# Examining Teacher Evaluation Validity and Leadership Decision Making Within a Standards-Based...

**Article** *in* Educational Administration Quarterly · February 2009

DOI: 10.1177/0013161X08327549

CITATIONS
27

READS
234

**2 authors**, including:

Anthony Milanowski
Westat
**46** PUBLICATIONS **736** CITATIONS

SEE PROFILE

# Examining Teacher Evaluation Validity and Leadership Decision Making Within a Standards-Based Evaluation System

Steven M. Kimball and Anthony Milanowski

*Educational Administration Quarterly* 2009 45: 34
DOI: 10.1177/0013161X08327549

**Educational Administration Quarterly**
Volume 45 Number 1
February 2009 34-70

10.1177/0013161X08327549
http://eaq.sagepub.com
hosted at
http://online.sagepub.com

# Examining Teacher Evaluation Validity and Leadership Decision Making Within a Standards-Based Evaluation System

Steven M. Kimball
Anthony Milanowski

**Purpose:** The article reports on a study of school leader decision making that examined variation in the validity of teacher evaluation ratings in a school district that has implemented a standards-based teacher evaluation system. **Research Methods:** Applying mixed methods, the study used teacher evaluation ratings and value-added student achievement data to identify 23 school leaders with "more" and "less" valid results. These leaders were interviewed to learn about their attitudes on teacher evaluation, their decision-making strategies, and school contexts. Results from interviews with a subset of eight school leaders with 2 years of consistent validity scores ($n = 4$ more valid and $n = 4$ less valid) were analyzed. **Findings:** Substantial variation was found in the relationship of evaluators' ratings of teachers and value-added measures of the average achievement of the teachers' students. The results did not yield a simple explanation for the differences in validity of evaluators' ratings. Instead, evaluators' decisions were found to be a complex and idiosyncratic function of motivation, skill, and context. **Conclusion:** The results suggested why overall validity of ratings was lower than expected and highlighted challenges in research on school leader decision making and cautions for using such decisions for high stakes purposes. Recommendations are provided for improving evaluation accuracy and validity, with considerations for performance evaluation policy and future research on school leader decision making and teacher evaluation.

***Keywords:*** *teacher evaluation; teaching standards; school leadership; decision making; validity; empirical paper*

**Authors' Note:** The research reported in this paper was supported by a grant from the U.S. Department of Education, Office of Educational Research and Improvement, to the Consortium for Policy Research in Education and the Wisconsin Center for Education Research, School of Education, University of Wisconsin-Madison (Grant No. OERI-R308A60003). The opinions expressed are solely those of the authors. The authors acknowledge the contributions of Brad White, who participated in the data collection and initial analysis phases for this study.

Recent federal action encourages experimentation with new compensation systems for teachers, calling for pay determinations to be based in part on evaluations of classroom teaching (U.S. Department of Education, 2006). One premise of this law is that school leaders can identify more effective teachers through performance evaluations. Recent research by Jacob and Lefgren (2006) found that principals' judgments of teacher effectiveness were related to variations in classroom-value-added student achievement in one district. Our own research (Kimball, White, Milanowski, & Borman, 2004; Milanowski, 2004; Milanowski, Kimball, & Odden, 2005) found that evaluators' ratings of teachers using standards-based teacher evaluation systems (Ellett, 1997; Ellett, Annunziata, & Schiavone, 2002; Danielson & McGreal, 2000) can have moderate correlations with classroom average value-added student achievement. These results suggest that school leader evaluations of teachers may in fact have some validity as measures of teacher effectiveness, providing some justification for consequential use of evaluation ratings.

In our research, however, we noted considerable variation in the strength of the evaluation rating-student achievement relationship (different levels of criterion-related validity) across grades and sometimes subjects. Part of this variation is likely because of differences in the characteristics of the tests. But if evaluators differ substantially in the degree to which their ratings correlate with student achievement, teachers could receive consequences that are not justified by the general validity evidence cited above.

This article reports on the results of our exploration of evaluators' decision-making practices and whether these differences are related to differences in the strength of the evaluation rating–student achievement relationship. The study was designed to better understand evaluator decision making to learn whether differences in decision making could help account for the differential validity we have observed in principal evaluations of classroom performance. The results could then lead to recommendations for improving teacher evaluation practices.

Although evaluator decision making and rating validity has been extensively studied in private sector organizations (Landy & Farr, 1980; Murphy & Cleveland, 1995), research on these topics in education organizations are much less common. With increased attention to teacher accountability and the potential use of assessments of teaching performance as a factor in teacher pay, there is a need to learn about how principals make decisions about teacher performance, especially when using increasingly popular standards-based evaluation systems. Differences in validity across principals

are clearly problematic as stakes are raised, and the decision making of those with less valid ratings needs to be improved. Research on principal decision making in teacher evaluation and evaluation validity can provide useful information for evaluator and principal training, teacher quality and instructional improvement, and school accountability. The two questions guiding our study were the following:

1. How much does the validity of the performance rating relationship vary across evaluators?
2. Are differences in evaluator decision making in a standards-based teacher evaluation system related to differences in the strength of the student achievement- performance rating relationship?

## Standards-Based Evaluation

Standards-based teacher evaluation, as exemplified in the work of Ellett and his colleagues (1997, 2002) and of Danielson (1996; Danielson & McGreal, 2000) has been growing in use and could contribute to more valid judgments of teacher effectiveness. It has also formed the foundations for systems designed to measure performance for teacher advancement or pay (Heneman, Milanowski, Kimball, & Odden, 2006). These systems contain public standards and detailed rating scales, which provide guidance to evaluators in making judgments, potentially lowering subjectivity by establishing a common criterion reference for evaluating teacher performance. Standards-based evaluation systems also typically call for more varied sources of evidence about teachers' practice than traditional evaluation approaches and for more extensive training of evaluators, who are typically school principals.

The framework developed by Danielson (1996) represents one commonly used standards-based teacher evaluation approach (Heneman et al., 2006). The evaluation standards consist of 22 components within four domains of teaching practice: planning and preparation (Domain 1), classroom environment (Domain 2), instruction (Domain 3), and professional responsibilities (Domain 4). There are 66 elements that list aspects of performance on the components and domains. A four-level rubric provides a range of performance descriptions from unsatisfactory to distinguished teaching practice. This system was designed to apply to all grade levels and subject areas and to inform both formative and summative decisions related to teaching practice (Danielson & McGreal, 2000).

The *Framework for Teaching* was developed from the knowledge base compiled for the PRAXIS III assessment series (Danielson, 1996), which includes job analysis, reviews of state licensing programs, field work, and reviews of the literature. The literature drawn on includes, but is not limited to, effective teaching (Brophy, 1986) and conceptions of teaching based on pedagogical content knowledge put forth by Shulman (1987). The system is largely classroom observation based but is intended to tap facets beyond the classroom that may impact teaching, including planning, use and modification of instructional materials, working relationships with members of the school community, and professional activities.

Despite the research and experiential basis for the *Framework,* use of such teaching standards for evaluation has been criticized for reducing the complex act of teaching to a simplistic level (Peterson, 2000). Because teaching is socially constructed and variable, systems based on classroom observation "open up or constrict one's view of teaching and given teaching occasions" (Stodolsky, 1990, p. 175). It has been argued that standards-based teacher evaluation suffers from "the practical difficulty of describing teacher performances in terms that are precise, clear, specific, detailed, and understandable" (Peterson, 2000, p. 227). Peterson (2000) also pointed out that such standards-based evaluation systems lacked empirical evidence on their uses and impact. Although subject to criticism, these systems are being used to foster and measure school district visions of instruction and increasingly as part of new teacher compensation systems. Their use merits careful examination.

We have conducted a number of studies on standards-based teacher evaluation systems based on the *Framework for Teaching* (Danielson, 1996) that have found acceptance by teachers and administrators on their uses and that evaluation ratings can have a moderate degree of validity (Kimball et al., 2004; Milanowski, 2004; Milanowski et al., 2005). In these studies, validity was represented by the extent to which evaluation ratings were related to the criterion of value-added measures of student achievement. In some cases, these studies found relationships that were substantially stronger than were found in earlier research on the validity of principal ratings of teacher performance (Medley & Coker, 1987). These findings applied both to systems designed for high-stakes, summative purposes (teacher pay) and lower stakes, formative purposes (professional growth). There were notable differences, however, in the strength of the evaluation rating–student achievement relationship across schools and districts and within organizations by grade and subject. In the course of investigating these differences, it became apparent that even within districts, there was

considerable variation among evaluators in the extent to which their ratings correlated with value-added student achievement. These differences raise questions about the reliability of using teacher evaluation ratings, either to inform professional development or to make high-stakes decisions.

