



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

## Relationships Among Dimension Scores of Standards-Based Teacher Evaluation Systems, and the Stability of Evaluation Score – Student Achievement Relationships Over Time

Anthony Milanowski  
Consortium for Policy Research in Education  
University of Wisconsin-Madison  
Madison, WI 53706  
(608) 262-9872

April, 2004  
CPRE-UW Working Paper Series  
TC-04-02

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 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

This paper explores additional issues relevant to the use and validity of standards-based teacher evaluation scores, using data from evaluation systems in the Cincinnati Public Schools and the Vaughn Next Century Learning Center. First, the interrelationships among the scores on the various dimensions of the teacher evaluation systems was analyzed. Second, the strength of the relationship between each dimension score and student achievement was compared across dimensions. Third, the relationship between teacher evaluation scores and average student achievement in the year after the teacher was evaluated was assessed. Dimension scores were moderately correlated in Cincinnati, and highly correlated at Vaughn. In Cincinnati, the correlations of dimension scores with student achievement were fairly similar, though instruction scores had a slightly weaker relationship, overall, than classroom management or planning. At Vaughn, subject-specific instruction scores were the best predictors of reading and math achievement, but the literacy dimension score was as good as the math dimension score in predicting math achievement. At both sites, the evaluation scores from one year did not predict subsequent year student achievement as well as they predicted achievement in the current year. The implications of these results for the use of evaluation scores in understanding teacher effects on student learning, as well as for administrative decisions, are discussed.

This paper explores two additional aspects of the standards-based teacher evaluation scores analyzed in the companion paper by Milanowski, Kimball, and White. First, I look at the relationship among the scores on the various dimensions of the teacher evaluation systems, and whether some dimension scores show a stronger relationship with student achievement than others. Then, I look at whether the teacher evaluation scores predict classroom average student achievement in the year after the teacher was evaluated. These questions are important to researchers seeking to understand teacher effects on student learning, as well as to practitioners intending to use the evaluation systems and their results for administrative decisions.

For researchers, the relationship among the evaluation dimensions (or domains), differences in the strength of dimension ratings with student achievement, and whether an evaluation score from one year is related to student achievement in subsequent years are relevant to understanding how teacher evaluation ratings represent the construct of teacher performance. The evaluation standards can be considered a model of teacher performance, and the scores as measures of how well a teacher's practice fits the model. To justify interpreting these scores as valid measures of how teachers perform in relationship to the model, there must be evidence of their construct validity. An analysis of the relationships among the dimension scores is one bit of construct validity evidence. We would expect that, because the concepts underlying the dimensions are related and because successful instruction requires a certain minimal level on all of the dimensions, the dimension ratings would be moderately correlated, and they would each have a substantial positive correlation with student achievement. However, we would not expect that the correlations among dimensions to be extremely large, if teacher performance is truly multi-dimensional, as these evaluation systems postulate. We might also expect that ratings on some dimensions might be more strongly related to student achievement than others. With respect to the systems based on the Framework for Teaching (Danielson, 1996) we might expect that scores on the domain that attempts to capture classroom instruction might be expected to have a stronger relationship to student achievement than the domain of professionalism, which includes many activities that are primarily supportive in nature. We might also expect that teacher evaluation scores from one year would

be at least somewhat predictive of the average achievement of subsequent classes of students. If we conceptualize teacher performance as being a function of skill, motivation, and context (Campbell, 1994; Rowan, Chiang, and Miller, 1997), and postulate that skill is a relatively stable trait, then if teacher evaluation scores correlate with classroom student achievement in the year of evaluation, we would expect that these evaluation scores would also correlate with student achievement in later years. If they do not, we might suspect that the correlations we found for the year of evaluation were an artifact of that year's tests, a reflection of special effort the teacher expended that year, or of contextual factors, rather than representing a stable characteristic of the teacher.

For those concerned with the administration of teacher evaluation systems, the question of how closely the dimensions correlate is important for two reasons. First, a very high level of intercorrelation can be a warning that evaluators are not distinguishing between performance on each dimension in making their decisions. If ratings that reflect differences in performance across dimensions are expected and desired (for example, for developmental purposes), raters may need additional training in avoiding 'halo' error (Woehr and Huffcutt, 1994; Cooper, 1981), rating scales may need to be modified to sharpen the distinctions among dimensions, or additional sources of evidence more specific to each dimension may need to be identified. Second, standards-based evaluation systems are expensive to operate. If the level of teacher performance does not differ across dimensions, systems could be simplified by dropping redundant dimensions and time taken to rate each teacher on redundant dimensions could be reallocated to more productive uses.

