Evaluating Principals Work: Design Considerations and Examples for an Evolving Field

Steven M. Kimball  
*University of Wisconsin–Madison*

Matthew Clifford  
*The American Institutes for Research*

Jenni Fetters  
*The American Institutes for Research*

Jessica Arrigoni  
*University of Wisconsin–Madison*

Kim Bobola  
*The American Institutes for Research*

March 2012
The Teacher Incentive Fund (TIF) would like to thank the following people for their time in reviewing drafts of this resource: Shen Lee, Westat; Jenna Aurand Scott, Westat; Carolyn Lampila, U.S. Department of Education.

The work described in this paper was supported by the U.S. Department of Education through the Teacher Incentive Fund. The opinions expressed are those of the authors and do not necessarily reflect the view of the U.S. Department of Education or the institutional partners. Comments and suggestions are welcome.

In 2010, the U.S. Department of Education awarded the Teacher Incentive Fund contract to Westat—in partnership with Learning Point Associates, an affiliate of American Institutes for Research; Synergy Enterprises, Inc.; J. Koppich and Associates; and the University of Wisconsin.

The primary purpose of the Teacher Incentive Fund (TIF) is to support TIF grantees in their implementation efforts through provision of sustained technical assistance and development and dissemination of timely resources.

This work was originally produced in whole or in part with funds from the U.S. Department of Education under contract number ED-ESE-10-0057. The content does not necessarily reflect the position or policy of TIF or the Department of Education, nor does mention or visual representation of trade names, commercial products or organizations imply endorsement by TIF or the federal government.

Allison Henderson, Director

Phone: 888-202-1513
E-mail: TIF3@westat.com
Introduction

Principal evaluation is at the forefront of the national education policy agenda, joining teacher evaluation as a means to accurately and effectively determine and support educator effectiveness. Federal initiatives such as the American Recovery and Reinvestment Act of 2009, Race to the Top (RTT), the School Improvement Grant (SIG) program, the Teacher Incentive Fund (TIF), and the waiver criteria for the Elementary and Secondary Education Act (ESEA) have prompted many states and school districts to develop new educator performance evaluation systems. However, district and state policymakers are creating these new systems without a strong research base to guide them. Notably, few rigorous research studies are currently available on the quality or impact of advanced principal evaluation systems.

State and district policymakers have raised questions about evaluation design to the TIF technical assistance team. This paper addresses the following common design questions on principal evaluation measurement, implementation, and support raised by state, district, and school leaders:

• Which leadership standards should drive evaluation design?
• How can states, districts, and charter organizations apply broad leadership standards to specific measures of principal performance?
• What are appropriate outcomes to use for principal evaluation?
• What constitutes an observation of principal performance?
• Should weights be different across measures when determining a summative evaluation score?
• How do evaluators combine measures to create an overall effectiveness rating?

Answers to these questions draw upon research on leadership effectiveness, research on leadership evaluation from within and beyond the education field, and features of principal evaluation design from states and districts that have substantively addressed these questions. This paper highlights practitioner experiences from profiles of state, district, and nonprofit education organizations. These examples illustrate possible solutions to the questions raised above, and readers are encouraged to critically examine research and practices cited in this paper for rigor and applicability to their contexts.\(^1\)

There are five main sections to this paper. The first section addresses leadership effectiveness literature and the documented need for principal evaluation

\(^1\) We used a purposive sample to select the sites for the case profiles in order to get a mix of states, districts, and nonprofit organizations that are either involved in the TIF grant or have developed new principal evaluation systems outside of TIF. We first reviewed documents available from organization websites, then carried out paper phone interviews to clarify questions about the documents and learn about any challenges confronted during the design and implementation process.
improvement. The second section includes the key components of principal evaluation systems, particularly those designed to meet the TIF requirements, supplemented with examples from innovative school districts and states. This section addresses challenges reformers often confront in the design process and solutions for some of those challenges. The third section summarizes the critical areas of evaluator training and pilot testing. The fourth section highlights stakeholder engagement and communication. The fifth section summarizes training and pilot testing. This work is not an exhaustive review or guide for principal evaluation design. Importantly, there are other sources and links available for the reader to reference in the pertinent sections of this paper.