## Conceptual Framework for Understanding Evaluator Decision Making

To guide our exploration of evaluator decision making, we reviewed the literature on performance evaluation to identify potential influences on evaluators that might vary enough to help explain differences in the strength of the relationship between performance ratings and student achievement. This literature has identified three broad classes of such influences (DeCotiis & Petit, 1978; Landy & Farr, 1980). These can be summarized as will (evaluator motivation), skill (evaluator expertise), and the evaluation context (i.e., the school environment).[1] The policy implementation literature has emphasized similar constructs. In studies of policy implementation, policy actions have been described as varying due to people's beliefs, knowledge, and the places in which they operate, as well as the nature of the policy design (Honig, 2006).

Evaluator motivation[2], or will, is likely to affect the strength of the rating criterion relationship in a number ways. First, motivation may affect the degree of leniency of the evaluator. For example, evaluators whose goal is to maintain good relationships with employees or to improve performance may be more lenient, because negative feedback can lead to lower actual performance (Kluger & DeNisi, 1996). Leniency attenuates the relationship between performance ratings and criterion measures; it restricts the range of evaluation scores and reduces discrimination between performance levels at the low end of the rating distribution. The ratings of a more lenient evaluator are thus likely to show a weaker relationship with student achievement.

Evaluators may also differ in the importance they place on distinguishing between individuals as opposed to identifying individuals' strengths and weaknesses. Cleveland, Murphy, and Williams (1989) argued that accuracy in the former may be unrelated to accuracy in the latter. Between-teacher accuracy is important in achieving a strong rating–student achievement relationship because if the levels of the evaluation system are related to student achievement, failure to accurately distinguish among teachers will obscure the connection.

Evaluator attitudes toward the evaluation system can affect evaluator motivation and thus accuracy and validity (Tziner, Murphy, & Cleveland, 2001). A school leader who views the performance evaluation system as too much work or just another mandate is likely to spend less time observing teaching behavior and making careful assessments than one who sees performance evaluation as a tool for instructional improvement.

Evaluator skill in observing and processing information about employee behavior is also likely to influence the performance rating–student achievement relationship. The more skilled the evaluator, the more likely that she will give ratings that accurately reflects how the teacher actually performs on the dimensions defined by the evaluation system. Thus, if there is a relationship between the teacher behaviors specified by the system and student learning, an accurate set of ratings will exhibit a stronger relationship with student achievement than an inaccurate set. A basic factor in evaluation accuracy is the ability to recall and process the information (Bernardin & Cardy, 1982; DeNisi, Cafferty, & Meglino, 1984). Another factor is the evaluator's own knowledge or familiarity with job content. Although there is some evidence that familiarity with job content is associated with more accurate rating (Smither, Barry, & Reilly, 1989), the research is somewhat mixed, and there has been little attention to whether evaluators with experience in performing the job or who have a knowledge base in evaluatee's occupation rate more accurately. This is a potentially important issue in teacher evaluation because school administrators may not have much knowledge or experience with all academic subjects, particularly at the secondary level (Nelson & Sassi, 2005). Evaluator training related to understanding the system, providing a frame of reference for ratings, conducting observations, and decision making has shown a positive effect on accuracy (Bretz, Milkovich, & Read, 1992; Hedge & Kavanagh, 1988; Smith, 1986; Woehr & Huffcutt, 1994).

Numerous context factors may affect the strength of the rating-criterion relationship across evaluators. Studies have identified differences in the opportunity to observe relevant behavior (Freeberg, 1969; Judge & Ferris, 1993), evaluators' perceived incentives for accuracy (Murphy & Cleveland, 1995; Napier & Latham, 1986), and the status or performance of the organizational unit (Murphy & Cleveland, 1995) as potential or actual causes of differences in accuracy. Of particular interest is the effect of the performance of others as a background against which a particular evaluatee's performance is judged. Evaluators tend to rate a moderate level of performance higher if other performers in the group are poor performers and lower if others are good performers (Grey & Kipness, 1976; Ivancevich, 1983;

Klein, 1998). In schools with greater concentrations of lower performing teachers, ratings would likely be inflated, and the district-wide teacher performance–student achievement relationship attenuated.

As identified in the literature on rater cognition, rating accuracy, and policy implementation, the broad categories of will, skill, and context appear important to explore in studies of evaluation decision making because of their logical connection to cognitive processes and enacted behaviors. Within these broad areas, we sought to further our understanding of evaluation decision making and identify possible explanations for validity of evaluation decisions. The results could then be used to help districts focus evaluation training as well as advance the field of teacher evaluation research.

## Method

The research reported here employed a sequential mixed methods design (Tashakkori & Teddlie, 1998) that led with a statistical analysis to determine whether evaluators in one school district using a standards-based evaluation system differed in validity of evaluation ratings and, if so, to identify more valid and less valid evaluators. This sample of evaluators (school administrators) then became the study focus on how evaluators might differ in conducting the teacher evaluation process and deciding on evaluation ratings. A statistical analysis alone would not explain how evaluators enact the evaluation process or why some evaluators provide more or less valid ratings. Therefore, we interviewed evaluators and teachers and collected evaluation documentation to examine potential differences. We used data and method triangulation to check the validity of our assumptions and to explore alternate possibilities and interpretations for evaluation rating–student achievement relationships. The remainder of this section provides information on the teacher evaluation system in the research site, details the statistical analysis that uncovered evaluator variation and yielded our sample for the qualitative study, and then describes the qualitative methods employed.

### Research Site

The study was carried out in a large school district in the western United States. This district educates more than 60,000 students in 88 schools employing about 3,300 teachers. The site was initially selected because it

had more than 3 years of experience with a standards-based teacher evaluation system adapted from Danielson's (1996) *Framework for Teaching* and had student achievement and performance evaluation results for a large number of teachers over several consecutive years. We had been conducting research on the teacher evaluation system in this district for several years as part of a larger study. The criterion-related validity studies we carried out in this district have shown consistent and statistically significant correlations between evaluation results and classroom-average student achievement (Milanowski et al., 2005). The overall correlation between value-added student achievement (in reading and mathematics) and teacher evaluation ratings averaged .22 throughout the 3 years studied. Prior case studies of this district (Kimball, 2001) and districts using similar evaluation systems (Halverson, Kelley, & Kimball, 2004; Kimball, 2002) exposed differences in principals' motivation to be accurate, their knowledge about instruction and the evaluation system, and school context. These differences appeared to influence how principals perceived and carried out their evaluation responsibilities. We were interested in exploring whether differences in motivation, knowledge and skill, and school context explained why some evaluators' ratings of teachers would show a stronger relationship with the achievement of the teachers' students than other evaluators' ratings. The study also sought to uncover how these factors made a difference.

## Teacher Performance Evaluation Measures

The school district implemented a new teacher evaluation system structured on the *Framework for Teaching* (Danielson, 1996) in 2000. The evaluation process was adopted by the district in response to dissatisfaction with the prior, non-standards-based approach and to comply with a state mandate for annual teacher evaluations. The district wanted a system that would represent a common framework for evaluation discussions among school leaders and teachers, promote instructional improvement through formative feedback, and encourage teacher reflection. Evaluation ratings are also used for summative evaluation decisions, such as interventions for substandard performance, contract renewal, and tenure.

Principals are the primary evaluators of teacher performance, but assistant principals also conduct evaluations at large elementary, middle, and high schools. Evaluator training was concentrated on the front end of program implementation, in the 1st and 2nd years. All evaluators were trained on basic aspects of the system, including understanding the performance standards and interpreting the different rubric levels, what procedures

were expected to be followed, and recommended sources of evidence to be applied to the rubrics in making teacher performance judgments. Evaluators were encouraged (but not required) to consider the following evidence sources: teachers' self-assessment, a preobservation data sheet (including a lesson plan), classroom and nonclassroom observations, a reflection form, and instructional artifacts (e.g., assignments and student work, logs of professional activities, and parent contacts). Training did not emphasize interrater consistency. Furthermore, school administrators were not scored for the accuracy of their evaluation ratings or compared to a standard as part of their training. However, one training session did have principals observe actual classroom teaching in small groups, after which a facilitator led follow-up discussions on how the evaluation rubrics could be used to describe the classroom instruction that was observed. In subsequent years, optional training was available to principals on how to manage the evaluation process (i.e., completing evaluations by their due date), and mandatory training was required only for new principals. Training for new principals focused primarily on understanding the procedures of the system and managing the process.

