CPRE Policy Briefs

Reporting on Issues and Research in Education Policy and Finance

May 2006 RB-45

Graduate School of Education

University of Pennsylvania

Standards-Based Teacher Evaluation as a Foundation for Knowledge- and Skill-Based Pay

By Herbert G. Heneman III, Anthony Milanowski, Steven M. Kimball, and Allan Odden

State accountability systems and the federal No Child Left Behind Act have put additional demands on schools and teachers to improve teacher quality and improve student achievement. Many researchers (e.g., Cohen, 1996; Corcoran & Goertz, 1995; Floden, 1997; Newman, King, & Rigdon, 1997) have argued that such improvements will require a substantial increase in the instructional capacity of schools and teachers. One strategy for capacity building is to provide teachers with incentives to improve their performance, knowledge, or skills. The incentive strategy requires the design and implementation of alternative teacher compensation systems that depart from the single salary schedule (Odden, 2000; Odden & Kelley, 2002). Though slow to take hold, the incentive strategy is currently being pursued by several states (Peterson, 2006). Most of these new or proposed plans link pay to combinations of assessments of teacher performance, acquisition of new knowledge and skills, and student test score gains. Denver's widely followed Pro Comp plan also contains these components.

The Teacher Compensation Group of the Consortium for Policy Research in Education (CPRE) has been studying the design and effectiveness of such systems for nearly a decade. We initially focused on school-based performance award programs, in which each teacher in a school receives a bonus for meeting or exceeding schoolwide student achievement goals (Heneman, 1998; Heneman & Milanowski, 1999; Kelley, Heneman, & Milanowski, 2002; Kelly, Odden, Milanowski, & Heneman, 2000). We then shifted our attention to knowledge- and skill-based pay (KSBP) plans, an approach that provides teachers with base pay increases for the acquisition and demonstration of specific knowledge and skills thought to be necessary for improving student achievement.

Our initial research described a variety of experiments with KSBP plans (see Odden, Kelley, Heneman, & Milanowski, 2001). We found plans that were rewarding numerous knowledge and skills, including (a) additional licensure or certification, (b) participation in specific professional development activities, (c) National Board Certification, (d) mastery of specific skill blocks such as technology or authentic assessment, (e) leadership activities, and (f) teacher performance as measured by a standards-based teacher evaluation system. We also found districts experimenting with standards-based teacher evaluation without an intended pay link. As described below, in standards-based teacher evaluation systems, teachers' performance is evaluated against a set of standards that define a competency model of effective teaching. Such systems replace the traditional teacher evaluation system and seek to provide a more thorough description and accurate assessment of teacher performance. Findings from our research on some of these systems are the focus of this issue of *CPRE Policy Briefs*.

A popular competency model of teacher performance is the Framework for Teaching, developed by Danielson (1996) and intended to apply to all grade levels and subjects. The Framework attempts to describe the full range of teacher performance, from beginner to expert. It defines four performance domains: planning and preparation, the classroom environment (classroom management), instruction, and professional responsibilities. For each domain there is a set of specific performance components, each of which has one or more elements (a total of 66). Each element has four performance levels/standards, defined by specific behavioral rubrics. An example of the rubrics associated with two elements of the component "communicating clearly and accurately" from the instruction domain is shown in Exhibit I.

Consortium for Policy Research in Education

University of Pennsylvania

Harvard University

Stanford University

University of Michigan

University of Wisconsin-Madison

Exhibit I. Example Rubric From Framework for Teaching

Domain 3: Instruction
Component 3a: Communicating Clearly and Accurately

	Level of Performance			
Element	Unsatisfactory	Basic	Proficient	Distinguished
Directions and procedures	Teacher directions and procedures are confusing to students.	Teacher directions and procedures are clarified after initial student confusion or are excessively detailed.	Teacher directions and procedures are clear to students and contain an appropriate level of detail.	Teacher directions and procedures are clear to students and anticipate possible student misunderstanding.
Oral and written language	Teacher's spoken language is inaudible, or written language is illegible. Spoken or written language may contain grammar and syntax errors. Vocabulary may be inappropriate, vague, or used incorrectly, leaving students confused.	Teacher's spoken language is audible, and written language is legible. Both are used correctly. Vocabulary is correct, but limited or is not appropriate to students' ages or background.	Teacher's spoken and written language is clear and correct. Vocabulary is appropriate to students' age and interests.	Teacher's spoken and written language is correct and expressive, with well-chosen vocabulary that enriches the lesson.

Adapted from Danielson (1996), p. 91.

The Framework for Teaching (with adaptation to the local context) can be used as the performance measure for a standards-based teacher evaluation system. Evaluators can gather evidence from various sources (e.g., classroom observation, portfolios, logs) about the teacher's performance and then rate the teacher's performance on each element. Written and verbal feedback can be provided, and action plans for improvement can be developed. Moreover, to integrate the evaluation system into a KSBP plan, the overall or average rating can be used to determine placement and movement on the KSBP plan salary schedule.

An example of such a schedule is shown in Exhibit II. Teachers are placed into five levels of increasing performance competency: apprentice, novice, career, advanced, and accomplished. Placement is based on performance ratings on the four domains from the Framework for Teaching. Movement into higher levels occurs as performance ratings improve, with the proviso that a teacher can remain an apprentice for only two years and a novice for only five years. Within

Exhibit II. Example Knowledge- and Skill-Based Pay Schedule

Level	Performance Ratings	Beginning Rate	Additional Salary Steps
Accomplished (no maximum years)	"Distinguished" on all domains	$60,000	2 steps to $65,000
Advanced (no maximum years)	"Distinguished" on Instruction and one other domain; "proficient" on the others	$53,000	3 steps to $57,000
Career (no maximum years)	"Proficient" on all domains	$38,000	5 steps to $50,000
Novice (5 years maximum)	"Basic" on all domains	$32,000	2 steps to $36,000
Apprentice (2 years maximum)	(Entry level)	$30,000	None

Additional base pay may be provided for advanced degrees, special certifications, or leadership activities.
Adapted from Odden and Kelley (2002).

each level there are salary steps based on years of service at that level, but once at the maximum step, teachers need to move to the next competency level to increase their base pay. There is no overlap in the ranges for each level, which encourages teachers to develop the competency needed to move to a higher level. In this schedule, additions to base pay can be earned for advanced degrees, special certifications, and leadership activities (specific amounts not shown).

