table 1. Coventry administered the Stanford 9 at the district level in grades 1-3 and 5-6, while the New Standards reference examinations were administered by the state of Rhode Island in grade 4. Scaled score results for reading and math were used for all assessments.

Table 1

Tests Used for Final Analyses in Coventry

Grade	Year	Posttest	Pretest
2	1999-2000	Stanford 9, Reading & Math	Grade 1 Stanford 9, Reading & Math
2	2000-2001	Stanford 9, Reading & Math	Grade 1 Stanford 9, Reading & Math
3	1999-2000	Stanford 9, Reading & Math	Grade 2 Stanford 9, Reading & Math
3	2000-2001	Stanford 9, Reading & Math	Grade 2 Stanford 9, Reading & Math
4	2000-2001	New Standards Reading & Math	Grade 3 Stanford 9, Reading & Math
4	2001-2002	New Standards Reading & Math	Grade 3 Stanford 9, Reading & Math
6	1999-2000	Stanford 9, Reading & Math	Grade 5 Stanford 9, Reading & Math
6	2000-2001	Stanford 9, Reading & Math	Grade 5 Stanford 9, Reading & Math

Additional student data. We also obtained data on relevant student demographics, the availability of which varied by year collected. Data from 1999-2000 include variables for student gender and student age in months, while data from 2000-01 and 2001-02 include variables for student gender, student age, student race, student free and reduced-price lunch status, and student special education status. Gender, race, and free or reduced-price lunch status were coded as dummy variables, while student age was recorded in months old.

**Analyses**

A two-level random intercept hierarchical linear model is used to estimate empirical Bayes (EB) intercept residuals representing the average student performance in a teacher’s classroom controlling for students’ prior achievement and demographics, which are then standardized and correlated with teacher evaluation scores. The basic level 1 model was  $Posttest = \beta_0 + \beta_1 pretest + \beta_2 X_2 + \dots + \beta_n X_n + r$ , where  $X_2 \dots X_n$  represent various student demographic variables. The availability of demographic data varied by year collected, as described above. All Level 1 predictors were grand mean centered and included in the model when available.

The Level 2 specification was simply:  $\beta_{0j} = \gamma_{00} + u_{0j}$ . Here, the  $u_0$  represents the teacher-specific differences from the average of the group intercepts. The standardized EB residuals from this model were used as the measure of the average student performance relevant to each teacher. Given the grand mean centering, the standardized EB intercept residuals represent the difference created for the student with the “average” prior test score and demographic characteristics. The slopes for all Level 1 variables were treated as fixed. These standardized

residuals were then correlated with teacher evaluation scores at each grade level and across all grade levels. These correlations represent estimates of the relationship between evaluation scores and student achievement and help to summarize the criterion-related validity of the evaluation system.

## Results

The correlations between teacher evaluation scores and the estimates of average student achievement based on standardized empirical Bayes intercept residuals obtained from the model are displayed in Table 2. The results indicate a small overall correlation in reading (.240) and essentially no correlation in math (.032). The results also indicate rather large fluctuations in correlations between years and across subjects and grade levels.

Table 2  
Correlations Between Teacher Evaluation Scores and Standardized EB Residuals

Grade	Year	N	Reading	Math
2	1999-2000	10	.387	.314
	2000-2001	6	.739	-.816*
3	1999-2000	10	.145	-.456
	2000-2001	11	.294	.079
4	2000-2001	9 reading, 8 math	-.146	.093
	2001-2002	11 reading, 10 math	.286	.505
6	1999-2000	5	-.521	.589
	2000-2001	8	.216	-.379
Weighted Average	1999-2002	70 reading, 68 math	.240	.032

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

In order to provide an idea of the potential importance of having a more highly-rated teacher, we also calculated the average estimated change in student achievement associated with a one level change in teacher evaluation score. We calculated the number of standard deviations in post-test score that were associated with a change in teacher ratings of one overall level (i.e. from Proficient to Distinguished), and found effect sizes of 0.13 in reading and 0.01 in math.

These results indicate that, for reading, the criterion related validity of the Coventry program is similar to the results obtained in Washoe County (Kimball, White, Milanowski, and

Borman, 2004), but somewhat less than the standards-based evaluation systems in Cincinnati, Ohio (Milanowski, 2004) and Vaughn Charter School in California (Gallagher, 2004). In mathematics, however, there is little to no consistent correlation or effect. Though the observed effects in reading are small, they would add up to a substantial advantage for a student with two or three consecutive teachers rated at the Distinguished rather than the Proficient level. These results should be interpreted with caution, however, due to the small sample size, relatively few years of data, and considerable differences across subjects and grades.

Two additional analyses were also conducted. The first models the data without the student demographic variables, since some researchers argue that pre-test data are sufficient student-level controls. This analysis of standardized EB residual intercepts and teacher evaluation scores yielded an average correlation of .211 in reading and -.003 in math. These findings are similar to the previous results, and indicate that controlling for demographic data has little effect.