The evaluation procedures are structured around three stages of evaluation: probationary, postprobationary major and postprobationary minor. Probationary teachers (those in their first two years) are evaluated on all four of the performance domains, where they must meet at least Level 1 ("target for growth" or basic) scores on all elements. Probationary teachers are observed at least nine times during the year. Teachers in postprobationary status undergo a major evaluation on two performance domains, one of which is teacher selected. They are formally observed three times throughout the course of the year. In the next 2 years, they receive minor evaluations, focusing on one domain they select and involving at least one observation during each year. During their two minor evaluation years, in lieu of formal classroom observations and related conferences, teachers with 5 years of experience may choose to conduct a self-directed growth activity, referred to as the "Track 2" minor option, which can include action research, peer mentoring, supervising student teachers, or seeking certification by the National Board for Professional Teaching Standards. Regardless of the cycle, if a postprobationary teacher is not evaluated on the instruction domain (Domain 3), evaluators are required to complete a supplemental evaluation form. The supplemental evaluation form includes a composite of elements from the planning and preparation and the instruction domains. Throughout the course of the 3-year cycle, teachers will have evaluation scores on all four domains.

Because most teachers are evaluated on different domains each year depending on their stage in the three-year major-minor evaluation cycle, capturing a large sample of teachers who could be compared on the same standards was fundamental to the study. To maximize the number of teachers who could be compared, we used scores from the district's supplemental evaluation form as the primary measure of teacher performance. The evaluation scores were obtained from the spring of 2002 and the spring of 2003. Like the individual element scores, teachers are rated on each of four composite scores as *unsatisfactory* = 0, *target for growth* = 1, *proficient* = 2, and *area of strength* = 3. The appendix includes descriptions associated with these categories.

We used the simple average of the four composite scores to obtain an overall measure of teacher performance. The average correlation among the composite scores was .72, and the coefficient alpha was .91 for the overall composite score based on teacher evaluations from 2002 to 2003. Despite using the composite measure, some teachers were not included in the analysis because of missing student achievement data or because they were evaluated on Domain 3 of the evaluation system (and thus did not have the supplemental evaluation scores).

## Student Achievement Measures

Because of district testing patterns and data availability, we were restricted to estimating the classroom average student achievement in reading and mathematics for teachers in Grades 3 through 5 for the 2001 to 2002 school year, and Grades 3 through 6 for the 2002 to 2003 school year. In total, 5,683 students and 328 teachers were included in the analysis for 2001 to 2002, and 9,873 students and 569 teachers for the 2002 to 2003 school year.

We derived an estimate of the classroom average achievement of the students of each teacher in the year the teacher was evaluated using the following two-level hierarchical linear model. At the student level (Level 1), the model was the following:

$$\text{Current year test score for student ``i'' in classroom ``j''} = \beta_{0j} + \beta_{1j}\ \text{prior year test score} + \beta_{2j}X_{2,ij} \ldots \beta_{nj}X_{n,ij}\ \text{student characteristics} + R_{ij}$$

where $X_{2,ij} \ldots X_{n,ij}$ represent student characteristics including gender, ethnicity, special education status, and free and reduced price lunch status for "i"

student in the classroom of "j" teacher. All Level 1 predictors were grand-mean centered.

The Level 2 (teacher-level) model was the following:

$$\beta_{0j} = \gamma_{00} + U_{0j.}$$

At Level 2, the $U_{0j}$ represented the teacher (classroom)-specific differences from the average of the group intercepts. The slopes for all Level 1 variables were treated as fixed. From this model, the empirical Bayes (EB) intercept residuals were obtained. The EB intercept residuals are the deviations of the intercepts representing the average achievement of a teacher's students from the average for all teachers. The EB residual is a weighted average of the classroom specific intercept and the average intercept, with weights reflecting the reliability of each. These reliabilities are a function of the variance within and between classrooms and the number of students (Snijders & Bosker, 1999). These residuals were taken as the measure of the average student performance relevant to each teacher. Given the grand mean centering, the EB intercept residuals represent the difference for the average student in prior year test score and other characteristics at Level 1. The residuals were obtained by grade and subject (reading and mathematics). To allow comparison across evaluators, we converted the value-added measures into *z* scores within grades, averaged *z* scores for reading and mathematics across teachers within grades then combined the data across grades into a single data set. This data set was used to calculate the overall correlation between evaluation scores and student achievement and the correlation between the ratings of individual evaluators and the average classroom student achievement of the evaluated teachers.

## Assessing Differences in Evaluators' Rating–Student Achievement Correlations

We calculated the Pearson and Kendall correlations between the average of the four performance component scores (described above) and the EB residuals representing value-added classroom student achievement for evaluators who rated at least five teachers. Using the 2001 to 2002 data, there were 39 evaluators who evaluated five or more teachers. Thirty evaluators were excluded by this criterion. Using the 2002 to 2003 data, there were 57 evaluators who evaluated five or more teachers, with 12 excluded.

## Exploration of Evaluator Decision Making

The correlations we calculated were used to identify evaluators with more and less valid ratings (i.e., those whose ratings had strong versus weak relationships with student achievement) as the main sample for our study. Based on the analysis of the 2001 to 2002 teacher evaluation and student test score data, evaluators were categorized as having high, average, or low (including negative or zero) correlations between their ratings and the average student achievement of the teachers they rated. This classification was based on the comparison of each evaluator's correlation with the correlation obtained from all evaluators with comparable results ($r = .22$ for 2001 to 2002) and the distribution of individual correlations. We looked for evaluators with correlations substantially above or below the group correlation. A sample of 23 evaluators with high and low correlations (21 principals and 2 assistant principals) was selected. These evaluators were sorted into two groups: Group A included 11 with more valid evaluation results (average $r = .68$), and Group B had 12 with less valid evaluation results (average $r = –.37$).

The evaluators were interviewed in the fall of 2003. A semistructured interview protocol was developed to explore issues of evaluator will, skill, and context. It included questions about evaluator background, evaluation goals, evaluation evidence, how the evaluation task was carried out, decision processes and considerations using the evidence and rubrics, and school climate. Three teachers in most of the evaluators' schools were selected at random and also interviewed. Teachers were not interviewed from the schools of five evaluators who had retired or moved to other positions in the district.

Transcripts and interview notes were coded, and content analyzed in several stages. Three researchers individually reviewed detailed interview notes from all principal and teacher interviews to identify themes across evaluators and schools and begin establishing a coding structure within the will, skill and context categories. Recognizing potential bias given the prior identification of evaluators, the notes and interview transcripts were coded blindly, without regard for whether the respondent was in the more or less valid group. The response categories were then compiled in a matrix for further analysis, allowing a search for patterns and differences in responses across all evaluators. The qualitative software package NVivo was used to manage the data during the analysis. Finally, transcripts and interview notes were divided into the more and less valid groups and analyzed to gain a more holistic perspective on possible differences or similarities in evaluation decision making within each group.

**Table 1**
**Distribution of Correlations Between Evaluator Ratings and Student Achievement Measure for Evaluators Evaluating Five or More Teachers**

| Correlation Interval | Number of Evaluators in Interval, 2001-2002 | Number of Evaluators in Interval, 2002-2003 |
|---|---|---|
| Below –.90 | 0 | 0 |
| –.90 to –.70 | 4 | 1 |
| –.69 to –.50 | 1 | 3 |
| –.49 to –.30 | 2 | 3 |
| –.29 to –.10 | 4 | 11 |
| –.09 to +.10 | 4 | 14 |
| +.11 to +.30 | 4 | 7 |
| +.31 to + .50 | 4 | 7 |
| + .51 to + 70 | 5 | 9 |
| + .71 to + .90 | 8 | 2 |
| Above +.90 | 1 | 0 |
| No Correlation (All teachers received same rating) | 2 | 2 |

Because we believed that individual evaluator correlations based on 1 year of data and relatively small numbers of teachers per evaluator were of questionable reliability, we used the 2002 to 2003 teacher ratings and student achievement data, which became available in the winter of 2003 (following the qualitative field work), to calculate a second set of correlations for the 23 evaluators. We then looked for evaluators who had high or low correlations in both years. Eight evaluators met this criterion: four with high correlations (average $r = .55$) and four with low correlations (average $r = -.28$). We reanalyzed the interview transcripts for these evaluators to learn whether new patterns or themes emerged or if the existing trends held. In addition, we examined all written evaluations these evaluators provided to teachers to obtain additional insight into their evaluation decision making and to triangulate findings with evaluator and teacher responses.

# Results

## Rating–Student Achievement Correlations

Table 1 shows the distribution of correlations for both the 2001 to 2002 and 2002 to 2003 data for evaluators rating five or more teachers. These

correlations can be compared to the overall correlation between evaluation scores and value-added student achievement of .22 in 2001 to 2002 and .19 in 2002 to 2003.