Given the promise and complexity of standards-based teacher evaluation systems, we felt it was necessary to study their design and effectiveness in detail to help determine if, and how, these systems could be used as part of KSBP plans. In order to decide on whether standards-based evaluation systems could be used for such plans, we believed that policymakers would want information on their validity, acceptability, and usability. Due to the newness of both standards-based teacher evaluation systems and KSBP plans for teachers, there were few sites available to study. We located and gained the cooperation of four sites throughout the country: Cincinnati, Vaughn Charter School in Los Angeles, Washoe County (Reno/Sparks, Nevada), and Coventry (Rhode Island). Cincinnati and Vaughn were beginning with a pilot study of a standards-based teacher evaluation system that would then later link to a KSBP plan. Washoe began with a pilot study and then moved to districtwide implementation of a standards-based teacher evaluation system, while Coventry did not conduct a pilot study before implementation; neither of these two sites linked implementation of its system with a KSBP plan. Exhibit III summarizes the standards-based teacher evaluation systems at each of these sites.

The research reported in this brief was conducted by The Consortium for Policy Research in Education (CPRE) and funded by the Institute of Education Sciences, United States Department of Education, under Grant No. R308A960003. Opinions expressed in this brief are those of the authors and do not necessarily reflect the views of the Institute of Education Sciences, the United States Department of Education, CPRE, or its institutional members.

Exhibit III. Evaluation Systems at the Four Sites

District	Cincinnati	Washoe	Coventry	Vaughn
District Background	Midwest, urban 40,000 students, primarily African American	Western, urban-suburban-rural 60,000 students, majority White, large Hispanic minority	East coast, suburban 6,000 students, predominantly White	West coast, urban 1,200 student, predominantly Hispanic
Teachers	2,500	3,300	475	40 pre-K through 5
Schools	81	88	9	1 charter school
Pilot test year	1999-2000	1999-2000	N/A	1999-2000
Year first full implementation	2000-2001	2000-2001	1998-1999	2000-2001
Competency model (evaluation standards)	Customized version of Framework for Teaching; 4 domains, 16 standards	Minor modifications to Framework for Teaching; 4 domains, 23 components, 68 elements	Modified version of Framework for Teachin g; 3 domains, 18 components, 56 elements	2 domains modeled on Danielson Framework; 12 other content-specific domains developed locally
Evaluation procedures	Comprehensive evaluation on all domains for new teachers and veterans who are identified as need ing improvement, at certain steps on the schedule, or desiring to become lead teachers. Original plan was for comprehensive evaluation once every 5 years. All others undergo less rigorous annual evaluation on one domain each year. Comprehensive evaluation consisted of 5-6 classroom observations and a teacher-prepared portfolio.	Nontenured teachers evaluated on all domains/elements, via 9 classroom observations. Tenured teachers evaluated on 1 domain (minor) or 2 domains (major) over 3-year cycle via at least 1 classroom observation. No portfolio required, but evaluators also look at artifacts like student work.	Nontenured teachers evaluated on a subset of the standards each year for 2 years, then receive a full evaluation in the 3rd year. Evaluation is ba sed on at least 2 observations. Tenured teachers are evaluated on all domains every 2, 3, or 4 years depending on prior rating level. At least 1 observation is required.	All teachers evaluated annually, with 2 ratings per year on selected domains, primarily via classroom observations. Observations are conducted as many times as is necessary over a 2-week period each semester. No portfolio is required, but evaluators also look at artifacts like student work. All rated on 5 core domains and selected content-relevant domains.
Evaluators	Peer evaluators, principals, and assistant principals	Principals and assistant principals	Principals and department heads	Self, peer, and assistant principals

At all four sites we focused our research on the standards-based teacher evaluation system. The teachers in Cincinnati and Vaughn were aware of the possibility that evaluation scores would be linked to their future pay via a KSBP plan; the teachers in Washoe and Coventry knew there was no intent to link evaluation scores to pay. We used multiple methods research methods, including both qualitative and quantitative data collection and analysis. A summary of our research approach at each site can be found in the appendix.

Our research focused on four questions:

1. *What is the relationship between teachers' standards-based teacher evaluation scores or ratings and the achievement of their students*?

This question is fundamental, since there is no point in encouraging teachers to develop and use competencies that are not related to student achievement. Further, if the scores are to be linked with pay increases as part of a KSBP plan, the evaluation system needs to be able to distinguish those teachers who facilitate greater levels of student achievement in order to justify rewarding them. Significant positive relationships would provide evidence not only that the standards-based evaluation ratings can be used to identify good teaching, but also that the teacher competencies underlying the system do help facilitate student achievement.

2. *How do teachers and administrators react to standards-based teacher evaluation as a measure of instructional expertise*?

This question is important because administrator and teacher reactions are a major determinant of the willingness of administrators to use the system as designed, and of teachers to agree to link pay with assessments of performance. The initial acceptance and long-term survival of the evaluation and KSBP systems will be jeopardized if administrators and teachers believe the evaluation system is unfair, overly burdensome, and not useful in guiding teacher efforts to improve performance.

3. *Is there evidence that standards-based teacher evaluation systems influence teacher practice*?

This question is important in order to assess the potential of a KSBP plan to motivate teachers to improve instructional practice. The evaluation system must provide guidance for teachers about districts' performance expectations and feedback on the current level of performance. Even if the evaluation will not be linked to pay, feedback from evaluators and the desire for a favorable evaluation provide incentives for improvement. So we investigated the influence of evaluation systems on what teachers report they do in the classroom.

4. *Do design and implementation processes make a difference*?

These processes can affect the overall viability and impact of the system. The system is not likely to have a sustained impact on teacher skill development, or survive for long, if it is cumbersome, prone to implementation glitches, and unaligned with other human resource management programs that affect instructional capacity.