Impetus for Principal Evaluation Reform

School principal evaluation is an education policy concern because principals have a strong impact on school outcomes. Research has demonstrated that, after teaching, school leadership is the second most influential school-level factor on student achievement (Hallinger & Heck, 1998; Leithwood, Louis, Anderson, & Wahlstrom, 2004; Waters, Marzano, & McNulty, 2003). Although research indicates leadership and teaching are important school-level factors, other factors, including student health and socio-economic status, parental education levels, and family literacy habits, may be more influential on student achievement.

Rather than directly working with students, principals influence student achievement indirectly by creating conditions for better teaching and learning to occur. For example, principals influence the degree to which schools implement instructional and other programs with fidelity (Halverson & Clifford, 2006; Seashore-Louis, Leithwood, Wahlstrom, & Anderson, 2010), and principals also shape teacher working conditions by acting as school-level human capital managers who may have the power to oversee school teacher hiring, placement, evaluation, and professional learning (Kimball, 2011; Milanowski & Kimball, 2010). However, not all principals have the same capacity to influence school conditions, instruction, or student achievement. District and state policies, school level (i.e., elementary, middle), and other factors shape principals’ roles and abilities to lead.

Independently conducted meta-analyses have identified certain principal practices associated with improving student achievement that might serve as a starting point for principal evaluation design. These practices include:

- Creating and sustaining an ambitious, commonly accepted vision and mission for organizational performance;
- Engaging deeply with teachers and using data on issues of student performance and the quality of instructional services;
- Ensuring programs and policies are adequately resourced and implemented;
- Efficiently managing resources such as human capital, time, and funding;
- Creating physically, emotionally, and cognitively safe learning environments for students and staff;
- Developing strong and respectful relationships with parents, communities, and businesses to mutually support children’s education; and
- Acting in a professional and ethical manner (Hallinger & Heck, 1998; Leithwood et al., 2004; Marzano, Walters, & McNulty, 2005; Murphy, Elliott, Goldring, & Porter, 2006; Stronge, Richard, & Catano, 2008).

Davis and colleagues (2011) note that principal effectiveness research continues to evolve and that

---

2 References to the TIF program requirements throughout this document pertain to the Notice of Final Priorities published on May 21, 2010, and other 2010 competition documents.
more sophisticated research methodologies have the potential to identify additional leadership practices that are influential on school conditions and student achievement.

Although principals can influence school organizations and student performance, principals commonly report that they are underprepared to do so (Davis et al., 2011) and that they have few opportunities to advance their practice in light of trusted feedback. Performance evaluation has the potential to support professional growth, hold principals accountable to established professional standards and performance expectations, and inform district and state reform agendas (Orr, 2011).

Despite the potential usefulness of performance evaluation, few research or evaluation studies are available on the design or effects of performance evaluation on principals, schools, or students (Clifford & Ross, 2011). However, those available indicate the need for a break from past practices. Current research suggests that:

- Principals view performance evaluation as having limited value for supporting professional learning or raising their sense of accountability (Portin, Feldman, & Knapp, 2006).
- Principal evaluations are inconsistently administered, which raises concerns about evaluation accuracy and impact (Kimball, Milanowski, & McKinney, 2009;). 
- Performance evaluations may not align with existing state or national professional standards for practice (Heck & Marcoulides, 1996) or standards for personnel evaluation (Goldring et al., 2009).
- Few widely available principal evaluation instruments display psychometric rigor or make testing public so that validity and reliability can be examined (Goldring et al., 2009; Heck & Marcoulides, 1996).