As reflected in the table, there is a substantial degree of variation across evaluators in the strength and direction of the relationship between teacher performance ratings and the achievement of those teachers' students. For the 2001 to 2002 school year, 11 of the 39 evaluators (28%) fall into the low validity category, with correlations below –.10. These evaluators have correlations .30 or more below the group average. On the other hand, 17 evaluators (44%) have correlations of .31 or higher, considerably higher than the average. For the 2002 to 2003 school year, during which student achievement data for more teachers was available, of the 59 evaluators there are 18 (31%) with correlations below –.10, and 25 (42%) with correlations of .31 or higher. The results imply that an overall estimate of the criterion-related validity of evaluation ratings may mask substantial differences in the validity of ratings produced by different evaluators and justified an in-depth analysis of evaluator decision making. Because most evaluators evaluated relatively few teachers and because the Pearson correlation coefficient is sensitive to outliers, we also calculated a rank order coefficient of association, Kendall's tau. The values of this statistic ranges from –.84 to +.89. The distribution of Kendall correlations (not shown) is quite similar and typically these correlations identify the same evaluators in the more or less valid categories.

These results confirmed our suspicions that there was considerable variation among evaluators in the relationship between their ratings and student achievement in the classrooms of the teachers they evaluated. We were also surprised by the number of negative correlations observed. Of course, it is important to remember that the limitations in the coverage of student testing to the elementary grades restricted the number of teachers for the evaluator correlations. Thus, the observed correlations and their differences may not be totally reliable as an indicator of evaluator quality, or differences in evaluator decision processes. Nevertheless, they do illustrate that in a specific set of evaluation ratings, the degree of association between the ratings of particular evaluators and a criterion can vary substantially.

## Analysis of Group Differences in Evaluation Decision Making

As discussed above, we used the conceptions of will, skill, and context to analyze the interview data and evaluation documents relevant to the four evaluators with more valid ratings and the four evaluators with less valid

ratings in both years. It became apparent that these categories were not enough to fully explain the decision-making process or provide consistent reasons for evaluator decision-making validity in our two samples. Our search for other themes and patterns provided some additional insight into rating variation. The findings relating to will, skill, and context are first represented in Tables 2, 3, and 4. The tables and their descriptions demonstrate the complexities of evaluation decision making and deficiencies of simple explanations based on these features. We then present other findings that help illuminate evaluation decision making.

*Evaluator will.* Motivation or will to conduct teacher evaluations was illustrated through evaluator descriptions of several aspects of the evaluation process, including the evaluation standards, purposes and goals, procedures, outcomes, and written evaluations. As Table 2 represents, there was no clear differences between the more and less valid groups.

Evaluation goals included both accountability and teacher development in each group, and development goals received primary emphasis in each. Accuracy and evaluation score validity were not typically identified as primary goals. This finding is consistent with the results of a 2002 district school administrator survey (not reported here) conducted by the authors.

Evaluator attitudes about the system's content and processes did not vary much between the groups. Most evaluators in each group were positive about the system, seeing it as tool for teacher development. Only one evaluator in the less valid group did not see the system as useful. One might expect that evaluators in the less valid group would have more negative attitudes about the system and it procedures, which would reduce their motivation to exert effort to follow the evaluation process. There is no clear difference, however, between the groups. With the exception of one evaluator in the less valid group (B3), attitudes were generally positive. The act of evaluating teachers and completing all the paperwork was described as a complex and time-consuming demand, yet despite the workload, none expressed strong opposition to the teacher evaluation process. As one principal commented,

> I think it is definitely a tool [for instructional leadership]. It is a guide and actually…made it a lot easier for me to have the rubric to follow when there is nothing left unsaid or they know exactly what it takes to get to Level 3 or Level 2 and I really like that" (Principal B4).

There were no differences in the reported compliance with the evaluation process.

*(text continues on page 56)*

## Table 2
## Evaluator Will

| Theme | | Evaluator | | | |
|---|---|---|---|---|---|
| Evaluation Goals | "More valid" | A1. Accountability; encourage and suggest changes to teachers; student awareness of learning goals | A2. Provide meaningful, nonthreatening feedback; validate teaching | A3. Help teachers develop goals aligned with school and district goals; monitor professional development use in classrooms | A4. Help teachers; get goals across to students; teach to the "whole" child, not for the sake of standards, per se |
| | "Less valid" | B1. Do what is best for students; replicate best teaching practices | B2. Help teachers reflect on and improve teaching; identify and assist low performers | B3. Monitor teachers use of multiple methods, attention to low- and high-performing students | B4. Help teachers improve themselves across range of expertise; ultimately improve student achievement |
| Attitudes about evaluation system and process | "More valid" | A1. Generally positive; evaluation is mandated; tries to bring in authentic observations thru walkthroughs | A2. Positive; uses system to improve teaching; values rubrics and weaves evaluation with school goals | A3. Positive; uses system for teacher professional growth and ongoing dialog about teaching and learning | A4. Generally positive; would rather keep process informal; likes dialogue with teachers but not writing up evaluations |
| | "Less Valid" | B1.Positive; works evaluation into leadership; puts effort into feedback to staff; provides praise to teachers | B2. Positive about system and rubrics; appreciates district flexibility in using system | B3. Negative about system and utility of teacher evaluation; rating system not helpful, no consequences for teachers; prefers Track II option | B4. Positive about system and evaluation as a leadership tool; likes specificity of rubrics |

*(continued)*

**Table 2 (continued)**

| Theme | | Evaluator | | | |
|---|---|---|---|---|---|
| Accuracy definition and incentives to be accurate | "More Valid" | A1. When evaluator is unbiased, has good feel for curriculum and effective teaching, and teachers aware of expectations<br>No incentive; it's unclear what happens after evaluation completed | A2. When based on long-term exposure to teachers<br>Can lead to changed instruction | A3. When there is mutual goal agreement with teacher, clear expectations and timeline, and it is based on observed data<br>Motivated to be accurate and also enjoys going into classroom, which matters for credibility and communication | A4. When based on teacher activities and behaviors in and out of classroom; must be honest and complete<br>Accurate evaluation doesn't make a difference; incentive is to be honest with people about their practice |
| | "Less valid" | B1. When there are multiple observations, both formal and informal, and teachers are clear about what evidence is to be collected<br>Matter of basic integrity and trust; no consequences or district oversight | B2. When rubrics are used and compared to observed behaviors, relate findings to teachers so they agree.<br><br>Accuracy is important for useful feedback; also for integrity and credibility with staff | B3. When evaluations are tied to rubric criteria; clarity on different performance levels<br><br>Ethical responsibility to be unbiased and fair; no consequences for teacher unless there is a real problem | B4. When based on frequent observations, with feedback tailored to teacher needs and goals and tied to evaluation system<br>It is morally correct to be accurate |

*(continued)*

## Table 3
## Evaluator Skill

| Theme | | Evaluator | | | |
|---|---|---|---|---|---|
| Background most helpful | "More valid" | A1. District training some help, but questions own evaluation skills | A2. District training; adaptation and use of evaluation tools; interaction with colleagues | A3. Experience teaching and as Title I coordinator; professional development on effective instruction | A4. District training; being a people person, ability to listen and work well with people |
| | "Less valid" | B1. Business education and business sense; walkthrough training | B2. Teaching and business experience | B3. Teaching experience | B4. Teaching and administrator experience |
| How prepare and tools applied | "More valid" | A1. Prepare by knowing teachers' goals and what's needed to meet goals<br>No specific tools described | A2. Prepare by goal setting and review with teachers, goal revisions after walkthroughs<br>Adapted forms to guide formal and informal observations; uses laptop during observations to take notes | A3. Prepare by reviewing prior evaluations, establish teacher comfort level with informal walkthroughs, meet with teachers to set goals and expectations, and provide evaluation forms<br>Keeping Danielson book at hand as reference; takes rubrics into observations | A4.Took classes and talked with peers about how they conduct evaluations; views evaluation as ongoing part of work with teachers; does not take many notes<br>No specific tools described |