Relationship Between Teacher Evaluation Scores and Student Achievement

We assessed the relationship between teachers' performance evaluation scores and student achievement by correlating teachers' overall evaluation scores with estimates of the value-added academic achievement of the teachers' students. Our value-added measure was estimated controlling for prior student achievement and other student characteristics, such as socioeconomic status, that influence student learning. At each site, we have analyzed multiple years of data. Exhibit IV summarizes the results for reading and mathematics.

We found positive relationships between teacher evaluation scores and student achievement, though the average relationship varied across the four sites. At the Vaughn school, the relationship was substantial, with an average correlation over the three years we studied of 0.37 in reading and 0.26 in mathematics.[1] In Cincinnati, the relationship was similar, with a three-year average correlation of 0.35 in reading and 0.32 in mathematics. In Washoe, the relationships were somewhat smaller; the average correlations were 0.22 and 0.21 for reading and math achievement, respectively. In Coventry, the average correlation between teacher ratings and student achievement in reading was 0.23, and 0.11 with mathematics

[1] The correlation is a quantitative indicator of the degree of association between two variables. It ranges from -1.00 to + 1.00, with a correlation of .00 indicating no association between the variables. Correlations of .20 to .40 are quite common in educational research, and correlations in this range are considered meaningful indicators of an association between variables.

achievement.[2] Note that one would not expect to find a perfect or even near-perfect correlation between evaluation scores and student achievement, given the various other factors that influence both. On the student achievement side, tests are not perfect measures of student learning, nor is teacher behavior its only cause. Teacher evaluation scores are also not perfect representations of teachers' actual classroom behavior. Given the size of recent estimates of likely teacher effects on student achievement (Nye, Konstantopolis, & Hedges, 2004; Rowan, Correnti, & Miller, 2002), the average correlations for Cincinnati and Vaughn are about what one might expect.

We speculate that Cincinnati and Vaughn have higher average correlations in part due to the use of multiple evaluators. In addition, Cincinnati evaluators received intensive, high-quality training. Vaughn evaluators could draw on a strong shared culture and history of working on instruction that fostered agreement on what good teaching looks like. In contrast, at the two other sites, a single evaluator (the school principal or an assistant principal) made the ratings, and less training was provided to the evaluators. Measurement error, relatively small samples in some grades and subjects, differences in the quality and coverage of student tests, and idiosyncrasies in different evaluators' interpretations of teacher performance are likely explanations for the variation in the strength of the relationship across years within each of the four sites.

Overall, our results suggest that the scores from standards-based performance evaluation systems can have a substantial positive relationship with student achievement and that the instructional practices measured by these systems contribute to student learning. The evidence supports the potential usefulness of a well-designed and rigorously implemented standards-based teacher evaluation as a basis for a KSBP pay system for teachers.

Exhibit IV. Average Correlations Between Teacher Evaluation Ratings and Estimates of Average Student Achievement in Reading and Mathematics

		Subject	
Site	Grades		
Cincinnati		Reading	Math
2001-2002	3 -8	.48	.41
2002-2003	3 -8	.28	.34
2003-2004	3 -8	.29	.22
3-year average:		.35	.32
Coventry			
1999-2000	2,3,6	.17	.01
2000-2001	2,3,4,6	.24	-.20
2001-2002	4	.29	.51
3-year average:		.23	.11
Vaughn			
2000-2001	2 -5	.48	.20
2001-2002	2 -5	.58	.42
2002-2003	2 -5	.05	.17
3-year average:		.37	.26
Washoe			
2001-2002	3 -5	.21	.19
2002-2003	4 -6	.25	.24
2003-2004	3 -6	.19	.21
3-year average:		.22	.21

[2] Our research is reported in several journal articles and a book chapter. Validity results can be found in Kimball, White, Milanowski, and Borman (2004); Milanowski (2004); Gallagher (2004); and Milanowski, Kimball, and Odden (2005). Research on the implementation of the systems and teacher reactions can be found in Heneman and Milanowski (2003); Kimball (2002); and Milanowski and Heneman (2001).

Teacher and Administrator Reactions

In addition to producing ratings which correlate with student achievement, a standards-based teacher evaluation system must be accepted by those who use it if it is to survive and contribute to performance improvement. Accordingly, we assessed teacher reactions by interviewing hundreds of teachers and conducting multiple surveys at three of our sites. We assessed administrator reactions through interviews. Exhibit V summarizes the results for teachers.

The most positive and least varied reactions were to the performance competency model embedded within the evaluation system. Teachers generally understood the standards and rubrics comprising the evaluation systems, and agreed that the performance described at higher levels described good teaching. Many teachers told us that this was the first time they ever had a clear and concise understanding of the district's performance expectations for their instructional practice. Additionally, many reported that the use of the teaching standards helped improve dialogue with their principals about teaching and performance expectations.For the other areas shown in Exhibit V, teacher reactions were more mixed. Numerous specific aspects of the new evaluation systems and their implementations contributed to the variety of reactions. This variety, including both positive and negative reactions, suggests that close attention to design and implementation issues is needed in order to maximize teacher acceptance.

Administrators also generally accepted the performance competency model and the evaluation system based on it. Like teachers, they often commented that evaluation dialogue was improved under the new system. Many principals valued the increased opportunity to discuss instruction with teachers and felt that the greater amount of evidence they collected, combined with the explicit rubrics describing the four levels of teacher performance, helped them do a better job as evaluators. However, principals also saw the