The second supplementary analysis compares the validity findings regarding teacher evaluation scores with those regarding teacher experience measures. Using the model with student demographics described above, the correlation between the standardized EB residual intercepts and teachers' years of experience was .175 in reading and -.004 in math. The correlation between experience and evaluation score was .24 and was not statistically significant. These findings indicate that, given the limitations of this study, teacher evaluation scores are a more valid indicator of teacher value-added than teacher experience, though substantial room exists for improving the relationship between evaluation scores and value-added student achievement measures.

## **Conclusions**

The results of these analyses indicate that the Coventry teacher evaluation system has some criterion-related validity in reading, though not in math. There is initial evidence that the evaluation system identifies teachers that produce higher than expected student assessment results in reading, and that the impact of having a Distinguished teacher as opposed to a teacher rated as Proficient equates to a difference of about 0.13 standard deviations in reading assessment results.

The findings with regard to reading achievement are similar to those observed at other sites we have studied. The lack of a consistent and discernable relationship between evaluation scores and math achievement in Coventry is somewhat different, however, and warrants further exploration. One reason for this may be that there simply is not substantial between-teacher variance in value-added student achievement in some grades and in some years, and preliminary analyses suggest this to be the case with our data.

In addition, this study provides some support for the notion that controlling for student demographics, in the context of such studies where prior achievement is already accounted for, does not significantly or substantially alter statistical results. The supplementary analysis also indicates that standards-based teacher evaluation scores can provide some evidence of teacher quality, in terms of value-added student achievement, beyond that offered by teacher experience alone.

It is important to note, however, that the findings presented in this study are tentative and limited. The small sample size in this study makes it difficult to draw any definitive conclusions based on the data. The sample was restricted by both the methodology employed, which required consecutive years of student assessments and was thus limited to teachers of reading and math in the elementary grades, and the design of the evaluation system, which did not call for all teachers to be evaluated annually. Moreover, many students and several teachers in the district were excluded from the analyses due to missing test scores and other data. For these reasons, among others, findings and conclusions should be accepted with caution and limited to this site.

## References

- Coventry Public Schools and Coventry Teachers Alliance, (2003). *Teacher Evaluation Handbook, Administrator's Guide: Performance Evaluation of Certified Staff*. Coventry, RI: Coventry Public Schools and Coventry Teachers Alliance.
- Danielson, C. (1996). *Enhancing Professional Practice: A Framework for Teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Danielson C., and McGreal, T. L. (2000). *Teacher Evaluation to Enhance Professional Practice*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Gallagher, H. A. (2004) *Vaughn Elementary's Innovative Teacher Evaluation System: Are Teacher Evaluation Scores Related to Growth in Student Achievement?* To be published in the *Peabody Journal of Education*, Spring, 2004.
- Kellor, E., Milanowski, A., Odden, A., and Gallagher, H. A., (2001) *How Vaughn Next Century Learning Center Developed a Knowledge- and Skill-Pay Program*. Madison, WI: University of Wisconsin, Wisconsin Center for Education Research, Consortium for Policy Research in Education.
- Kellor, E. (2003). *Catching Up With the Vaughn Express: Four Years of Performance Pay and Standards-Based Evaluation*. Madison, WI: University of Wisconsin, Wisconsin Center for Education Research, Consortium for Policy Research in Education.
- Kimball, S. (2002). *Washoe County Teacher Performance Evaluation System: A Case Study*. Madison, WI: University of Wisconsin, Wisconsin Center for Education Research, Consortium for Policy Research in Education.
- Kimball, S., White, B., Milanowski, A., and Borman, G. (2004). *Examining the Relationship between Teacher Evaluation and Student Assessment Results in Washoe County*. To be published in the *Peabody Journal of Education*, Spring, 2004.
- Milanowski, A. (2004). *The Relationship between Teacher Performance Evaluation Scores and Student Achievement: Evidence from Cincinnati*. To be published in the *Peabody Journal of Education*, Spring, 2004.
- Odden, A., Archibald, S., Milanowski, A., and Conti, E. (2001). *A Case Study of the Implementation of a Knowledge and Skill-Based Pay System: Coventry, Rhode Island*. Madison, WI: University of Wisconsin, Wisconsin Center for Education Research, Consortium for Policy Research in Education.
- Odden, A. and Kellor, E. (2000). *How Cincinnati Developed a Knowledge- and Skills-Based Salary Schedule*. Madison, WI: University of Wisconsin, Wisconsin Center for Education Research, Consortium for Policy Research in Education.

## **APPENDIX I: Revised Coventry Teaching Standards (2003)**

### Domain 1: Planning and Preparation

Component 1a: Demonstrating knowledge of content and pedagogy

Component 1b: Demonstrating knowledge of students

Component 1c: Selecting instructional goals

Component 1d: Demonstrating knowledge of resources

Component 1e: Designing coherent instruction

Component 1f: Assessing student learning

### Domain 2: Instruction

Component 2a: Communicating clearly and accurately

Component 2b: Using questioning and discussion techniques

Component 2c: Engaging students in learning

Component 2d: Providing feedback to students

Component 2e: Demonstrating flexibility and responsiveness

Component 2f: Establishing a culture for learning

### Domain 3: Professional Responsibilities

Component 3a: Organizing and managing the classroom

Component 3b: Reflecting on teaching

Component 3c: Communicating with families

Component 3d: Contributing to the school and district

Component 3e: Growing and developing professionally

Component 3f: Showing professionalism