*(continued)*

**Table 3 (continued)**

| Theme | | Evaluator | | | |
|---|---|---|---|---|---|
| | "Less Valid" | B1. Reviews teacher goals and domains, has teacher complete self-evaluation, and sets schedule and timelines for evaluations<br>Uses rubrics for relevant domain when observing | B2. Has teachers fill out goal sheet with domains of focus and meets with teachers to review goals<br>Brings rubrics to reference during observation note taking | B3. Not much preparation; has all non-probationary teachers do Track II evaluations that do not include regular class observations<br>No tools described | B4. Shares expectations for class observations (e.g., lesson plans), hands out sheet with "look-fors" (e.g., student engagement, student work displayed) for walkthroughs<br>Uses observation forms and laptop for note taking and record keeping |
| Evidence | "More Valid" | A1. Formal and informal classroom observations with detailed scripting of lessons; some evidence is domain dependent; for domain 4 will look at professional development and how applied in classroom; if domain 3 will collect work samples (e.g., lesson plans, student work) | A2. Formal and informal classroom observations with detailed scripting of lessons; includes frequent walkthroughs; views lesson plans; focus on teaching to objectives, effective instruction, time on task, rapport with students | A3. Observes series of lessons throughout year; specific evidence depends on goals; looks at class management, transitions, materials, use of space, student writing, feel of class, student-teacher dialog, differentiated instruction | A4. Frequent formal and informal observations; wants complete picture of teacher from staff meetings, parent conferences, walkthrough observations, and child study meetings; does not take notes |

*(continued)*

**Table 3 (continued)**

| Theme | | Evaluator | | | |
|---|---|---|---|---|---|
| | "Less Valid" | B1. Depends on lesson; will script classroom observations, also collects student work. teacher grade book and attendance folder, teacher self-evaluation, and professional development records | B2. Formal and informal observations; collects evidence on student engagement, teaching to standards, checking for understanding, assessing prior student knowledge, and teacher behaviors in and out of classrooms | B3. Primarily uses walkthroughs; for probation teachers, collect evidence on planning to standards, mapping curriculum, feedback from students and parents; avoids formal observations for postprobation teachers | B4. Depends on domain; classroom observations for Domains 1 to 3; Domain 4 requires more evidence; also uses daily walkthroughs, attends all grade-level meetings, and meetings for school reading program |
| Scoring decisions | "More Valid" | A1. Compare evidence to standards and rubrics, then ask for more evidence to confirm or refute initial judgments; scores must match performance; may adjust score if marginal or if teacher presents additional evidence | A2. Continually references rubrics while assessing evidence; rates initially while observing; reviews all evidence, then writes narrative and final scores; questions self on evidence for score; makes sure has justification for lower scores; may | A3. Active use of rubrics during observations and scoring; looks at evidence related to teacher goals; brings evidence from formal and informal observations to avoid "dog and pony," starts writing some evaluations during observations but | A4. Based on recollection of observations and interactions with teachers, but scores according to rubrics; does not collect artifacts; teachers get benefit of the doubt unless not doing their jobs; will change score if teacher presents new, convincing evidence on something missed |

*(continued)*

**Table 3 (continued)**

| Theme | | Evaluator | | | |
|---|---|---|---|---|---|
| | | | occasionally give teachers benefit of doubt, but also not afraid to confront them on scores | finalizes score after rechecking evidence | |
| | "Less valid" | B1. Starts with "blank slate;" uses teacher self-evaluation to assist; rely on documentation with as much data as needed; may ask for additional evidence from teacher; compares evidence to rubric | B2. Use rubric as guide; applies common sense and "gut" feeling; adjusts some rubric interpretations for classes with high needs students; combines impressions from walkthroughs with formal observations | B3. Applies rubric to evidence; does not spend time on supplemental evaluation form; views supplemental evaluation form as meaningless and marks all proficient | B4. Starts as team process, with discussion of teacher self-evaluation; begins scoring during formal observations; uses ongoing process; may alter score if teacher challenges and presents new evidence |

## Table 4
## Evaluation Context

| | | Evaluator | | | |
|---|---|---|---|---|---|
| Student composition | "More valid"<br>"Less valid" | A1. High SES school<br>B1. Low SES school, high English as Second Language | A2. Low to mid SES school<br>B2. High SES school | A3. High SES school<br>B3. Mid SES school | A4. Mid SES school<br>B4. Low SES school, high English as Second Language |
| Student achievement levels | "More valid"<br>"Less valid" | A1. Middle<br>B1. Low | A2. Middle<br>B2. High | A3. High<br>B3. Middle | A4. Low<br>B4. Low |
| Years as teacher | "More valid"<br>"Less valid" | A1. Seventeen<br>B1. Seven | A2. Seven<br>B2. Nine | A3. Fourteen<br>B3. Twenty-two | A4. Fifteen<br>B4. Five |
| Years as principal or assistant principal | "More valid"<br>"Less valid" | A1. Thirteen<br>B1. One | A2. Five<br>B2. Nine | A3. Nineteen<br>B3. Ten | A4. Thirteen<br>B4. Four |
| Relationship with teachers | "More valid"<br>"Less valid" | A1. Good; only three teachers have left in nine years<br>B1. Good; direct leadership, but has open door policy | A2. Good; some annual teacher turnover (5 per year); considered "starter" school for teachers<br>B2. Good; many teachers recruited by principal; sees all teachers as strong; uses direct leadership, but solicits input | A3. Good; proud of relationship with teachers<br>B3. "Rocky"; high principal turnover at school; teachers mentioned labor grievances, and infrequent classroom visits from principal | A4. Good; friend and mentor; team-based approach; first year at school<br>B4. Good; sees self as equal; more facilitator than a boss; fosters open communication |

NOTE: SES = socioeconomic status.

Many evaluators did, however, circumvent district intentions relating to the use of the supplemental evaluation form, a finding that has implications for the validity of evaluation scores across evaluators. Although the district required evaluators to complete the supplemental evaluation form to assess teachers on instruction and hold them accountable for their performance in this area, most evaluators did not appear to rate the supplemental scores as carefully as the regular evaluation scores. Indeed, several referred to the supplemental evaluation as an added burden and tried to complete the form quickly. One thought the form was meaningless and marked all teachers proficient (B3).

There was also little evidence of differences in attitudes toward accuracy and perceived incentives for being accurate between the groups. Evaluators in both groups tended to equate accuracy with sufficient observations of the teacher. Two evaluators in the less valid group also mentioned following the rubrics, an important aspect of accuracy not mentioned by evaluators in the more valid group. Reported incentives for accuracy were also similar across groups. What is interesting is that evaluators commented that there was little oversight or consequences from the district on teacher evaluations in general. One principal commented that no one knows what happens after the evaluations are submitted to the district office and that it would be good to know that evaluations were taken seriously (A1). When asked whether it made a difference how accurately a principal evaluated teachers, another stated, "No," but went on to explain the following:

> When I started doing it first, the first couple of years, I'm going "gosh, someone is looking over my shoulder"…it probably took me a while to understand…you are doing this for the teacher, you're not doing it for [district administration]. (Principal A4)

These findings do not show that differences in evaluator motivation are an explanation for differences in evaluation validity. With the exception of the supplemental evaluation form, most evaluators in each group appeared to more or less follow the process. We next turn to evaluator skill for potential differences between the two groups.

*Evaluator knowledge and skill*. Table 3 summarizes results relating to knowledge and skill, beginning with evaluator background and evaluator training. These are considered precursors or skill proxies to the actual knowledge and skills employed in the evaluation process, which were not directly observed. To get a sense of how skills were applied to the evaluation

process, however, we did ask evaluators to explain how they prepared for evaluations, the evidence collected, and how they used evidence and the rubrics to decide on scores.

Evaluators in both groups discussed a variety of trainings they attended, both within and outside the district. It is interesting to note that all but one of the evaluators in the more valid group cited district training as helping build their evaluation skills. In the less valid group, only one mentioned district training. Respondents in this group indicated that their experience teaching, in administration, or in business, was most helpful in conducting evaluations. Because we do not have evidence of their effectiveness in these capacities, these findings are not considered strong, but they are interesting and could provide some reason for group differences.

With respect to evaluator preparation, there is again little evidence of consistent differences between the groups. Two in each group mentioned setting or understanding teachers' goals. Only one (in the less valid group) mentioned review of the rubrics. Most of the preparation in each group seems to be focused on preparing teachers for the process. Some evaluators in both groups described using tools to help them focus their evaluations, including forms to guide observations or taking rubrics into observations. Several principals talked about sharing observation or evidence collection forms with their colleagues and adapting them to meet their needs.

Most evaluators in both groups did tap multiple, similar, evidence sources to meet evaluation purposes. The district evaluation system specified at least one formal classroom observation under the minor evaluation, three formal observations for teachers on the major evaluation, and nine observations for probationary teachers. Classroom observations were the primary source of evidence used by evaluators in their decisions on teaching performance, and most used several observations, often going beyond minimal district requirements by collecting evidence from informal classroom visits. Some in both groups also included other interactions with teachers, discussions with peers, and meetings with parents. Two evaluators in the valid group (A2 and A3) have a highly structured process and organization system that they applied to their evaluations. They started the evaluation process early by setting expectations for teachers and gathering comprehensive sources of performance evidence. These evaluators put considerable attention on the evaluation rubrics and used them in a variety of ways. They tended to refer to the rubrics before evaluation discussions, during evaluation observations, taking notes directly on worksheets for the rubrics and domains selected by the teacher, and during the evaluation scoring process. Yet three in the less valid group (B1, B2 and B4) also were

fairly structured in their approach and actively applied the evaluation rubrics to decisions.