Exhibit V. Summary of Teacher Reactions in Cincinnati, Vaughn, and Washoe County

Issue	Findings
Competency model (standards and rubrics)	Teachers in general accepted the model as appropriate and as an adequate description of good teaching. There were some concerns about the attainability of the highest performance levels and the appli cability of the model to some subject areas (art, music, special education).
Evaluation evidence	Teachers generally favored multiple classroom observations but were not enthusiastic about portfolios. These were perceived as burdensome and the requirements confusing.
Evaluators	Reactions were mixed, influenced by perceptions of trust, subject matter expertise, familiarity with the school context, and preparation. Reaction to peer evaluators from outside the school (used in Cincinnati and to a small extent at Vaughn) was often negative. There were concerns that the workloads of principals and other administrators left little time to evaluate, provide feedback, and coach.
Rating accuracy and fairness	Most teachers perceived the ratings as fair and accurate . Some concerns were raised about evaluator competence, strictness, and leniency.
Feedback and assistance for improvement	Teachers recognized the potential of standards -based systems to generate better feedback, but this potential was not always realized . Teachers generally wanted more feedback, more timely feedback, and assistance for improvement. Teachers were more positive about the system when they were provided with useful feedback about how to improve their performance. Feedback and assistance tende d to focus more on classroom management and basic instructional issues than on more complex aspects of instruction.
System as a whole	Teachers had mixed reactions to the system as a whole. In Washoe and Vaughn, reactions were mostly positive. In Cincinnat i, many teachers had strong negative reactions, especially veteran teachers. They found the system stressful and disruptive.

new evaluation procedures as more work to implement, requiring many of them to lengthen their work day and complete evaluations on weekends. Some principals attempted to shortcut the process by reducing the number and length of observations, providing relatively general written feedback to teachers or focusing their time primarily on new or struggling teachers.

Impact on Teacher Practice

We assessed effects on teaching practice primarily through interviews with teachers and evaluators. Many teachers reported that the new system had positive impacts on their instructional practice. But the evidence indicated that the initial effects of standards-based teacher evaluation on teacher practice tended to be broad, but relatively shallow. Engaging in more reflection, improved lesson planning, and better classroom management were commonly cited impacts. These are basic but critical features of instructional practice that can create conditions for student learning. Teachers were less likely to report changing their instruction to a pedagogy characterized by student-initiated activities or empowerment, as emphasized in some of the "distinguished" levels of the Framework rubrics. This is not surprising since the level of feedback and assistance provided in most cases emphasized classroom management and general pedagogy. Nor was professional development highly focused on the teaching practices assessed by the evaluation systems. One interesting impact was that many teachers being evaluated began to take *student* standards more seriously, because of the emphasis on teaching to those standards in the evaluation systems. One principal we interviewed remarked that one year of standards-based evaluation had done more to motivate teachers to pay attention to the student standards than years of workshops.

Design and Implementation

At three of our sites, the design process included not only gathering input from representatives of teachers, principals, and central office administrators, but also a pilot test of the system prior to full implementation. These practices helped build support and uncover potential implementation problems. In many cases, these problems were addressed before all teachers were required to undergo the new evaluation process or shortly after full implementation. Despite the best intentions, however, there were still some aspects of the evaluation process or related systems that caused problems. And in the case of Cincinnati, the teacher association sought substantial revisions of the evaluation process, and the members ultimately voted to reject the new teacher KSBP pay schedule with raises to depend in part on the results of evaluation.

A problem most apparent at Cincinnati and Vaughn was the difficulty of ironing out all the glitches in the performance evaluation process. Because of the complexity of these systems and the accompanying KSBP plans, it proved hard to foresee all potential problems and address them to the satisfaction of teachers and administrators in the first year of implementation. Implementation glitches led teachers to perceive that "they are building the airplane while it flies." This was unsettling to many teachers and provoked doubts about the validity and fairness of the system. Vaughn was more successful in addressing these concerns, being able to take quicker action on implementation problems and having a higher level of trust between teachers and administrators.

Training for teachers and administrators was also an issue. Principals were generally responsible for training teachers in the new system, resulting in uneven quality. Teacher training tended to be process oriented, with limited emphasis on understanding how to develop and demonstrate the performance competencies. Training for evaluators varied considerably. Cincinnati invested considerable resources in teaching all evaluators how to collect evidence and apply the rubrics, and produced good interrater agreement and a relatively strong relationship between evaluation scores and student achievement. Coventry and Washoe spent less time on training, and had weaker relationships. While all sites provided training in the first year of implementation, one site failed to train principals or teachers who joined the district after the initial implementation. At all sites, administrator training did not appear to put much emphasis on providing useable feedback, setting performance goals, and coaching.

One factor that limited effects of the new evaluation systems was that the competency model was not part of a coherent effort to drive the development of a new performance culture. Except at Vaughn, the systems were not sufficiently linked to broader strategies for improving instruction or student achievement, nor to other parts of the human resource management system. Although there was some attempt at alignment, districts generally aligned only one or two of their other human resource systems (i.e., recruitment, selection, induction, mentoring, professional

development, compensation, performance management, and instructional leadership) with the instructional vision in the competency model, forfeiting the opportunity to have these programs reinforce the teacher performance evaluation system and the vision of good instruction embodied in it.

At each site, there was at least one key district administrator who shepherded the new system through the design, pilot, and implementation phases, and helped rally others behind the cause. Despite their efforts, in two districts at least one of the leaders of other important functions (curriculum and instruction, principal supervision, professional development, or human resources) remained resistant or disengaged. This seemed to be due to the lack of active superintendent engagement with the performance competency model underlying the teacher evaluation system. So instead of tightly integrating the performance evaluation system with other efforts to improve teacher quality, these districts treated the performance evaluation or pay system as just another isolated reform.

Another issue was the lack of alignment between the teacher performance evaluation system and the performance expectations and evaluation process for school administrators. Most sites did not hold administrators accountable for the quality of their efforts to evaluate and support teachers through the new system, or even for completion of the evaluation process. Though most evaluators worked hard to accommodate the new system, lack of accountability led some administrators to minimize involvement in the new teacher evaluation system, fail to evaluate teachers in a timely way, or fail to provide the feedback teachers desired.

Using Standards-Based Teacher Evaluation as the Foundation for Knowledge- and Skill-Based Pay

Our results provide evidence that ratings from standards-based teacher evaluation systems can have a meaningful relationship with measures of student achievement. There is thus evidence that these evaluation systems are holding teachers accountable for competencies related to student achievement. These results also reassure us that holding teachers accountable for their performance is worthwhile, in terms of the outcome policymakers and much of the public feels is most important improved student achievement. These findings suggest that standards-based teacher evaluation systems could be used as the foundation of a KSBP plan, but only if the evaluation system is designed and implemented properly to support this use.