Research suggests that many principals do not experience a performance evaluation “system,” and that principal evaluations may not produce results that accurately assess their abilities or assist with their development (Clifford & Ross, 2011; Davis et al., 2011; Heck & Marcoulides, 1996; Portin et al., 2006).

**Building Blocks for Comprehensive Principal Evaluation**

The federal government requires comprehensive principal evaluation reforms under RTT, SIG, and the TIF program. According to 2010 TIF program requirements, grant recipients must either have or develop principal evaluation systems that are rigorous, transparent, and fair. The evaluation systems must clearly differentiate effectiveness using multiple rating categories. Specifically, the principal evaluation process must:

1. Use an objective, evidence-based rubric aligned with professional leadership standards and the local education agency’s coherent and integrated approach to strengthening the educator workforce;
2. Provide for observations of each principal at least twice during the school year by individuals (who may include peer reviewers) who receive specialized training;
3. Incorporate the collection and evaluation of additional forms of evidence (e.g., school climate surveys, parent surveys, school improvement plan progress reports); and
4. Ensure a high degree of inter-rater reliability (i.e., two or more raters who score a principal approximately the same.
To obtain flexibility under the ESEA, states have also embraced principal evaluation system changes that include similar requirements. This section addresses these requirements and uses examples provided from TIF grant recipients and other sites that have pursued comprehensive principal evaluation reforms.

**Evaluation System Goals and Uses**

Districts and states should articulate specific evaluation purposes in addition to compliance with grant requirements. Articulating evaluation system purposes, including formative and/or summative uses, will help in the identification of leadership standards and performance measures. In the process of determining the uses of evaluations, it is critical that states, districts, or other organizations focus on key leadership actions that will affect important educational outcomes such as improved teacher and principal quality, staff retention, distribution of educator effectiveness, and student learning (Clifford et al., 2012b).

Principal evaluation can provide formative performance feedback and that can be its primary purpose. Systems designed for this purpose could include principal self-reflection on performance standards; self-directed activities; work with a coach, mentor, or peer group; and the identification of professional development or tasks that will lead to improvement on the performance standards. These low-stakes decisions may allow for flexibility in how the measures are applied and the rigor of evaluation training.

Principal evaluation can also inform summative decisions about principal performance, which can affect continued employment and performance pay. Summative systems have higher stakes and require a higher level of rigor on measurement design and evaluator training. In addition, higher stakes requires a thorough pilot test to work out issues with the system and to examine evaluation reliability, validity, and impact. This paper addresses training and pilot testing in more detail later.

In most cases, principal evaluation has features of both formative and summative assessment. To establish fairness and increase impact, systems must include opportunities for principals to improve on performance standards and measures before principals receive their final rating. Following the formal performance assessment, professional development may be required. Notably, TIF sites must demonstrate that their educators receive training so that they understand the features of the performance measurement system and that the evaluation results inform professional development.

**Anchoring to Leadership Standards**

After identifying evaluation purposes, the next important step in the design of principal evaluation is to determine the leadership standards used to measure and support principal performance. The TIF program calls for an “evidence-based rubric aligned with professional leadership standards.” A number of recent versions of leadership standards can inform the design process. Standards from professional associations or consortia include the Interstate School Leader Licensure Consortium standards (CCSSO, 2008), the National Association of Elementary School Principals, and the National Board for Professional Teaching Standards. Other sets of leadership standards are also available for adoption and modification (e.g., Reeves, 2009; Waters et al., 2003). These standards draw from the research base and expert opinion on the behaviors, knowledge and skills that school leaders need to positively affect school conditions and outcomes.
Most states and many districts have adopted a set of leadership standards, with a large number built on the Interstate School Leaders Licensure Consortium (ISLLC) standards. However, these standards serve primarily as policy documents written to inform pre-service training programs, licensure, individual or district professional development, and other policies. They were not designed primarily to assess principal practice and do not typically include rubrics that can measure principal performance according to the standards. Although there are many examples of leadership standards, evaluation design teams will need to narrow in on key leadership standards to assess and develop rubrics that accurately measure principal performance, as illustrated by the following example.