With respect to scoring decisions, the evaluators in the more valid group did tend to mention the use of rubrics in a more analytical way. However, three evaluators in each group also mentioned factors that motivated them to adjust scores, in each case likely toward a more lenient assessment. It was also apparent that most evaluators in each group had not consciously thought about how they make evaluation decisions until the interviews conducted for this study. At times, they struggled to systematically articulate their decision processes, evidence sources, or accuracy conceptions. As one principal commented, "I don't think I've ever sat down [to] really reflect or analyze how I really do this" (A3). Another stated that "I think it comes down to your judgment too. I've been a principal for 9 years; I think I know good teaching when I see it" (B2).

Overall, there are few striking systematic differences between the groups on the components of evaluator skill on which we collected data. There are some differences in training and experience, and some tendency for the more valid group to pay more attention to the rubrics. But most in both groups used similar evidence and indicated tendencies to be lenient.

*Evaluation context*. Potentially important context factors include the schools' overall socio-economic status (SES), student achievement levels, school administrator experience, and relationships between school administrators and teachers. These also did not appear as consistent explanations for differences between the two groups. Table 4 includes summaries of these context features.

One might expect that principals would tend to be lenient in their teacher evaluations in schools with more students that face economic or academic challenges. Rather than focusing on the quality of teaching described in the standards and observed in practice, principals might give teachers the benefit of the doubt given the difficult conditions in which they work. Our interview data did not reveal that this was the case. Furthermore, two schools from the less valid group (B1 and B4) did have students from the lowest SES in our sample, but the other two schools from the same group were of higher SES status. In addition, one school in the more valid group had relatively low SES (A2). Three schools in the valid group had higher percentages of proficient students in Grade 3 reading and math, but so did two in the less valid group.

There were no clear differences between the groups in the experience of the evaluators as teachers (nature of assignment and length of teaching) or

as school administrators. It was apparent from the teacher interviews that some evaluators in the less valid group were seen by their teachers as knowledgeable and skilled instructional leaders, so perceived credibility was not a primary factor for differences in our sample.

Furthermore, based on evaluator and teacher interviews, the teacher-evaluator relationships were similar and positive across the groups. There was one notable exception. In the school of principal B3, there was evident strain in the relationships between teachers and the principal. The school had frequent turnover of principals and there was apparent lack of trust among the teachers of the current principal.

*Interaction of will, skill, and context*. Upon finding few distinguishing aspects of will, skill, and context between evaluators in our two groups, we turned to an examination of possible interactions among these aspects that might better explain differences in evaluation decision-making validity. When we looked at the features of will, skill, and context together, we found idiosyncrasies that helped explain why some evaluators had more or less valid evaluation decisions.

In one case (B3), there are indications from the teacher interviews that problems existed in teacher and administrator relationships in the school (context). Also, this evaluator held more negative opinions about the evaluation system (will), conducted less direct observations of teacher performance, and relied more on reports from teachers on their self-directed growth activities (will and skill) and indicated that there was a history of high turnover of principals in the school and that teachers were jaded in their opinions of the school's administration (context). This principal wrote vague narratives with few critical comments, and some teachers received written summaries with identical language. This example presents the clearest indication of the interaction of will, skill, and context potentially influencing evaluation validity.

In another case, an evaluator in the low validity group (B2) had a long-term working relationship with many teachers in her school (context). In fact, the principal had personally recruited and initiated many teachers into the school. As the principal commented,

> My teachers are all very experienced, so I would never walk in thinking they are a [Level] 1. To me it is a matter of, OK, what types of things will I see now that will have me decide between [Level] 2 and [Level] 3 (principal B2).

This principal operated with a predetermined rating in mind and could not conceive of rating a teacher as "basic." It could be that this evaluator was

more lenient than others because of the impression that these teachers were of higher quality because they were personally selected to match the administrator's educational philosophy (will).

Among those in the more valid group, there were two cases where the interaction of will, skill, and context clearly suggested why the evaluators had ratings that were higher in validity. Evaluators A2 and A3 were quite positive about the system (will), gathered extensive evidence and took careful notes of observations (skill), and appeared to foster open working environments with teachers in their schools (context). Yet the same cannot be said for the other two evaluators in this group.

Considering the will, skill, and context framework on a case-by-case basis helps shed more light on the decision-making process and suggests why some evaluators produced ratings that were more or less valid than others. Even though these factors do not clearly explain why the evaluators as a group had more or less valid ratings, they do help highlight potential problems with using ratings of single evaluators (in this case, school principals and assistant principals) for high-stakes purposes.

## Other Findings

*Formative focus and leniency.* Evaluators in both groups reported multiple goals for their evaluations and emphasized using the system for formative purposes related to teacher growth rather than summative assessments of teaching performance. Among the goals, most discussed trying to help teachers improve, foster teacher reflection, acknowledge the hard job of teaching, or give teachers some specific feedback. These dimensions are illustrated in the following quote:

> When I evaluate teachers, my goal is to observe, to be that second set of eyes, that can really provide feedback in a meaningful way, hopefully not punitive, threatening, but someplace where they feel safe enough, and trust me enough that I really am trying to help them be the best that they can be. But also to recognize the things that are going well, because I think there are a lot of great things happening everyday in those classrooms and teachers can go for days at a time without someone recognizing that (A2).

As another indication of the formative focus, evaluators allowed considerable input from teachers on the focus of evaluations and on the evidence that would help demonstrate evaluation goal completion (A3, A4, B1, B4). This finding is also illustrated in the analysis of written evaluations and teacher comments about evaluation feedback. The written evaluations focus

primarily on praise, with minimal description of decision rationale or feedback for teacher improvement. For example, the following is the entire written report for one teacher of Principal A4 and typifies the evaluators' written narratives:

> [Teacher name] is a hard working and dedicated teacher. She has a pleasant mannerism in the classroom that helps take the pressure off her students. During walkthroughs and observations, I found [her] classroom a place where learning and hard work is part of the norm. [She] has worked with our [literacy] mentor to help her with her knowledge and teaching techniques. This has been a big help for [her]. This helps her students become better readers and better students. During teacher visitation, [she] has put the information she has observed into her bag of teaching skills to assist her in becoming a better teacher. [She] is a very effective classroom teacher and staff member. Her students benefit from her hard work.

This teacher did receive four "Level 2" ratings, which presumably could be raised to the "Level 3" category through feedback and support, but no suggestions for improvement were provided. In this case, although the evaluators' ratings were accurate according to our analysis, the written feedback provided was not specific. During interviews, several teachers expressed concerns about the nature and depth of instructional feedback, which often focused on affirmation and encouragement rather than constructive criticisms or recommendations on specific instructional strategies. Principals confirmed that they wrote evaluation summaries using careful language and with the understanding that their words would go into the teachers' permanent record.

Evaluators did not suggest that producing an accurate assessment of performance for the purpose of differentiating among teachers was their primary goal. Accuracy was considered important more to help foster useful feedback or to maintain credibility with teachers. Evaluators also saw little consequences from the evaluation system and tended to use a more formal, accuracy-focused approach only for the weakest teachers. In such situations, they expressed the need to clearly document all evidence and justifications for ratings. Such actions were quite rare; some indicated that they saw no reason to give a teacher an unsatisfactory rating on any element and considered most of their teachers to be strong educators.

Evaluators in both groups tended to give teachers leeway in the evidence gathered or in situations where teachers contested their scores. They were either more concerned about providing suggestions for improvement or, in one case (B1), used teacher input and teacher self-evaluations to guide their

final scores. Although this could be because of a greater emphasis on formative aspects of performance evaluation, it could also indicate problems in the school climate, a desire not to complicate working relationships by providing lower scores than teachers expected, or by the evaluator's bias in favor of teachers. The tendency to be lenient can certainly impact validity results across all evaluators.