Guidelines for Design and Implementation

Based on our findings, we suggest the following guidelines for designing and using a standards-based teacher evaluation system as the basis for a KSBP system.

1. Specify that performance improvement is a strategic imperative. This will first require identification of performance gaps (e.g., student learning relative to state proficiency standards, achievement gaps), followed by a conclusion that improvement in teachers' instructional practice will be a key lever for closing these gaps. In this way teacher performance competency becomes identified as a factor of strategic importance, and a standards-based evaluation system and a KSBP system will logically follow as key tools to be used in the drive for performance improvement. If these initiatives are not embedded within a strategy of performance improvement, they will likely be viewed as “just another program” by teachers and administrators. In turn, the evaluation system will be lost among other priorities that come along, not have a designated champion for its success, be underfunded in both monetary and time terms, meet resistance or rejection by teachers and administrators, and gradually lose its potency.

2. Develop a set of teaching standards and scoring rubrics (i.e., a competency model) that reflects what teachers need to know and be able to do to provide the kind of instruction needed to meet the district's student achievement goals. The model is the foundation of the program and everything about the program will flow from it. Active teacher participation in the construction and refinement of the model is essential. The Framework for Teaching represents one possible starting point, but other models such as those developed by the National Board for Professional Teaching Standards, the state of Connecticut (see http://www.state.ct.us/sde/dtl/t-a/index.htm), the National Council of Teachers of Mathematics, and the National Council of Teachers of English, as well as a new set of standards being developed by Allan Odden to reflect research-based instructional practice (see Odden & Wallace, forthcoming) should also be considered. The Connecticut and Odden models place more emphasis on specific instructional practices derived from the most current research on how students learn (e.g., Brans-

ford, Brown, & Cocking, 1999; Donovan & Bransford, 2005a, 2005b, 2005c; Cunningham & Allington, 1994).

Our research also suggests that some additions to the content of Framework-based systems may be useful in improving instruction and may also yield evaluation scores with a stronger relationship to student achievement. First, while the generic teaching behaviors emphasized in the systems based on the Framework are important, it may also be useful to explicitly evaluate teachers on their skill in implementing specific instructional programs important to the jurisdiction's strategy to improve student achievement. For example, if models like Success for All or Direct Instruction, or a specific curriculum, are part of the strategy to achieve school or district goals, teachers should be evaluated on how well these are implemented in the classroom. Second, if more skill in content-specific pedagogy and higher levels of pedagogical content knowledge are needed to facilitate a major boost in student achievement, evaluation systems may need to place more emphasis on those approaches to instruction.

3. Be prepared for additional workload for teachers being evaluated and those doing the evaluation. System designers need to carefully review what is required of teachers to minimize burden. This is especially an issue if teachers will be required to prepare a portfolio as part of the evaluation. Perhaps some small reduction in other responsibilities while teachers are undergoing evaluation would decrease the perception of burden and sense of stress. Similarly, while evaluating teachers is already a part of school leaders' jobs, doing high-quality evaluations is often not rewarded by districts nor easy to do given the current structure of the jobs. While the addition of peer evaluators can reduce administrator workload, districts should review the design of administrators' jobs and consider incentives for them to allocate more time to teacher evaluation, feedback, and coaching.

4. Prepare teachers and administrators thoroughly. Simply communicating about the system is not enough. Training will be necessary for both teachers and administrators. For teachers, early training should focus on the nature of the performance competencies on which the system is based, the purposes and mechanics of the evaluation system, and knowledge and skills needed to function effectively within the new system. A key here will be providing guidance on the specifics of what good teaching looks like according to the evaluation system, so teachers have a clear idea of what they need to do to get a good evaluation before the process starts. Administrators will also need early training on the performance competency model and on the purposes and mechanics of the new evaluation system. In addition, it will be critical to provide training in observational skills and accuracy, as well as providing timely, useful feedback and coaching. Further into the life of the new system, training can shift to broader issues of performance management and instructional leadership centered on the teacher competency model.

5. Consider using multiple evaluators if “live” observations are part of the system. The burden of effective standards-based evaluation may be too great for many school administrators to shoulder alone. Not only do they have many other demands on their time, but few can be experts in all grades and subjects, nor can all resist the temptation toward giving lenient ratings to preserve working relationships. Having a second evaluator provides expertise, reduces workload, and can help reduce leniency when scores have to be compared and discussed. Alternatively, systems could use curriculum-unit based instructional portfolios, with videos of teachers' instructional practice rather than live observations. This is the approach of the National Board, Connecticut, and the Odden and Wallace (forthcoming) proposal.

6. Provide evaluators with high-quality training. For example, Cincinnati began with a three-day session for evaluators on system goals, procedures, and rating pitfalls, and followed up with having raters view, discuss, and rate several videotapes of teaching at various performance levels. Raters had to meet a standard of agreement with the ratings of a set of experts, and received follow-up training to help them do so. The training should include the use of a structured scoring process to guide evaluator decision making and to discourage “gut-level” decisions. Clarify for evaluators issues such as what evidence is to be collected, how that evidence should be compared to the rubrics or rating scales, or how to deal with evidence that falls between two rubric categories.

7. Support teachers in acquiring the knowledge and skills needed to reach high performance. These efforts need to go beyond orienting teachers or training them in the new process to providing resources for improvement. Feedback needs to be concrete and specific, telling the teacher not only her/his rating but also exactly what prevent-

ed her/him from getting a higher score, and what specific behaviors or results would raise the score. This could be followed by information about relevant professional development, suggestions about techniques to try and whom to observe to see good performance exemplified, and even modeling of aspects of desired performance. This in turn requires that evaluators be trained in providing feedback and that teachers have a coach or mentor to go to for help. It may be necessary to get school leaders and teaching peers more involved in providing developmental feedback and coaching.

It may also be necessary to restructure professional development programs on the teaching knowledge and skills underlying the evaluation system, so that they to provide the skills teachers need to do well in both the classroom and the evaluation process. The link between specific professional development activities and the performance competencies in the standards then needs to be made clear to teachers. Having the professional development system "look" like the evaluation system also can help bring alignment. Emphasizing the development and use of standards-based curriculum units, which has been shown to be a powerful form of professional development (Cohen & Hill, 2001), aligns nicely with a curriculum-unit approach in the evaluation system.