The State of Indiana examined a number of examples of leadership standards and rubric formats as it developed a state model for districts to use in principal evaluations. The Indiana Rise approach places an emphasis on two main domains of leadership: Teacher Effectiveness and Leadership Actions. This emphasis is intentional in that the state is indicating that these two areas of leadership practices are paramount and focus “…on evaluating the principal’s role as driver of student growth and achievement through their leadership skills and ability to manage teacher effectiveness in their buildings” (Indiana Department of Education, 2011, p. 4). Indiana districts can adapt other standards that focus on building or operational management, but every district using the state model must focus on the two domains. Within each main domain are several indicators or standards of practice. The domain and standards are: (1) Teacher Effectiveness—Human Capital Manager, Instructional Leadership, and Leading Indicators of Student Learning and (2) Leadership Actions—Personal Behavior, Building Relationships, and Culture of Achievement. Evaluators use a four-level rubric to assess principal performance on a number of indicators under each standard. (For more information on Indiana’s approach, see http://www.riseindiana.org.)

Many other states and districts have gone through the process of identifying leadership standards and developing rubrics to help assess performance on the standards. In some districts, these are the primary tools used in principal evaluation (Kimball et al., 2009). These leadership standards center principal evaluation systems and can also provide the foundation of an aligned district human capital management system for principals (Kimball, 2011; Milanowski & Kimball, 2010). The evaluation standards spell out the key knowledge, skills, and attributes (also called competencies) principals should bring to their role or develop on the job. Internal assessments using those standards of principal competencies can then inform the selection, induction, development, performance management, and compensation (i.e., core elements of human capital management systems) of the school leaders carrying out district and state school improvement strategies. The extent to which leadership competencies are applied across human resource functions refers to human resource alignment, which can be assessed for principals using a process described by Heneman and Milanowski (2011).

**Practice Measures**

The next step is to identify the aspects of principal practice for use in performance assessment. Research on leader effectiveness (Grissom & Loeb, 2009) and teacher effectiveness measures (Gates Foundation, 2012) indicates the need for multiple forms of evidence to identify effective educator practices. A host of possible evidence sources can help measure principal practice. They typically fall in five categories: document review, direct observations, surveys, interviews, and principal evaluation conferences. Evaluation designers need to identify the most appropriate evidence sources for use in measuring practice based on the selected leadership standards and rubrics. Table 1 provides an example of how to apply these evidence sources to a leadership standard.
Table 1: Possible evidence sources for one of the six ISLLC standards

<table>
<thead>
<tr>
<th>ISSLC Leadership Domain</th>
<th>Examples of evidence</th>
</tr>
</thead>
</table>
| School Culture and Instructional Program | **Observations:** Evaluator observes principal activities such as conducting student data analysis sessions with faculty, making classroom walkthroughs, leading & engaging faculty in curriculum planning meetings, professional development sessions, & teacher evaluation conferences. Could also observe use of technology, differentiation of instruction in a sample of classrooms.  
**Documents:** School improvement plan, professional development plan, school walkthrough data, assessment plans/grading policy, working documents from curriculum alignment or data analysis sessions, student assignment and course-taking patterns, descriptions of implementation of new initiatives, written teacher evaluations.  
**Surveys:** Teacher perceptions of leadership practices (e.g., Val-Ed), working conditions survey.  
**Interviews:** Evaluator interviews a sample of staff about perceptions of the instructional leadership of principal and their opportunities to develop instructional capacity and students about motivation to learn and expectations held for them by staff.  
**Principal evaluation conference:** Evaluator & principal review instructional leadership activities such as curriculum development and monitoring, promotion of staff collaborative work, data-based decision making, differentiation of instruction. |

A number of states’ and districts’ evaluation materials list possible evidence sources for use in principal evaluation. However, the evaluation materials may not be explicit about the most important evidence to collect or how that evidence should inform assessments of practice. Clear guidance on the alignment between standards, measures, and uses can help principals and their evaluators better understand what to expect from performance evaluation, and can help establish and maintain inter-rater reliability.