*Evaluation complexity.* Our research exposed the complexities principals perceived in making evaluation decisions. For example, one principal spoke of the potential ambiguities in evaluating teaching and competing demands (i.e., accuracy verses promoting performance) that may influence evaluation accuracy and illustrates one reason why evaluators tended to inflate ratings:

> But what if…my evaluation is glowing of that person and I hadn't spent time going through [their] previous evaluations. So, I [then] go through them and there are some rotten evaluations…Am I way off base, or have I just given someone the key to success and confidence that they've never even had before. I mean it's all just so, I don't want to say that it's arbitrary, I know we're looking for accuracy, but sometimes that can happen. And it can happen with students too and the teacher that never gave them a shot, never found that thing they could get [excited about]. But I do spend time with the other evaluations because I do want to know whether I'm way off base or that other people have evaluated this person in another way (Principal A2).

Clearly, evaluators in our study have not sought to construct their use of the evaluation process to yield valid results, at least in the sense of relations of the results to measures of student achievement. Rather, they have constructed their own meaning by adapting an evaluation process that is acceptable to them, their teachers, and their school environments, in the context of flexible district guidelines.

## Discussion

The analysis of evaluator will, skill, and evaluation context provided insight into evaluation decision making but did not give rise to clear patterns of decision-making practice that might explain why the ratings of our evaluator groups were more or less strongly related to student achievement. We do not interpret our results to mean that evaluator will, skill, and context are unimportant, but rather that they interact in complex ways that are idiosyncratic across evaluators. These idiosyncrasies exist despite the guidance provided by

standards, rubrics, and district training. It became apparent to us when considering the qualitative findings that a complex interaction of will, skill, and context could be invoked to explain the ratings of evaluators, but these interactions are not strongly similar to those of other evaluators within the two sample categories.

Another explanation for why simple differences in will, skill, and context did not appear across the groups is that the recalled or espoused practices of the evaluators may have had little influence on their ratings. The evaluators may have relied on intuition or gut-level feelings about teachers, without even being completely aware of doing so. This possibility is supported by the fact that evaluators generally had difficulty thinking analytically about how they make decisions. The differences in the rating–student achievement relationships between evaluators may be because of the fact that some evaluators' intuitions were simply more attuned to performance factors related to student achievement than other evaluators.

Both of these explanations are consistent with our observation that the district's teacher evaluation process is a weak situation for evaluators (Mischel, 1977). A weak situation is one in which incentives, support, or normative expectations for defined behaviors or outcomes are ambiguous or absent. In weak situations, individuals do not share a common perception of what is expected of them. Therefore, they may fall back to whatever approach they are most comfortable with, rather than the one ostensibly designed to be used. In contrast, strong situations generate uniform perceptions concerning appropriate behaviors.

In general, in this district, there is little to make evaluation decision making a strong situation. Relatively little emphasis is placed on following a uniform process; there is a low level of accountability for accurate evaluation unless a teacher's job is at stake; evaluators are not required to take follow-up training; and the ratings have little consequence for most teachers. Evaluator training emphasized management of the task rather than evaluation accuracy or quality of feedback. All of these factors contribute to a situation that allows unique combinations of evaluator and context factors to govern decision making. They also help explain the tendency for evaluations on average to be lenient.

In contrast, where accurate assessment is paramount (e.g., professional certification, licensing or compensation), the process is often structured as a strong situation for evaluators, with a clear incentive structure and normative expectations for accuracy, and supports for learning to accurately rate. Yet because principals have to work with teachers after their evaluation is complete, principals may still tend to inflate ratings even in high stakes situations to maintain collegiality. Indeed, research on private sector

appraisal for compensation has found such practices common (Murphy & Cleveland, 1995). Our findings in this low-stakes situation clearly indicated that leniency was prevalent.

Another explanation is that our interviews did not uncover the influences of evaluator will, skill, and context in the detail required to reveal clear differences and similarities. This is most likely to be a problem with skill, because we focused on training and experience rather than assessing skill directly. Yet the literature does identify preparation and experience as important in evaluator decision making, along with evaluator attitudes and the process followed. It would appear that at the relatively gross level at which we were able to asses them, a simple combination of will, skill, and context does not explain differences in the rating–student achievement relationship across evaluators.

It might also be argued that the differences in validity on which we based our categorization of evaluators were simply artifacts of the relatively small number of teachers for whom we could obtain both evaluation ratings and student achievement data. We think this is unlikely because we chose evaluators whose ratings had strong positive or negative relationships with student achievement in two consecutive years. Also, though a statistical significance test does not apply in a strict sense, the difference between a correlation of .55 and –.26 would be statistically significant at the .05 level with $n = 8$. Our study did highlight the importance of checking such statistical measures with qualitative data sources. A statistical test of evaluator accuracy should not be relied on as a sole basis for determining evaluation decision-making accuracy.

A less plausible explanation is that the evaluators we interviewed distorted their responses to tell us what they thought we or the district would like to hear. This is doubtful because all respondents were quite willing to admit when they did not follow district-prescribed procedures. Few seemed to have any concern that such actions would be disapproved by us or the district.

## Policy and Practice Implications

Our interpretations of these results have several implications for evaluation policy and practice. We had hoped that we could identify evaluator practices associated with higher validity, which districts could then use to train evaluators to follow. Although disappointing, our failure to find such practices is important because it shows the complexity in identifying and assuring the use of good evaluation practice. Providing evaluators with relatively detailed rubrics or rating scales describing generic teaching behaviors

thought to promote student learning, coupled with initial training in applying them, is not enough to ensure that all evaluators' ratings will be positively related to student achievement. If policy makers and program designers want evaluation scores to be more highly related to some criterion such as student achievement, it will take more than specific rubrics and basic training of evaluators in the process to achieve a strong relationship.

The strength of the situation as perceived by evaluators may also need to be increased. This would involve adding incentives for accurate evaluation, oversight focused on encouraging evaluators to differentiate among teachers when making ratings, and ongoing practice with feedback in making accurate evaluations. Evaluators need to perceive that district expectations and peer practices are centered on applying a uniform evaluation process and a consistent interpretation of the rubrics to lessen the influence of idiosyncratic combinations of will, skill, and context or evaluator intuition.

Training may also need to focus on the type of instruction that has a deep impact on student learning, such as the approach described by Nelson and Sassi (2005). This approach starts with training principals (or other evaluators) to develop a firm understanding of effective teaching and learning in at least one content area. This knowledge can form a foundation for effective instruction that can be applied to professional development and to observations and feedback provided to teachers through evaluation practices. Whether this type of training in fact leads to more accurate evaluations, especially across content areas and grade levels, would be important to examine in future research.

Our findings suggest caution in using principal evaluation decisions for teacher compensation outcomes. Although this district did not design the evaluation process for teacher pay, the results here suggest that considerable attention to evaluator training, oversight, and improving evaluation validity is needed to promote more consistency among evaluators before compensation is tied to evaluations of teaching behaviors. Such consistency would be needed to overcome the common teacher suspicion that evaluation results depend on who is doing the evaluating. Though the evaluators we studied were not capricious, careless, or obviously biased against particular teachers, the potential for inconsistency across evaluators was clearly present.

## Research Implications

This study also has implications for future research on teacher evaluation and school leader decision making. In weak situations, generalization

across evaluators may be quite difficult, with the consequence that it will be difficult to further theory about evaluation decision making. Additionally, more intensive methodologies may be needed to get a better understanding of decision making. We found it was harder than expected to get evaluators to discuss their decision-making processes. This may be because of asking them to retrospectively describe what they did. Another reason may be the inexperience of evaluators' in thinking analytically about how they make decisions. One way to address this would be to observe evaluators during the evaluation tasks and having them think aloud while making evaluation decisions. This could provide more accurate and detailed information about decision processes (Ericsson & Simon, 1993; Martin & Klimoski, 1990). Another approach could have evaluators retain evaluation logs or diaries to note evidence they have collected and how they might interpret that evidence as it is gathered. These would serve as reference points during discussions with evaluators and may help evaluators recall data to make evaluation decisions.

Because a shared school culture and vision of good teaching may help to reduce leniency and influence what evaluators look for, it may also be useful to add these to the conceptualization of school context. More detailed analysis of context, including such factors as instructional focus and professional community could shed additional light on the evaluator's motivation to be accurate, and the validity of evaluation ratings. Future research might incorporate measures of school culture, such as relational trust (Bryk & Schneider, 2002) and professional community (Louis & Marks, 1998) into studies of evaluation rating validity and evaluator performance.

Nonetheless, because of the complexity of the evaluation process, it may not be possible to fully capture all of the influences on evaluators' decisions. Our interviews with the initial group of 23 principals and assistant principals showed that the evaluation decision making took place throughout the school year and contained multiple decision points. The evaluators had numerous contacts with teachers to discuss evaluation goals and evidence, observe teaching performances, discuss observations, and provide written or verbal feedback. Individually or cumulatively, each of these decision points could influence evaluation accuracy and validity. Capturing the nuances involved with each decision point is a daunting research challenge. At best, researchers may be able to narrow the focus to the most critical decision points. Identifying these decision features are worth exploring through additional research and could yield important information for teacher evaluation practice and training.