The question of whether evaluation scores are related to student achievement in subsequent years is also important to designers of teacher evaluation systems. Since it is costly to intensively evaluate teachers every year, and is stressful to teachers, it would be advantageous if a complete evaluation could be done only once every few years. This is just what Cincinnati and Washoe do for tenured teachers. But if teacher evaluation scores are correlated substantially with classroom student achievement only in the year of evaluation, it would be problematic to assume that teachers who received a positive evaluation were those whose students were performing better in the years when no evaluation was done. There

would then be less justification for any actions taken by the district based on evaluation scores that had long term consequences. For example, Cincinnati's original proposal to increase teacher pay for a five year period based on evaluation scores obtained during the comprehensive evaluation year would be less justified if higher rated (and higher paid) teachers did not help to produce more student achievement for the full period.

### **Method**

To address these questions, data from the Cincinnati and Vaughn sites were used. Dimension evaluation scores were available for Cincinnati for the 00-01, 01-02, and 02-03 school years, and for Vaughn from the 00-01 and 01-02 school years. The origin and content of these scores were described in the companion paper by Milanowski, Kimball, and White. Average student achievement was represented by the Empirical Bayes (EB) intercept residuals from a random intercept model of student test scores, with controls for prior year test scores in the same subject and student characteristics, as described in the Milanowski, Kimball, and White paper. These residuals represent an estimate of the relative average 'value-added' by the teacher. These data were available for the 01-02 and 02-03 school years for Cincinnati, and the 01-02 school year for Vaughn. Most of the analyses involved bivariate correlations between scores, and for the Cincinnati data, combination of correlations across grades as described in the companion paper by Milanowski, Kimball, and White.

### **Results**

#### Dimension Interrelationships

Table 1 shows the correlations between ratings on the four performance domains for Cincinnati teachers evaluated in 2001-02 and 2002-03.

Table 1  
Correlations Among Domain Ratings, All Cincinnati Teachers Evaluated in 2000-01 and 01-02

Domain	Planning	Classroom Management	Instruction	Professionalism
Planning	-	.56	.56	.77
Classroom Management	.49	-	.68	.54
Instruction	.52	.61	-	.56
Professionalism	.75	.43	.54	-

Upper triangle: teachers evaluated in 2001-02, N=335. Lower triangle, teachers evaluated in 2002-03, N=318.

The correlations are moderate, except for that between the Professionalism and Planning domains. The pattern of correlation is likely influenced by the design feature of the Cincinnati system that has school administrators rate on the Planning and Professionalism domains, while the Teacher Evaluators from outside the school and the administrator both rate the Classroom Management and Instruction domains. The average correlation between ratings on the Planning and Professionalism domains (.76) is higher than the average correlation between the scores on the Classroom Management and Instruction domains (.65) and substantially higher than across the domains rated by different types of raters (.53). Exploratory factor analyses with orthogonal rotations (results not shown) show two factors: one for the dimensions rated by the building administrators and one for those rated by the teacher evaluators, though all domain scores load to some extent on each factor.

Table 2 shows the correlations between ratings on the five basic evaluation system dimensions for the Vaughn teachers evaluated in 2000-01 and 2001-02.

Table 2  
Correlations Among Domain Level Scores, Vaughn Teachers Evaluated in 2000-01 and 01-02

	1	2	3	4	5
1. Lesson Planning	-	.83	.65	.77	.74
2. Classroom Mgmt	.84	-	.80	.80	.84
3. Literacy	.84	.92	-	.93	.90
4. Language Development	.83	.90	.96	-	.90
5. Math	.84	.86	.95	.96	-

Upper triangle: teachers evaluated in 2000-01, N=34. Lower triangle, teachers evaluated in 2001-02, N=35

At Vaughn, there is a high level of correlation among ratings on the domains for both years. Unlike the case in Cincinnati, here each type of rater (self, administrator, and peer) rates each teacher on each domain. Somewhat unexpectedly, the ratings for the subject specific domains (literacy, language development, and math) are more highly correlated with each other than with the two “generic” domains (lesson planning and classroom management). If different knowledge, skills and techniques are needed to teach these different subjects, one might have expected moderate correlations among the subject-specific domains. What we see is that the subject specific domain ratings have the highest average correlation (.93), the two generic domains have a lower average correlation (.84), and the average intercorrelation of the generic with the specific domains is lowest, but still substantial (at .82). These correlations suggest that a considerable amount of halo may be present and suggest the need to examine the rating process and domain rating scales.