The Rewarding Excellence in Instruction and Leadership initiative overseen by Maricopa County Educational Service Agency in Arizona developed a guide to practice-based evidence linked to its evaluation standards. Table 2 below summarizes evidence sources used in its principal observation process.
### Table 2: Rewarding excellence in instruction and leadership: Principal observation process

<table>
<thead>
<tr>
<th>Rubric</th>
<th>Element</th>
<th>Observation settings</th>
<th>Documentation settings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Leadership Team meeting</td>
<td>Business meeting</td>
</tr>
<tr>
<td>Setting &amp;</td>
<td>Shared purpose</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td>Communicating</td>
<td>Goal setting</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td>Direction</td>
<td>Continuous improvement planning</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td></td>
<td>Consensus-building around CIP</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td>Building</td>
<td>Accessibility</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
| Relationships   | Celebrating individual & team achievements | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ | ✓ ✓ ◆
Table 2: Rewarding excellence in instruction and leadership: Principal observation process (Cont’d)

<table>
<thead>
<tr>
<th>Rubric</th>
<th>Element</th>
<th>Observation settings</th>
<th>Documentation settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leading Instruction</td>
<td>Observing &amp; analyzing instruction</td>
<td>Leadership Team meeting √</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>Pre- &amp; post-conferencing</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>Educator goal plans</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>Supervision of written, taught, &amp; tested curriculum</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Securing Accountability</td>
<td>Accountability for goals</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>Accountability systems</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>Manages day-to-day operations</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>Manages school resources</td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>

Source: REIL/MCESA Principal Practice Evidence Sources

The Rewarding Excellence in Instruction and Leadership guide begins with the assessed leadership standards and then shows the specific elements that make up each standard. Each element links to observation settings (up to six) and documents (up to eight) for review. The relevant observation or document evidence sources change based on the leadership standard and element assessed.

Challenge of Principal Observation

Many district, state, and other principal evaluation systems are required to include principal observation, but requests for technical assistance suggest that many organizations struggle to articulate principal observation measures. TIF guidelines require at least two observations of principal practice.
Many aspects of principal practice make systematic observations of practice difficult. Principals typically work in multiple locations, such as hallways, classrooms, faculty rooms, community meetings, district offices, and, occasionally, the principal’s office. While observers have ample opportunity for observation, principals’ work often varies from day to day, which presents challenges to systematic observation. Additionally, principal responsibilities may be carried out at multiple venues based on the leadership context, thus requiring evaluators to make multiple observations. (Milanowski & Kimball, 2011).

The above example from Maricopa provided one way to chart possible observational evidence sources and relate them to leadership standards. Table 3 provides other possible scenarios for principal observation and the potential evidence available from the observation.

