Abstract Title Page
Not included in page count.

Title:
Redesigning teacher evaluation: Lessons learned from a pilot implementation in New Hampshire

Authors and Affiliations:
Julie Riordan, Education Development Center
Natalie Lacireno-Paquet, WestEd
Karen Shakman, Education Development Center
Candice Bocala, WestEd

Abstract Body

Background/Context:

Many studies have called attention to the limitations of current teacher evaluation systems and the need for reform nationwide (Gordon, Kane, & Staiger, 2006; Heneman, Milanowski, Kimball, & Odden, 2006; Measures of Effective Teaching Project, 2012; Toch & Rothman, 2008; Weisberg, Sexton, Mulhern, & Keeling, 2009). These studies have critiqued teacher evaluation systems for neither differentiating among teachers and the quality of their instruction nor emphasizing teachers' influence on student achievement (Daley & Kim, 2010; Measures of Effective Teaching Project, 2010; Weisberg et al., 2009). Driven by federal policies and incentives, including Elementary and Secondary Education Act waivers, School Improvement Grants (SIGs)[2] , and Race to the Top grant requirements, increasing numbers of state policymakers are changing teacher evaluation policies to impose more frequent evaluations and greater rigor in evaluation measures—for example, assessments of student achievement growth that aim to measure teacher contributions to student learning. In 2009 only 14 states required annual teacher evaluations, but by 2012 that number had increased to 23, and by 2012, 43 states required annual evaluations of all new teachers (National Council on Teacher Quality, 2012).

As redesigned teacher evaluation systems have emerged across the country, recent studies have begun to examine their effectiveness, reliability, and validity. But most of the empirical studies have focused on the reliability or performance of specific instruments; few have documented their implementation. Studying implementation is important because local context influences outcomes and because implementation may reshape policy in practice (see for example, Fowler, 2004, and McLaughlin, 1990).

Purpose/Objective/Research Question/Focus of Study:

This study addresses three research questions: (1) What are the features of the new teacher evaluation systems in New Hampshire's districts with SIG schools? (2) To what extent did schools implement the evaluation system as intended? And (3) What factors affected implementation during the pilot year?

Setting:

New Hampshire introduced new, more rigorous teacher evaluation guidelines for districts in 2011. Under guidance from the New Hampshire Department of Education, districts with schools that had received SIG funding were asked to design new teacher evaluation systems in the 2011/12 school year.[3] The systems were developed and then piloted in 2012/13 in the state's 15 SIG schools. These 15 schools were located in eight districts across the state.

Population/Participants/Subjects:

The study included a sample of 35 evaluators and 277 teachers who responded to surveys about their experiences and perceptions of the evaluation system. It also included a small sample of

[2] School improvement grants (SIGs) are federal funds distributed by states to local educational agencies to provide financial assistance for school improvement activities. In awarding SIG grants, the states must give priority to the lowest-achieving schools that also demonstrate the greatest need for the funds and a strong commitment that the funding will be used toward meeting school improvement goals.

[3] The U.S. Department of Education provides funds to state education agencies to award district subgrants for local school improvement efforts. These grants are provided to the lowest achieving schools in each district.

district administrators (5), principals (8) and teachers (6) who participated in semi-structured interviews.

Intervention/Program/Practice:
The New Hampshire Task Force on Effective Teaching created a Blueprint for Effective Teaching (New Hampshire Department of Education, 2011) in October 2011. The blueprint identifies four pillars of effective teaching: preparation, induction and mentoring, professional development, and evaluation. It provides a framework for evaluation while allowing for local flexibility in the design. However, districts with SIG school were required to design new systems that (1) were based on a framework that includes at least three components: classroom environment, instruction, and professional responsibilities; (2) used a four-point performance rating scale; (3) included different teacher tracks for different levels of experience; and (4) used multiple measures, including student learning objectives (SLOs).

Research Design:
The study used a variety of methods to address the research questions including document review of district evaluation plans; descriptive statistics to analyze evaluator and teacher surveys and thematic coding of interviews.

Data Collection and Analysis:
Data for this study came from district administrative guidance documents and other administrative data, including evaluation plans and instruments; survey data from evaluators and teachers about evaluator experiences and perceptions about the evaluation system; and interview data from district administrators, principals, and teachers. All data were from the 2012/13 school year. Evaluators included any administrative staff responsible for conducting teacher evaluations. In some schools only principals were evaluators; in others, other administrators, such as assistant principals, shared this responsibility.