#### Relationship Between Domain Scores and Student Achievement

Since the domain scores are correlated to a considerable degree, it might not be necessary to rate all domains in order to predict which teachers will have students with higher achievement. Also, one might expect that certain domains (i.e. those representing classroom instruction) would be better predictors of classroom student achievement than others. To investigate the feasibility of using fewer dimensions, and the question of whether some dimensions are more strongly associated with student achievement than others, the correlations between domain scores and EB intercept residuals were calculated. The results for Cincinnati, combining across grades, are shown in Table 3.

Table 3  
Domain Level Correlations with Student Achievement (Combined Across Grades) – Cincinnati 3

Domain	Reading	Math	Science
Planning			
01-02	.38	.39	.36
02-03	.26	.33	-.09
Classroom Mgmt			
01-02	.37	.41	.18
02-03	.28	.35	.07
Instruction			
01-02	.33	.35	.14
02-03	.16	.29	-.04
Professionalism			
01-02	.42	.37	.37
02-03	.21	.17	-.13

Considering both years of data, no domain stands out as the best or worst predictor of student achievement. It is somewhat surprising that instruction scores have lower correlations with student achievement in both 01-02 and 020-03 than do classroom management and planning. It is also notable that Professionalism, which had relatively high correlations with student achievement in 01-02, has much lower correlations in 02-03.

Table 4 shows the correlations between domain ratings and student achievement in 2001-02 at Vaughn. As expected due to the high correlations among domain scores, all the domains are reasonably good predictors of classroom student achievement, though for reading and mathematics, the subject-specific domain scores are the best predictors.

Table 4  
Domain Level Correlations with Student Achievement – Vaughn

Domain	Reading	Math	Language Arts
Lesson Planning	.41	.38	.34
Classroom Mgmt	.57	.38	.45
Literacy	.61	.45	.44
Language Development	.57	.34	.38
Math	.62	.45	.41

Interestingly, it would appear that the Literacy evaluation scores alone could suffice as predictors of student achievement in each of the three subjects.

Another way to look at the importance of each domain in predicting student achievement is to consider what predictive power is lost when that domain is dropped from a total or average score combining all the domain scores. The difference between the correlation with student achievement of the composite score containing the domain and that without the domain provides an estimate of the importance of that domain to the predictive power of the composite and an estimate of what that domain ‘adds’ to the composite. Table 5 shows the results of this analysis for the Cincinnati data.

Table 5  
Change in Correlation with Student Achievement Measure When Domain is Dropped From Total Score, Cincinnati

Domain	Change in Correlation		
	Reading	Math	Science
Planning			
01-02	-.01	+.00	-.03
02-03	-.02	-.04	-.02
Classroom Mgmt			
01-02	-.02	-.03	+.02
02-03	-.02	-.04	-.04
Instruction			
01-02	+.01	-.01	+.02
02-03	+.02	-.02	-.00
Professionalism			
01-02	-.00	+.00	-.03
02-03	-.00	+.07	+.04

Note: positive values indicate an increase in the correlation between evaluation scores and student achievement, or a gain in predictive success when dropping the domain score. While negative values indicate a decrease in the correlation when dropping the domain score.

These results show that there is not much change in the strength of the relationship with student achievement if any one domain is dropped from a composite evaluation score. Dropping the Planning or Classroom Management domain appears to result in the most loss of information related to student achievement, while dropping Professionalism seems the least harmful, in terms of predicting student achievement.

A different analysis was conducted with the Vaughn data. Our original expectation was that the ratings on the subject-specific domains would be most strongly related to student achievement, but this

has not proven to be so. Can the correlation between teacher performance and student achievement be strengthened by adding the scores of other domains to make a composite evaluation score? Table 6 shows the correlations between student achievement and the relevant subject-specific domain score, with a composite including that domain and the planning and classroom management scores, and with a composite of all five domain scores.