8. Align the human resource management system with the performance competency model underlying the teacher evaluation standards. To reinforce the importance of the performance competency model and create a shared conception of competent instruction, the content of human resource programs should reflect the content of the model (Heneman & Milanowski, 2004). This applies to all eight major human resource program areas: recruitment, selection, induction, mentoring, professional development, performance management, compensation, and school leadership. For example, recruitment of new teachers can be targeted toward applicants likely to possess the competencies in the model, and applicants can be informed of the competencies expected. Another example is teacher induction. If induction programs are based on the performance competencies underlying the evaluation system, new teachers are likely to have a better understanding of performance requirements and to be better prepared for future evaluations, as well as to be less likely to leave the district or profession in response to a negative evaluation experience. Professional development needs to be aligned, so that teachers have the means to obtain the knowledge and skills rewarded by the pay system. Alignment reinforces the importance of the performance competencies, sends consistent messages about the district's vision of good teaching, and provides a framework for instructional leaders to use in helping teachers improve practice.

9. Work out details, pilot the system, and monitor implementation. We found that at least one pilot year was needed to work the glitches out of the evaluation systems. A single test year may not be enough. At some sites, going to scale after the pilot revealed implementation problems which in turn lowered the credibility of the system to teachers and reduced acceptance.

10. Conduct validity and interrater agreement (reliability) analyses. This will help assure all stakeholders that evaluation scores are based on observable and agreed-upon features of teacher performance and that higher evaluation scores are connected with important student achievement outcomes.

Many of these recommendations imply that a standards-based evaluation system should be designed, tested, and implemented before the link to pay is made. Pay change would thus follow the change in the performance evaluation system and the development of aligned human resource practices. While new pay and evaluation systems could be introduced all at once, before doing so program designers need to realistically assess their organization's capacity for implementing change in a number of major human resource systems, and the readiness of teachers for major changes in how they are evaluated and paid.

Guidelines Caveats

Several caveats about our guidelines for using standards-based teacher evaluation systems are pertinent. First, there are generalizability bounds on our research in terms of the teacher performance competency model studied (the Framework for Teaching), the heavy emphasis on classroom observation as the method of evidence gathering, and the types of training provided to teachers, administrators, and special evaluators. Alternative design and delivery features should be experimented with (for examples, see Odden & Wallace, forthcoming; Tucker & Stronge, 2005). Different performance competency models could be tried, varying features such as the number and types of standards (e.g., only ones that focus on instructional practice), or using alternative intact

competency models such as those identified above. Greater usage and weight could be accorded to portfolios and videos of teacher instruction, as opposed to classroom instruction. More intensive training, emphasizing accuracy of evaluation and feedback and coaching skills, could be given to administrators or other evaluators. Whatever the specific nature of the experimentation, evaluation of its effectiveness is paramount.

Second, the systems we evaluated, as well as our recommendations for practice, entail increased administrative workload for teachers, administrators, and human resource staff. Such an effect is a natural byproduct of serious attempts to improve teacher quality. Ways to minimize the workload effect, or help incorporate it into standard practice, should be experimented with. Suggestions here include streamlining implementation and evidence-gathering processes, automating them using web-based technologies, and standardizing and sharing processes, such as through district consortia or state-funded and conducted activities (e.g., training for administrators).

Finally, using standards-based evaluation results for KSBP systems will likely continue to generate resistance from some teachers and administrators. For some teachers, familiarity and comfort with the single salary schedule, aversion to performance pay, fears of pay fluctuations and uncertainty, skepticism about the stability and survival of funding for the pay program, and lack of self-confidence and assistance for meeting high performance standards all combine to make a new KSBP program a less than welcome addition to their educational lives. Resistance among some administrators also may run deep, particularly due to a loathing to make significant performance differentiations among teachers that will lead to significant pay differences among them. Mechanisms for lessening resistance must be incorporated into the initial design of the plan. These include communicating extensively and continually with teachers and administrators about the plan, making the plan prospective so that current teachers have the option of staying with the old plan, guaranteeing that there will be no pay cuts as the new plan is implemented and that there will be no artificial limits on the amount of pay that can be earned, ensuring stability of funding for the plan, and showing teachers the actual dollar impacts of the plan on their individual pay. The Denver Pro Comp plan incorporated all of these elements and was voted on favorably by the teachers. Many of these elements were missing from the proposed Cincinnati plan, and it was voted down by teachers.

Conclusion

Our research shows the promise of standards-based teacher evaluation as a foundation for KSBP systems. In order to make the most of this approach, moving toward standards-based evaluation should be more than a fine-tuning of the existing evaluation system. Indeed, the system should be made an integral component of a general performance improvement strategy. Then a commitment to a transformation in how teacher performance is defined, measured, and supported is needed. Such commitment needs to extend not only to the teacher evaluation process, but also to aligning the human resource management system, linking the aligned system to state or district instructional strategies, and addressing teacher and administrator apprehensions about changing the pay system. This commitment is not for the faint of will, time, or budget; it is for those who want to invest in creating a high-quality teaching force with the competencies needed to help kids learn in a standards-based world.

About the Authors

Herbert G. Heneman III is the Dickson-Bascom Professor (Emeritus) in Business, Management, and Human Resources at the University of Wisconsin-Madison. He also serves as a Senior Research Associate in the Wisconsin Center for Education Research. His research is in the areas of staffing, performance management, union membership growth, work motivation, and compensation systems. Heneman is the senior author of four textbooks on human resource management.

Anthony Milanowski is an Assistant Scientist with CPRE at the University of Wisconsin-Madison. Since 1999 he has coordinated the CPRE Teacher Compensation Project's research on standards-based teacher evaluation and teacher performance pay. His research interests include performance evaluation, pay system innovations, teacher selection, and the teacher labor market. Before coming to CPRE, he worked for many years as a human resource management professional.