Researchers analyzed district evaluation plans and other documentation from all eight districts with SIG schools and created a series of tables to compare various features of the plans. They then compared the documented features against the reported use of the features from teacher surveys to create an index of implementation fidelity for each district. To analyze factors affecting implementation, researchers thematically analyzed survey responses on evaluator and teacher perceptions about the new evaluation system. Findings from survey analysis were supplemented with the interview data.

Findings / Results:
Comparison of features of the teacher evaluation system

- Summative rating scales were similar across districts.
- All districts proposed to employ the Danielson Framework for Teaching, although domain components and weighting varied.
- The frequency of evaluations and the specific measures on which teachers were rate depended on teacher years of experience and tenure.
- The weight given to measures of student learning varied the most in the district plans, ranging from 5–20 percent of a teacher's summative rating.

Fidelity of implementation

Although the measurement of fidelity has multiple dimensions, including adherence, exposure or coverage, quality of program delivery, participant responsiveness, and program differentiation (Carroll et al., 2007; Dane & Schneider, 1998; Knoche, Sheridan, Edwards, & Osborn, 2010). this study examined only exposure or coverage. Researchers compared each district plan's required features with teachers' exposure to the features, as reported in the teacher survey.

Implementation fidelity ranged from moderate to high. Overall implementation fidelity was high (80 percent or higher) in three districts and moderate (60–79 percent) in the other five (insert table 3). Average fidelity was around 74 percent. Implementation fidelity to specific features was lowest for classroom artifacts (about 49 percent)3 and highest for SLOs (almost 89 percent), which are required in all districts.

Factors related to implementation

Capacity: Scheduling, paperwork, and personnel

Many evaluators and teachers reported that the new evaluation systems took too long to complete and that there were too few evaluators to complete the required number of teacher evaluations. About 70 percent of evaluators and 62 percent of teachers reported that the system required too much time to implement. Interview data revealed more information about evaluators' and teachers' time limitations. From the principal perspective, the system required considerable time to schedule and conduct classroom observations, walkthroughs, and conferences; compile the results from multiple measures for each teacher; and complete and maintain paperwork for all teachers. Teachers commented that it was cumbersome and time-consuming to complete paperwork and prepare for meetings with evaluators.

Training

Training of evaluators was another factor related to implementation. While the state provided training support early in the summer before implementation in the following school year, especially for the Danielson Framework for Teaching,[4] classroom observations, and calibration of evaluations, not all evaluators participated in this training. Evaluators who had participated in any trainings reported higher levels of preparedness to implement the features on which they had received training than evaluators who had not participated in trainings (insert figure 1).

Student learning objectives

Principals reported that SLOs were more challenging to implement than other features of the new evaluation system (for example, conferences, walk-throughs, development of professional growth plans, etc.), particularly because student measures of learning had not previously been part of the evaluation process. Incorporating SLOs required considerable time, resources, and training to implement this new component of teacher evaluation, in part because there is little empirical research about the statistical properties of SLOs or their use to measure student growth as a component of teacher evaluation (Gill, Bruch, & Booker, 2013).

[4] The Danielson Framework for Teaching is a set of 22 components of instruction (for example, setting instructional outcomes) that are aligned to the Interstate Teacher Assessment and Support Consortium standards. The components are divided across four domains of teaching responsibility: planning and preparation, classroom environment, instruction, and professional responsibilities. All SIG schools in the study employed the Danielson Framework and its domains, although components and weighting varied.

Stakeholder support

The majority of teachers and evaluators supported the evaluation system, with 83 percent of evaluators and 69 percent of teachers reporting that they think the evaluation system is fair (figure 2). Similarly, 74 percent of evaluators and 71 percent of teachers indicated that the teacher unions in their districts support the new evaluation systems. And 89 percent of evaluators and 87 percent of teachers reported that teachers in schools are complying with the new evaluation requirements. However, teachers and evaluators did not have the same level of agreement in their perceptions of the long-term benefits of the new evaluation systems: 67 percent of evaluators and 45 percent of teachers believed that the new evaluation system would result in accurate ratings of teachers.[5] Similarly, 83 percent of evaluators and 54 percent of teachers think that the new system will improve teaching (insert figure 2).[6]

Teacher support for the new evaluation system seems to be related with implementation fidelity. The three districts with the highest average fidelity also had the highest means on the survey for fairness/compliance and support of desired implementation outcomes. However, it is unknown whether higher stakeholder support facilitated higher implementation fidelity or whether higher implementation fidelity led to higher stakeholder support.