Table 6  
Correlation Between Student Achievement and Subject-Specific Evaluation Score and Multi-domain Composite Scores, 2001-02

	Reading	Math	Language Arts
Subject Domain Score	.61	.45	.38
Adding Planning & Classroom Mgmt	.55	.43	.41
Composite of All 5 Domains	.58	.42	.42

An average of the five domain scores is about as good a predictor of reading or math student achievement as the subject-specific domain evaluation score, and somewhat better for language arts achievement. The composite formed by adding the lesson planning and classroom management domain scores to the subject specific domain has a slightly stronger relationship with language arts achievement, but for other subjects the relationship is slightly weaker than that between the subject specific domain score and student achievement. These results show that a more complex composite is only marginally better, overall, than the relevant subject specific domain score alone, and is not quite as good as the literacy domain score alone. In terms of predicting average student achievement, it appears that the classroom management and lesson planning scores add little.

Relationship of Evaluation Scores to Later Year Student Achievement

In order to assess the degree to which evaluation scores are related to student achievement in later years, I matched the total evaluation scores of Cincinnati teachers evaluated in 2001-02 with the estimate of average student achievement (the EB residuals from the random intercept models with controls for

student demographic characteristics) of the students they taught in 2002-03. Similarly, I was able to match up total evaluation scores of the grade 3-8 teachers evaluated in 2000-01 with the estimate of average student achievement of the students they taught in 2001-02. The correlation between the evaluation scores and EB residuals was calculated, and combined across grades. Table 7 presents these correlations, and shows the correlations of total evaluation scores with EB residuals from the year of evaluation for the cohort of teachers evaluated in 2001-02 for comparison.

Table 7  
Comparison of Correlations Between Total Teacher Evaluation Scores and Student Achievement in Year Evaluated, and Following Year, Cincinnati

	Reading	Math	Science
Student Achievement from 01-02, Teachers Rated in 01-02	.48	.41	.26
Student Achievement from 02-03, Teachers Rated in 01-02	.19*	.19*	.15*
Student Achievement from 01-02, Teachers Rated in 00-01	.21*	.33	.28*

\* Confidence interval includes 0

The first point to note is that evaluation scores do have a positive relationship with the student achievement of classes taught in the next year, though that relationship is stronger for the cohort rated in 2000-01. For the cohort rated in 2001-02, the evaluation scores are more strongly related to student achievement in the year of evaluation than to the achievement of the students taught in the next year. While some difference in the size of the correlations would be expected, due to the relatively small sample of teachers involved, the different tests used for measuring student achievement, changes in teachers' circumstances, and variation in student cohorts from year to year, it is notable that the correlations with student achievement in the following year are lower than those with achievement in the year of evaluation. The pattern of correlations is consistent with an interpretation that teachers 'let down'

in terms of effort in the years in which they are not comprehensively evaluated.<sup>1</sup> Our field work with Cincinnati teachers does suggest that many saw the comprehensive evaluation as an experience requiring increased effort.

Because teachers are evaluated using the same process every year at Vaughn, there is less likely to be a decrease in effort due to less intensive evaluation. For the Vaughn data, the correlations between the subject specific domain evaluation scores given in 2000-01 and the achievement of students taught in 2001-02 can be computed. For the 28 teachers from 2000-01 teaching at Vaughn in 2001-02, the correlations were .62 for reading, .23 for mathematics, and .21 for language arts. The corresponding correlations between the 2001-02 subject domain ratings for these 28 teachers and 2000-01 student achievement were .76, .43, and .42.<sup>2</sup> Here, even though teachers are evaluated intensively each year, one year's evaluation scores do not seem to predict student achievement for subsequent classes as well as the current year scores do. This evidence suggests that factors other than a decrease in effort should be explored to explain this and the similar pattern of relationship in Cincinnati.

At Vaughn, the evaluation scores themselves are strongly correlated across years. Table 9 shows the correlations between the evaluation scores on each domain from 2000-01 and 2001-02 for the 28 teachers evaluated in both years. Teachers' relative performance, or at least their relative performance rating, was quite stable over time. This is evidence for the validity of the teacher evaluation scores as a measure of an underlying stable performance construct (It should be noted, however, that some of the stability may be due to a halo effect across years; once a teacher has performed well for a period, raters may tend to assume continued good performance, and not look as critically at the teacher's behavior.)