Steven M. Kimball is a researcher with CPRE at the University of Wisconsin-Madison and with the Wisconsin Center for Education Research. For the CPRE Teacher Compensation Project, he has researched the impact of school-based performance award programs, National Board Certifica-

tion, and standards-based teacher evaluation and compensation systems. Before joining CPRE, Kimball held legislative analyst positions in the U.S. Congress and the Texas State Office in Washington, DC.

Allan Odden is a Professor of Educational Leadership and Policy Analysis at the University of Wisconsin-Madison. He is also a Co-Director of CPRE, where he directs the Education Finance Research Program. His research and policy emphases include school finance redesign and adequacy, effective resource allocation in schools, the costs of instructional improvement, and teacher compensation. Odden has published widely on his research. His newest book, *New Directions in Teacher Pay*, co-authored with Marc Wallace, will appear this year.

References

Bransford, J. D., Brown, A. L., & Cocking, R. (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Cohen, D. K. (1996). Rewarding teachers for student performance. In S. H. Fuhrman & J. A. O'Day (Eds.), *Rewards and reform: Creating educational incentives that work* (pp. 60-112). San Francisco: Jossey-Bass.

Cohen, D. K., & Hill, H. C. (2001). *Learning policy: When state education reform works*. New Haven, CT: Yale University Press.

Corcoran, T., & Goertz, M. (1995). Instructional capacity and high performance schools. *Educational Researcher*, *24*(9), 27-31.

Cunningham, P., & Allington, R. (1994). *Classrooms that work: They can all read and write*. New York: HarperCollins.

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

Donovan, M. S., & Bransford, J. D. (Eds.). (2005a). *How students learn: History in the classroom*. Washington, DC: National Academies Press.

Donovan, M. S., & Bransford, J. D. (Eds.). (2005b). *How students learn: Mathematics in the classroom*. Washington, DC: National Academies Press.

Donovan, M. S., & Bransford, J. D. (Eds.). (2005c). *How students learn: Science in the classroom*. Washington, DC: National Academies Press.

Floden, R. E. (1997). Reforms that call for teaching more than you understand. In N. C. Burbules & D. T. Hansen (Eds.), *Teaching and its predicaments* (pp. 11-28). Boulder, CO: Westview Press.

Gallagher, H. A. (2004). Vaughn Elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement? *Peabody Journal of Education*, *79*(4), 79-107.

Heneman, H. G., III. (1998). Assessment of the motivational reactions of teachers to a school-based performance award program. *Journal of Personnel Evaluation in Education*, *12*(1), 43-59.

Heneman, H. G., III, & Milanowski, A. T. (1999). Teachers' attitudes about teacher bonuses under school-based performance award programs. *Journal of Personnel Evaluation in Education*, *12*(4), 327-341.

Heneman, H. G., III, & Milanowski, A. T. (2003). Continuing assessment of teacher reactions to a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education*, *17*(2), 173-195.

Heneman, H. G., III, & Milanowski, A. T. (2004). Alignment of human resource practices and teacher performance competency. *Peabody Journal of Education*, *79*(4), 108-125.

Kelley, C., Heneman, H., III, & Milanowski, A. (2002). School-based performance rewards: research findings and future directions. *Educational Administration Quarterly*, *38*(3), 372-401.

Kelley, C., Odden, A., Milanowski, A., & Heneman, H. G., III. (2000). *The motivational effects of school-based performance awards* (CPRE Policy Brief No. RB-29). Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Kimball, S. M. (2002). Analysis of feedback, enabling conditions, and fairness perceptions of teachers in three school districts with new standards-based teacher 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.

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., & Heneman, H. G., III. (2001). Assessment of teacher reactions to a standards-based teacher evaluation system: A pilot study. *Journal of Personnel Evaluation in Education*, *15*(3), 193-212.

Milanowski, A. T., Kimball, S. M., & Odden, A. (2005). Teacher accountability measures and links to learning. In L. Stiefel, A. E. Schwartz, R. Rubenstein, & J. Zabel, (Eds.), *Measuring school performance and efficiency: Implications for practice and research. 2005 Yearbook of the American Education Finance Association* (pp. 137-159). Larchmont, NY: Eye on Education.

Newman, F. M., King, M. B., & Rigdon, M. (1997). Accountability and school performance: Implications from restructuring schools. *Harvard Educational Review*, *67*(1), 41-74.

Nye, B., Konstantopoulos, S., & Hedges, L.V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(4), 237-257.

Odden, A. (2000). New and better forms of teacher compensation are possible. *Phi Delta Kappan*, *81*(5), 361-366.

Odden, A., & Kelley, C. (2002). *Paying teachers for what they know and do: New and smarter compensation strategies to improve schools* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Odden, A., Kelley, C., Heneman, H., III, & Milanowski, A. (2001). *Enhancing teacher quality through knowledge- and skills-based pay* (CPRE Policy Brief No. RB-34). Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Odden, A., & Wallace, M. (forthcoming). *New directions in teacher pay*.

Peterson, K. (2006). Teacher pay reform challenges states. *Stateline.org*: *Where policy & politics news click*. Retrieved March 13, 2006 from www.stateline.org/live/viewpage.action?siteNodeId=137&languageID=15contentID=93346.

Rowan, B., Correnti, R., & Miller, R.J. (2002). *What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools* (CPRE Research Report No. RR-051). Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Tucker, P. D., & Stronge, J. H. (2005). *Linking teacher evaluation and student learning*. Alexandria, VA: Association for Supervision and Curriculum Development.

Appendix. Summary of Research Activities at the Four Sites

CINCINNATI

A. 1999-2000 Pilot Test in 10 Schools
1. Interviews with teachers and principals regarding understanding and acceptance of teaching standards, system implementation, perceptions of fairness and impacts on teaching, evaluation scheduling and time de mands.
2. Teacher survey: understanding of system, acceptance of teaching standards, fairness of evaluation process and ratings, evaluator, impact on teaching, favorableness toward new system.

B. 2000-2001: Full Implementation Throughout District
1. Interrater agreement in ratings (peer and principal).
2. Interviews with teachers and principals: standards and rubrics, portfolio, evaluator qualifications, implementation, feedback, rating fairness, reaction to peer evaluators, impact on practice, time demands and burd ens, stress, pay-at-risk.
3. Teacher survey: fairness of rating process and results, utility, accuracy of ratings, satisfaction with system, stress, effort versus benefits.
4. Teacher performance ratings as predictor of student achievement —value-added analysis for reading, math, and science test scores.
5. Human resource alignment —degree of alignment of eight human resource practice areas with teacher performance competencies.