Professional climate

A fifth factor related to implementation was the professional climate of schools. The teacher survey used in the study included items designed to measure perceptions of professional climate. These items were adapted from the Chicago Consortium for School Research (2012) survey on school climate. It included constructs of leadership, teacher influence, and trust among peers and leaders. Schools with a more favorable climate—for example, schools in which teacher trust in administrators and influence in school-level decisions was high—had greater implementation fidelity (defined as the percentage of teachers in a district that reported being evaluated on the required features of the system).

Conclusions:

Findings from this study suggested the following implications for policy: The need to assess and address capacity issues for evaluators; provide adequate planning time or introduce components incrementally to support the implementation of new and complex initiatives; allow for adequate early and ongoing training on the system; provide additional training for SLOs; engage stakeholders. Foster a positive professional climate.

The research team met with the Commissioner and the New Hampshire Department of Education's Advisory Committee for the study to discuss these findings and implications for policy. The group discussed several opportunities to share findings with the wider community in the state, including: (1) a briefing during a quarterly SIG meeting; (2) a discussion of findings with the Commissioner's extended cabinet for the purpose of preparing a document that would describe the state department of education response to findings; (3) presentations to both the principals and superintendents associations in the state; and (4) a meeting with IHE group, Professional Standards Board, and Council of Teacher Education .

[5] This difference was statistically significant ($p < .05$).

[6] This difference was also statistically significant ($p < .05$).

Appendices

Appendix A. References

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science, 2*(40). Retrieved August 22, 2013, from http://www.biomedcentral.com/content/pdf/1748–5908–2-40.pdf

Chicago Consortium for School Research. (2012). *Survey documentation and measure information and statistics*. Retrieved March 28, 2014, from https://ccsr.uchicago.edu/surveys/documentation

Daley, G., & Kim, L. (2010, August). *A teacher evaluation system that works* (Working Paper). Santa Monica, CA: National Institute for Excellence in Teaching. http://eric.ed.gov/?id=ED533380

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, *18*(1), 23–45.

The Danielson Group. (2013). *The Framework for Teaching.* Princeton, NJ: Author. Retrieved July 29, 2013, from http://www.danielsongroup.org/article.aspx?page=frameworkforteaching

Danielson, C. (2011). *The Framework for Teaching evaluation instrument.* Princeton, NJ: The Danielson Group. Retrieved July 29, 2013, from http://www.danielsongroup.org/article.aspx?page=FfTEvaluationInstrument

Fowler, F. (2004). Looking at policies: Policy instruments and cost effectiveness. In F. Fowler (Ed.), *Policy Studies for Educational Leaders: An Introduction* (2nd ed.) (pp. 238–268). Upper Saddle River, NJ: Pearson: Merrill Prentice Hall.

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job.* Washington, DC: The Brookings Institution. Retrieved November 17, 2010, from http://www.brookings.edu/~/media/Files/rc/papers/2006/04education_gordon/200604hamilton_1.pdf

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 (*Policy Brief No. RB-45). Philadelphia: Consortium for Policy Research in Education. http://eric.ed.gov/?id=ED493116

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel.* Seattle, WA: Bill & Melinda Gates Foundation. http://eric.ed.gov/?id=ED540957

Knoche, L. L., Sheridan, S. M., Edwards, C. P., & Osborn, A. Q. (2010). Implementation of a relationship-based school readiness intervention: A multidimensional approach to fidelity measurement for early childhood. *Early Childhood Research Quarterly, 25*(3), 299–313.

McLaughlin, M. W. (1990). Embracing contraries: Implementing and sustaining teacher evaluation. In J. Millman & L. Darling-Hammond (Eds.), *The New Handbook of Teacher Evaluation* (pp. 403–415). Newbury Park, CA: Sage.