## Conclusion

This study demonstrated that there can be substantial variation in the criterion-related validity of evaluators' ratings of teacher performance. This suggests that estimates of criterion-related validity should be interpreted with caution and that the quality of ratings may vary considerably across evaluators. Our study does not dismiss will, skill, and context as potentially important factors in evaluation decision making, but it does illustrate the complexity in fully uncovering these factors. Simple differences in evaluator will, skill, and context were not found between more and less valid groups of evaluators. This suggests that extensive evaluator training and other interventions to standardize the rating context are needed to ensure consistency. There remains considerable ground to cover in research relating to accuracy and validity in teacher evaluation. Future research is needed to explore evaluator cognitive processes and organizational conditions to unfold more nuanced characteristics in the use of different evaluation system features that may help improve evaluation validity.

## Appendix

The teacher performance evaluation composite measure is made up of the following standards:

- The teaching displays solid content knowledge and uses a repertoire of current pedagogical practices for the discipline being taught (includes 10 elements from two performance domains).
- 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 (includes nine elements from two performance domains).
- 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 (includes three elements from one domain).
- The teaching engages students cognitively in activities and assignments; groups are productive; and strategies are congruent to instructional objectives (includes three elements from one domain).

## Note

1. We also acknowledge the contribution of our colleague Carolyn Kelley, whose ideas about will, skill, and structure in teacher evaluation practice helped inform this study. Halverson, Kelley, and Kimball (2004) also considered these aspects in their study of principals' sense making in teacher evaluation practices.

## References

Bernardin, H. J., & Cardy, R. L. (1982). Appraisal accuracy: The ability and motivation to remember the past. *Public Personnel Management*, *11*(4), 352-357.

Bretz, R. D., Milkovich, G. T., & Read, W. (1992). The current state of performance appraisal research and practice: Concerns, directions, and implications. *Journal of Management*, *18*(2), 321-352.

Brophy, J. (1986). Teacher influences on student achievement. *American Psychologist*, *41*(1), 1069-1077.

Bryk, A., & Schneider, B. (2002). *Trust in schools: A core resource for improvement*. New York: Russell Sage.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology*, *74*(1), 130-135.

Danielson, C. (1996). Enhancing professional practice: A framework for teaching. Alexandria, VA: Association for Supervision and Curriculum Development.

Danielson, C., & McGreal, T. (2000). *Teacher evaluation to enhance professional practice.* Alexandria, VA: Association for Supervision and Curriculum Development.

DeCotiis, T., & Petit, A. (1978). The performance appraisal process: A model and some testable propositions. *Academy of Management Review*, *3*, 635-646.

DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Performance*, *33*, 360-396.

Ellett, C. D. (1997). Classroom-based assessments of teaching and learning. In J. Stronge, *Evaluating teaching: A guide to current thinking and best practice* (pp. 107-28). Newbury Park, CA: Sage.

Ellett, C. D., Annunziata, J., & Schiavone, S. (2002). Web-based support for teacher evaluation and professional growth: The Professional Assessment and Comprehensive Evaluation System (PACES). *Journal of Personnel Evaluation in Education*, *16*(1), 63-74.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Revised ed.). Cambridge, MA: The MIT Press.

Freeberg, N. E. (1969). Relevance of rater-ratee acquaintance in the validity and reliability of ratings. *Journal of Applied Psychology*, *53*, 518-524.

Grey, R. J., & Kipness, D. (1976). Untangling the performance appraisal dilemma: The influence of perceived organizational context on evaluative processes. *Journal of Applied Psychology*, *61*, 329-335.

Halverson, R., Kelley, C., & Kimball, S. (2004). Implementing teacher evaluation systems: How principals make sense of complex artifacts to shape local instructional practice. In W. Hay & C. Miskel (Ed.), *Educational administration, policy, and reform: Research and measurement. a volume in research and theory in educational administration* (Vol. 3, pp. 153-188.). Greenwich, CT: George F. Johnson.

Hedge, J. W., & Kavanagh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. *Journal of Applied Psychology*, *73*(1), 68-73.

Heneman, H. G., III, Milanowski, A., Kimball, S. M., & Odden, A. (2006). Standards-based teacher evaluation as a foundation for knowledge- and skill-based pay (CPRE Policy Brief, RB-45). Philadelphia: Consortium for Policy Research in Education, University of Pennsylvania.

Honig, M. I. (2006). Complexity and policy implementation: Challenges and opportunities for the field. In M. I. Honig (Ed.), *New directions in education policy implementation: Confronting complexity* (pp. 1-24). Albany, NY: State University of New York Press.

Ivancevich, J. M. (1983). Contrast effects in performance evaluation and reward practices. *Academy of Management Journal*, *26*(3), 465-476.

Jacob, B. & Lefgren, L. (2006). When principals rate teachers. *Education Next*, *6*(2), 58-64.

Judge, T. A., & Ferris, G. R. (1993). Social context of performance evaluation decisions. *Academy of Management Journal*, *38*(1), 80-105.

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 Dissertation 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*(4), 241-268.

Kimball, S. M., White, B., Milanowski, A.T., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education*, *79*(4), 54-78.

Klein, S. J. (1998). Standards for teacher tests. *Journal of Personnel Evaluation in Education, 12*(2), 123-138.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback intervention on performance: A historical review, meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254-284.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, *87*(1), 72-107.

Louis, K. M., & Marks, H. M. (1998). Does professional community affect the classroom? Teachers' work and student experiences in restructuring schools. *American Journal of Education*, *106*, 532-575.

Martin, S. L., & Klimoski, R. J. (1990). Use of verbal protocols to trace cognitions associated with self- and supervisor evaluations of performance. *Organizational Behavior & Human Decision Processes*, *46*(1), 135-154.

Medley, D. M., & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *Journal of Educational Research*, *80*(4), 242-247.

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

Milanowski, A. T., Kimball, S. M., & Odden, A. R. (2005). Teacher accountability measures and links to learning. In R. Rubenstein et al. (Eds.), *Measuring school performance & efficiency: Implications for practice and research. Yearbook of the American Education Finance Association* (pp. 133-188). Larchmont, NY: Eye on Education.

Mischel, W. (1977). The interaction of person and situation. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology* (pp. 333–352). Hillsdale, NJ: Lawrence Erlbaum.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives.* Thousand Oaks, CA: Sage.

Napier, N. K., & Latham, G. P. (1986). Outcome expectancies of people who conduct performance appraisals. *Personnel Psychology*, *39*, 827-837.

Nelson, B. S., & Sassi, A. (2005). The effective principal: Instructional leadership for high-*quality learning.* New York: Teachers College Press.

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practice* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1-22.

Smith, D. E. (1986). Training programs for performance appraisal: A review. *Academy of Management Review*, *11*(1), 22-40.

Smither, J. W., Barry, S. R., & Reilly, R. R. (1989). An investigation of the validity of expert true score estimates in appraisal research. *Journal of Applied Psychology*, *74*(1), 143-151.

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

Stodolsky, S. S. (1990). Classroom observation. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 175-190). Newbury Park, CA: Sage.

Tashakkori, A., & Teddlie, C. (1998). Mixed methodology: Combining qualitative and quantitative approaches. *Applied Social Research Methods Series* (Vol. 46., pp. 297-319). Thousand Oaks, CA: Sage.

Tziner, A., Murphy, K. R., & Cleveland, J. N. (2001). Relationship between attitudes toward organizations and performance appraisal systems and rating behavior. *International Journal of Selection and Assessment*, *9*(3), 226-239.

U.S. Department of Education. (2006). *Teacher incentive fund, notice inviting applications for New Awards for Fiscal Year (FY) 2006.* Retrieved July 17, 2006, from http://www.ed.gov/legislation/FedRegister/announcements/2006-2/050106e.html

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, *67*, 189-205.

**Steven M. Kimball,** PhD, is a researcher with the Consortium for Policy Research in Education (CPRE) and the Wisconsin Center for Education Research at the University of Wisconsin-Madison. His research has focused on standards-based teacher evaluation and pay systems, school leadership, school-based performance award programs, and National Board Certification.

**Anthony Milanowski,** PhD, is an assistant scientist with the Consortium for Policy Research in Education and the Wisconsin Center for Education Research at the University of Wisconsin–Madison. He coordinated the CPRE Teacher Compensation Project's research on standards-based teacher evaluation and teacher performance pay. He has taught human resource management courses for the Schools of Business and Education at University of Wisconsin–Madison.