C. 2002-2003 Follow-Up
1. Teacher survey: fairness of rating process and results, utility, accuracy of ratings, satisfaction with system, stress, effort versus benefits.
2. Interviews with teachers and principals: standards and rubrics, portfolio, evaluator qualifications, implementation, feedback, rating fairness, impact on practice, time demands and burdens, stress, benefits versus costs.
3. Teacher exit survey.
4. Teacher performance ratings as predictor of student achievement —value-added analysis for reading, math, and science test scores.

D. 2003-2004 Follow-Up
1. Teacher performance ratings as predic tor of student achievement —value-added analysis for reading, math, and science test scores.

WASHOE COUNTY (RENO/SPARKS)

A. 2000-2001 Full Implementation Throughout District (Pilot Year Not Covered by Our Research)
1. Interviews with teachers and principals: evaluation feedback, enabling conditions, fairness, acceptance of purpose and standards, acceptance of ratings, human resource alignment.
2. Teacher survey: fairness of rating process and results, utility, accuracy of ratings, satisfaction with system, stress, benefits versus costs.

B. 2001-2002 Follow-Up
1. Interviews with teachers and principal s: fairness of rating process and results, utility, accuracy of ratings, satisfaction with system, stress, teacher efficacy, principal evaluation practices.
2. Teacher performance ratings as predictor of student achievement —value-added analysis for reading and math test scores.

C. 2002-2003 Follow-Up
1. Human resource alignment: alignment of eight human resource practice areas with the teacher standards.
2. Survey of teachers and principals: understanding and acceptance of standards and rubrics, fairness of rating process and results, utility, accuracy of ratings, satisfaction with system, stress, teacher efficacy, principal evaluation practices and decision processes.
3. Teacher exit survey.
4. Teacher performance ratings as predictor of student achievement —value-added analysis for reading and math test scores.

D. 2003-2004 Follow-Up
1. Interviews with principals: evaluation practices and decision processes.
2. Teacher performance ratings as predictor of student achievement —value-added analysis for reading and math test scores

Appendix (continued). Summary of Research Activities at the Four Sites

VAUGHN CHARTER SCHOOL

A. 1999-2000 Pilot Test
1. Teacher survey: understanding and acceptance of standards and rubrics, fairness of rating process and results, utility, accuracy of ratings, satisfaction with system, stress, attitudes toward performance pay.

B. 2000-2001 Full Implementation
1. Interviews with teachers: alignment of evaluation rubrics with state standards, curriculum materials, and professional development.
2. Teacher survey: understanding and acceptance of standards and rubrics, fairness of rating process and results, utility, accuracy of ratings, satisfaction with system, stress, attitudes toward performance pay.
3. Teacher performance ratings as predictor of student achievement: value-added analysis for reading, math, and language arts test scores.

C. 2001-2002 Follow-Up
1. Teacher survey: understanding and acceptance of standards and rubrics, fairness of rating process and results, utility, accuracy of ratings, satisfaction with system, stress, benefits versus costs, attitudes toward performance pay.
2. Teacher performance ratings as predictor of student achievement: value-added analysis for reading, math, and language arts test scores.

D. 2002-2003 Follow-Up
1. Teacher survey: understanding and acceptance of standards and rubrics, fairness of rating process and results, utility, accuracy of ratings, satisfaction with system, stress, benefits versus costs, attitudes toward performance pay.
2. Teacher performance ratings as predictor of student achievement: value-added analysis for reading, math, and language arts test scores.

E. 2003-2004 Follow-Up
1. Interviews with teachers: fairness, accuracy, impact on teaching, peer evaluators.
2. Interviews with evaluators: training for evaluation, evaluation process, impact on teaching.
3. Teacher survey: understanding and acceptance of standards and rubrics, fairness of rating process and results, utility, accuracy of ratings, satisfaction with system, stress, benefits versus costs, attitudes toward performance pay.

COVENTRY

1. Teacher performance ratings as predictors of student achievement—value-added analysis for reading and math test scores.

Nondiscrimination Statement

The University of Pennsylvania values diversity and seeks talented students, faculty, and staff from diverse backgrounds. The University of Pennsylvania does not discriminate on the basis of race, sex, sexual orientation, religion, color, national or ethnic origin, age, disability, or status as a Vietnam era veteran or disabled veteran in the administration of educational policies, programs, or activities; admissions policies, scholarships, or loan awards; and athletic or University-administered programs or employment. Questions or complaints regarding this policy should be directed to Executive Director, Office of Affirmative Action, 1133 Blockley Hall, Philadelphia, PA 19104-6021 or 215-898-6993 (Voice) or 215-898-7803 (TDD).

About CPRE

The Consortium for Policy Research in Education (CPRE) studies alternative approaches to education reform in order to determine how state and local policies can promote student learning. Currently, CPRE's work is focusing on accountability policies, efforts to build capacity at various levels within the education system, methods of allocating resources and compensating teachers, instructional improvement, finance, and student and teacher standards. The results of this research are shared with policymakers, educators, and other interested individuals and organizations in order to promote improvements in policy design and implementation.

CPRE unites five of the nation's leading research institutions to improve elementary and secondary education through research on policy, finance, school reform, and school governance. Members of CPRE are the University of Pennsylvania, Harvard University, Stanford University, the University of Michigan, and the University of Wisconsin-Madison.

CPRE Policy Briefs are published by CPRE. To learn more about CPRE research or publications, please call 215-573-0700 or access CPRE publications at www.cpre.org; www.wcer.wisc.edu/cpre/; or www.sii.soe.umich.edu.

Policy Briefs

Graduate School of Education
University of Pennsylvania
3440 Market Street, Suite 560
Philadelphia, PA 19104-3325

NON PROFIT
U.S. POSTAGE
PAID
PERMIT NO. 2563
PHILADELPHIA, PA