Measures of Effective Teaching Project. (2010). *Learning about teaching: Initial findings from the measures of effective teaching project.* Seattle, WA: Bill & Melinda Gates Foundation. http://eric.ed.gov/?id=ED528388

Measures of Effective Teaching Project. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill & Melinda Gates Foundation. http://eric.ed.gov/?id=ED540961

Measures of Effective Teaching Project. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study.* Seattle, WA: Bill & Melinda Gates Foundation. http://eric.ed.gov/?id=ED540958

National Council on Teacher Quality. (2012). *State of the states 2012: Teacher effectiveness policies.* Washington, DC: Author. http://eric.ed.gov/?id=ED536371

New Hampshire Department of Education. (2011). *New Hampshire Task Force on Effective Teaching: Phase 1 report.* Concord, NH: Author. Retrieved June 25, 2012, from http://www.education.nh.gov/teaching/documents/phase1report.pdf

New Hampshire Department of Education. (2012). *School Improvement Grant (SIG) criteria and review for Teacher Evaluation System.* Concord, NH: Author. Retrieved June 15, 2012, from http://www.education.nh.gov/instruction/integrated/documents/teacher_effectiveness_evaluation_system_rubric_final.doc

Toch, T., & Rothman, R. (2008). *Rush to judgment: Teacher evaluation in public education* (Education Sector Reports). Washington, DC: Education Sector. Retrieved March 10, 2011, from http://www.educationsector.org/sites/default/files/publications/RushToJudgment_ES_Jan08.pdf

U.S. Department of Education. (2011). *Guidance on fiscal year 2010 school improvement grants.* Washington, DC: Office of Elementary and Secondary Education. Retrieved June 25, 2012, from http://www2.ed.gov/programs/sif/sigguidance02232011.pdf

Appendix B. Tables and Figures

Table 3. Percentage of teachers who reported experiencing each required element, by pilot district, 2012/13

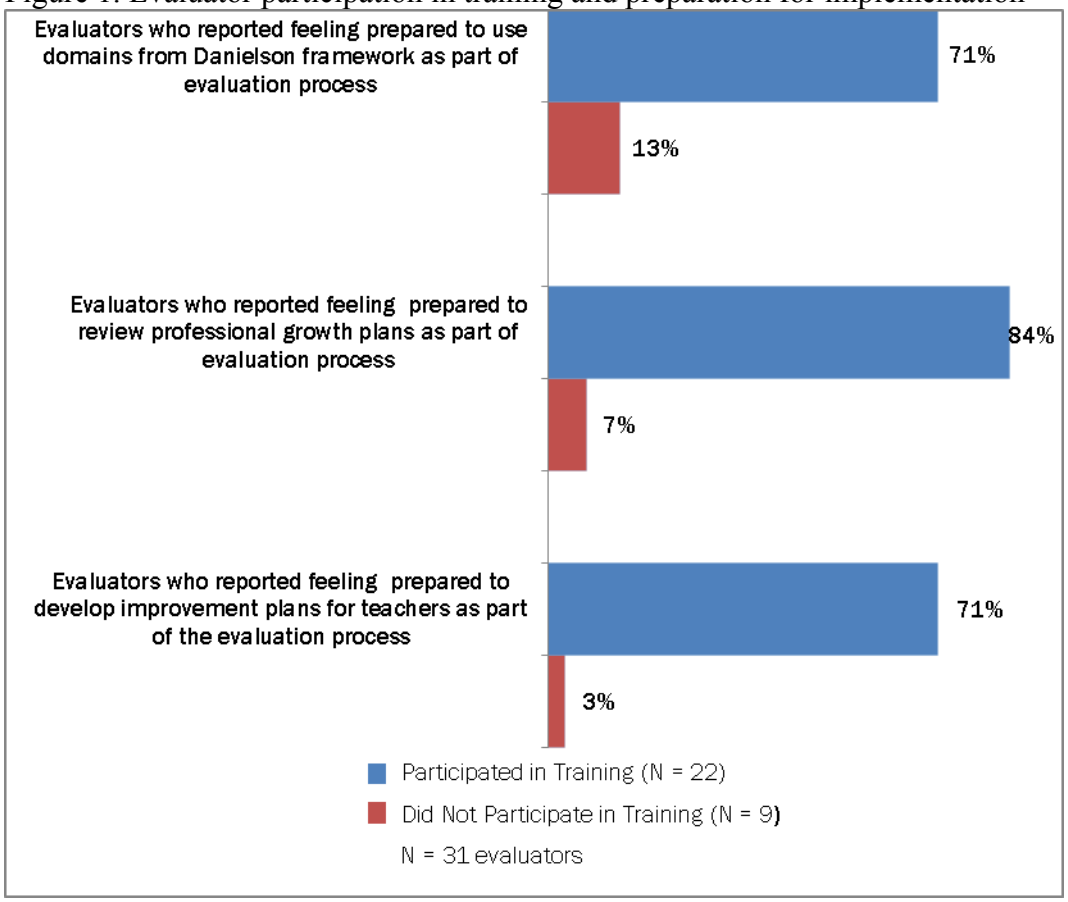
Required element	District								
	A	B	C	D	E	F	G	H[a]	Total
Formal classroom observation	100.0	82.4	100.0	72.0	55.6	56.7	83.3	81.8	79.0
Pre-/post-conference	100.0	65.7	90.5	79.0	75.0	58.3	66.7	45.5	72.6
Walkthrough	100.0	61.8	90.5	100.0	60.0	72.6	38.9	na	74.8
Classroom artifacts	36.0	44.1	19.1	87.5	na	24.1	33.3	100.0	49.2
Teaching portfolio	na	na	28.6	88.0	82.9	na	33.3	90.9	64.7
Self-assessment	81.8	na	na	83.3	na	70.4	77.8	100.0	82.7
Professional growth plan	100.0	79.4	90.5	82.6	77.1	83.1	55.6	81.8	81.3
Student learning objectives	100.0	94.1	100.0	91.7	62.9	90.2	89.0	90.9	89.8
Mean (implementation fidelity)	88.3	71.2	74.2	85.5	68.9	65.1	60.0	84.4	74.3