Table 9  
Correlations Between Domain Evaluation Scores Across Years - Vaughn

---

<sup>1</sup> Recall that most teachers are not comprehensively evaluated each year, so in the following year they are likely to receive a much less intensive annual assessment involving one observation by a building administrator, unless the teacher did so poorly on the comprehensive evaluation that another one is scheduled in the following year.

<sup>2</sup> Unfortunately, the student achievement data from 2000-01 was not available to compare these correlations with those between 2000-01 evaluation scores and the achievement of students taught in that year for just these 28 teachers.

Domain	Correlation Within Dimension Across Years	Average Cross-Year Correlation with Other 4 Dimensions
Lesson Planning	.72	.54
Classroom Mgmt	.69	.72
Literacy	.84	.70
Math	.80	.79
Language Development	.83	.75

Reflecting the high degree of intercorrelation among domain scores in each year, the correlations between scores on different domains across years are also relatively high, nearly as high as the within-domain cross-year correlations, providing another bit of evidence that the evaluation system may not be measuring separate dimensions.

### Discussion

The results reported above have implications for the use of standards-based teacher evaluation scores for administrative decisions and as representations of teachers' practice. In Cincinnati, dimension scores were moderately correlated and the highest correlations were between the two dimensions rated by building administrators (the Planning and the Professionalism domains). At Vaughn, scores on all five of the core dimensions were highly correlated, with the highest correlations between the scores on the subject-specific domains related to literacy, mathematics, and language development instruction. In Cincinnati, no domain score had a relationship with student achievement that was clearly stronger than the others, though instruction scores had a slightly weaker relationship, overall, than classroom management or planning. At Vaughn, the subject-specific domain scores were the best predictors of reading and math achievement, but the literacy domain score was actually a better predictor of language arts achievement and as good as the math domain score in predicting math achievement.

From an administrative perspective, these results suggest that, if the primary purpose of the evaluation systems is to identify those teachers who contribute most to facilitating student achievement, then some of the domains could be de-emphasized, simplified, or even dropped from the systems. In

Cincinnati, the professionalism domain seems the least useful in predicting student achievement, while at Vaughn, the literacy evaluation score would appear to suffice. However, while these results do suggest that some simplification of these systems would be justified, it should also be recognized that performance standards serve important communicative and developmental functions. They tell teachers what specific behaviors are desired, and can be used by skilled mentors or coaches to help teachers improve performance in areas of deficiency. New teachers, especially, are unlikely to have mastered every dimension of the complex work of teaching, and may need guidance on some aspects of performance that are automatic for experienced teachers. Therefore, though scores on all the domains are not needed to predict student achievement and provide evidence of the criterion-related validity of the systems, organizations should be cautious about deleting dimensions from these systems, even though doing so would make them easier to administer.

With respect to using evaluation scores to represent teacher practice, the results for Cincinnati present mixed evidence of the construct validity of the evaluation scores as measures of pedagogical practice. Domain scores are correlated, as expected, but not so highly as to cast doubt on the system's ability to differentiate performance across dimensions. However, given the importance attached to instruction in improving student achievement, one might have expected instruction domain scores to have been somewhat more strongly related to student achievement, in comparison with scores on the other domains. The strong relationship between the planning and professionalism domain scores also raises some construct validity issues. Two factors, besides the possibility that teachers who are good planners are also high in professionalism, could be contributing to this correlation. First, as noted above, these two domains are both evaluated by a single rater, a building administrator, based largely on the same source of evidence, the teacher's portfolio. The common rater and common evidence collection method are sources of common method variance, a common contaminant of judgmental measures. Second, a third factor, such as teacher conscientiousness, could be partly responsible for the inter-domain correlation. It is likely that conscientious teachers spend more time and effort on their portfolios, creating a document that shows them to best advantage. And it may be that conscientiousness contributes to student achievement more

than skill in lesson planning or professionalism. Thus a good portfolio could be a sign of a good teacher, but the domain-relevant practice the portfolio is intended to represent may not have a strong causal relationship to student achievement.