na is not applicable because the district does not require the element.

a. Results should be interpreted with caution due to a low response rate to the teacher survey in District H (39 percent)

Source: Authors.

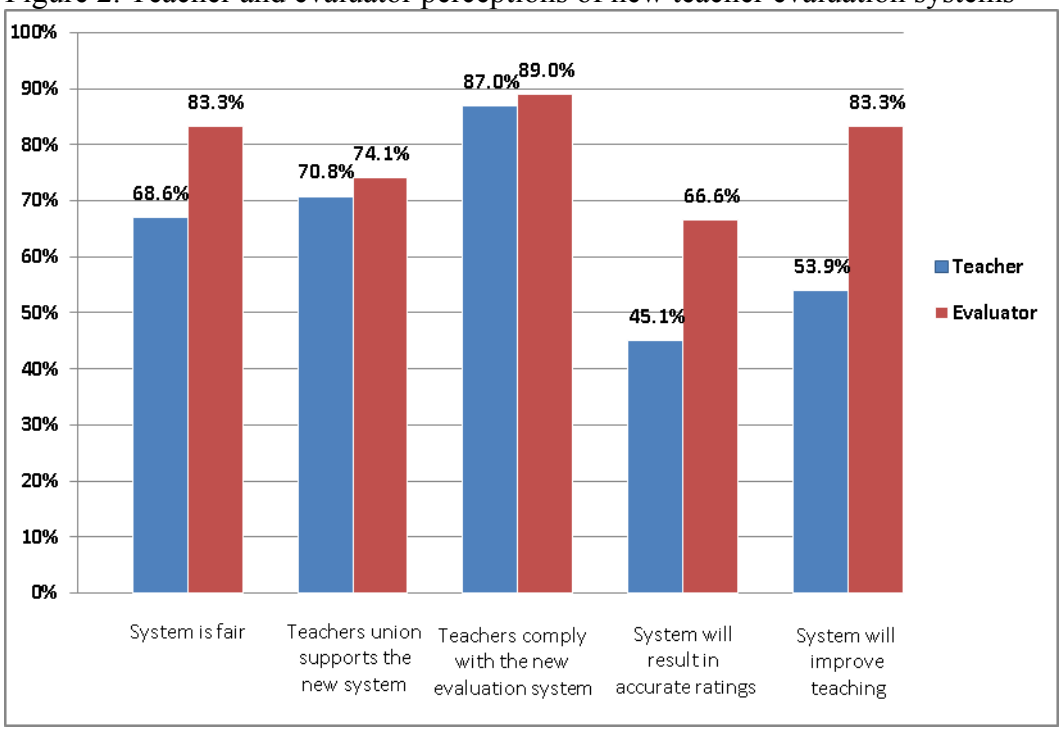
Figure 1. Evaluator participation in training and preparation for implementation



Note: The sample included 31 evaluators.

Source: Authors' analysis of New Hampshire Department of Education teacher evaluator data, 2013.

Figure 2. Teacher and evaluator perceptions of new teacher evaluation systems



* Differences between teachers and evaluators are statistically significant at the $p < .05$ level.

Source: Authors' analysis of New Hampshire Department of Education teacher and evaluator survey data, 2013.