At Vaughn, the large domain score intercorrelations suggest a considerable amount of halo error in teacher performance ratings. It is important to recognize that it is possible that teachers who perform well on one domain do so on others as well. As Murphy et al (1993) argued, it is not possible to separate ‘true’ halo (real similarities in performance across dimensions) from halo error (inability or unwillingness of raters to recognize differences in performance level across dimensions) in a field setting. Nevertheless, these high correlations suggest the need for research on evaluator decision making at this site before we can be comfortable that the domain scores accurately represent domain-specific practice as intended. That the literacy domain score is such a good predictor of student achievement in all three subjects is also grounds for recommending additional research. It might be expected that the literacy and language development domain scores would be highly correlated, as are the reading and language arts test scores, due to the conceptual relationships across domains. However, the correlation between the literacy and math domain scores does not seem so natural. Though it may be possible that a good literacy teacher is also a good math teacher, it may also be that raters are focusing more on general pedagogy when evaluating. That is, the raters, who are not subject specialists, may be more aware of differences in general pedagogical practices between teachers than differences in content-specific pedagogy. If so, the Vaughn subject-specific performance scores, though strongly related to student achievement, may not be good representations of subject specific teaching practice.

The paper also explored whether the teacher evaluation scores predicted average achievement for students taught in the year after the teacher was evaluated. At Vaughn, the evaluation scores from one year predicted subsequent year students’ reading achievement in reading rather well, almost as well as evaluation scores from the later year. Subsequent year math and language arts achievement was not as well predicted by the evaluation scores, nor predicted as well as by scores from the year of evaluation. In Cincinnati, a considerably weaker relationship was found. If it turns out that teacher evaluation scores do

not predict the relative average achievement of subsequent classes as well as current classes' achievement in additional years of data and at our Washoe site, the practical significance will be that evaluation may be needed every year if organizations are interested in using teacher evaluation scores as the basis for performance pay. The construct validity implications of these findings are not as clear. There is some stability in the evaluation score-student achievement relationship, as expected if teacher skill is a strong determinant of teacher performance. But scores can also reflect context and motivation. Because Cincinnati does not evaluate rigorously every year, the possibility that teachers lowered effort in the year after evaluation was suggested as a plausible explanation. The Vaughn results tend to cast doubt on this, however, since a similar, though less marked, weakening was found at Vaughn. We need to explore potential changes in motivation and context, as well as in rater decision making and student testing, that might explain the results obtained.

At both sites, further research is planned, especially around the issue of the construct validity of teacher evaluation scores as measures of the degree to which teaching practice conforms to the model of instruction on which the evaluation systems are based. Interview-based research on rater decision processes is planned, and we will also be looking at how domain scores of teachers with differing levels of experience are correlated. Since inexperienced teachers are less likely to have mastered all of the domains, their domain score correlations should be lower than those for experienced teachers, if the system is able to distinguish performance differences across conceptually different domains. We also intend to analyze the relationship between evaluation scores and classroom student achievement in subsequent years, using the data from the Washoe County site. Since Washoe does not evaluate intensively every year, it will be interesting to see if the correlation between evaluation score and the achievement of the subsequent year's students is also lower than the current year relationship. Another analysis will compare average classroom student achievement across years for teachers evaluated and not evaluated. If there were a reduction in effort, one would expect to see a decline in average student achievement for evaluated teachers but not for those not evaluated. The stability of relative average student achievement over time can also provide a context in which to interpret the stability, or lack

thereof, in teacher evaluation scores and the evaluation score – student achievement relationship. We hope that further analyses, and one more year of data, will provide more evidence about the usefulness and validity of these evaluation systems.

## References

- Campbell, J.T. (1994). Alternative models of job performance and their implication for selection and classification. In M.G. Rumsey, C.B. Walker, and J. H. Harris (Eds.) Personnel Selection and Classification. Hillsdale, NJ: Lawrence Erlbaum Associates, 33-51.
- Cooper, W. H. (1981) Ubiquitous halo. Psychological Bulletin, 90(2), 218-244.
- Danielson, C. (1996). Enhancing Professional Practice: A Framework for Teaching. 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.
- Murphy, K.R., Jako, R.A., and Anhalt, R.L. (1993). Nature and consequences of halo error: A critical analysis. Journal of Applied Psychology 78(2), 218-225.
- Rowen, B., Chiang, F, and Miller, R.J. (1997). Using research on employees' performance to study the effects of teachers on student achievement. Sociology of Education, 70, 256-284.
- Woehr, D.J., and Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. Journal of Occupational and Organizational Psychology, 67, 189-205.