<table>
<thead>
<tr>
<th>Observation Venue</th>
<th>Potential Evidence About</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Faculty Meetings</td>
<td>Ability to run an effective meeting</td>
</tr>
<tr>
<td></td>
<td>Communication with staff</td>
</tr>
<tr>
<td></td>
<td>Involvement of staff in decisionmaking</td>
</tr>
<tr>
<td></td>
<td>Focus on instruction</td>
</tr>
<tr>
<td></td>
<td>School culture</td>
</tr>
<tr>
<td>Leadership Team Meetings</td>
<td>Principal’s process for setting goals, planning, using data, &amp; monitoring, relative to the focus of the meeting (e.g., curriculum planning, teacher scheduling, analysis of student test data)</td>
</tr>
<tr>
<td></td>
<td>Distribution of leadership</td>
</tr>
<tr>
<td></td>
<td>School culture</td>
</tr>
<tr>
<td>Teacher Coaching/Feedback Sessions</td>
<td>Principal understanding of and focus on instruction</td>
</tr>
<tr>
<td></td>
<td>Effectiveness as coach</td>
</tr>
<tr>
<td>Principal-led Group Professional Development Sessions</td>
<td>Principal involvement with professional development</td>
</tr>
<tr>
<td></td>
<td>Principal skill in facilitating teacher learning</td>
</tr>
<tr>
<td>Classroom Walkthroughs</td>
<td>Principal ability to observe and analyze instruction</td>
</tr>
<tr>
<td></td>
<td>Principal efforts to monitor instructional quality</td>
</tr>
<tr>
<td>IEP Meetings</td>
<td>Principal understanding of and assistance with the IEP process</td>
</tr>
<tr>
<td>School Assemblies</td>
<td>Student discipline</td>
</tr>
<tr>
<td></td>
<td>School culture</td>
</tr>
<tr>
<td>Teacher Hiring Interviews/Hiring Process</td>
<td>Principal’s use of effective hiring methods</td>
</tr>
<tr>
<td>Individual Parent Meeting</td>
<td>Principal communication skills</td>
</tr>
<tr>
<td></td>
<td>Quality of parent relationships</td>
</tr>
<tr>
<td>School Steering Committee or PTA/PTO Meetings</td>
<td>Principal communication skills</td>
</tr>
<tr>
<td></td>
<td>Level of parental input into school decisions</td>
</tr>
<tr>
<td>Parent Night or Other Group Meetings with Parents or Community Members</td>
<td>Principal communication about school program</td>
</tr>
<tr>
<td></td>
<td>Principal advocacy for children and school program</td>
</tr>
<tr>
<td></td>
<td>Quality of community relationship</td>
</tr>
</tbody>
</table>

Source: Milanowski & Kimball, 2011. Teacher Incentive Fund memo on principal observation
Several states, districts, and charter school organizations, such as the example provided below, are taking creative approaches to include principal observation in the overall evaluation of principal effectiveness. Some sites have designed observation components to assess how well principals carry out teacher evaluation tasks, seen as a critical part of both improving teaching practice and, because principals are the primary evaluators of teacher performance, generating reliable data for use in teacher effectiveness systems.

**Reach Teacher Incentive Fund Consortium**

The Reach Institute for School Leadership, along with ARISE High School and three other charter schools in their network, created a principal observation approach that focuses on principal skills in teacher observation and post-observation conferencing. The observation component uses a structured set of activities principals should go through to carry out the Reach-Arise teacher evaluation system. Each principal should go through each step, but for the formal assessment, the principals document each step.

First, principals identify a teacher to observe and log basic information about that teacher and classroom. In this step, the principal notes prior work with the teacher and any professional development that might be pertinent to her or his practice. The principal then reviews the teacher evaluation standards and rubrics and makes initial ratings based on prior evidence of the teacher’s practice, noting the specific evidence used for the initial rating.

The second step is observation of the teacher. Principals are required to assess practice using a standard observation protocol supported with related evidence. The following evidence is required: prior observations, performance in the school learning community (i.e., department or cross-departmental teams), professional development interactions, student achievement data, student assessment results, and referral patterns. Principals also collect and include any items that the teacher has prepared such as lesson plans or handouts.

Step three is the post-observation conference with the teacher. The conference revolves around the six teaching standards and includes planning for instructional coaching when necessary. Videotaping of the conference also takes place.

The final step is a formal reflection exercise. Principals reflect on the first three steps, review their plan for the conference, what actually happened during the conference, and why they think these things occurred. Principals also describe their evaluation of the teacher, including strengths, areas for growth, evidence used to assess the teacher, the next steps for the teacher to improve, and instructional leadership that would aid in teacher improvement. The six teaching standards inform the reflective piece.

Once completed, entering of the video and related documentation (i.e., principal notes and reflections) into an on-line data system takes place. External raters then score this evidence using a rubric developed to assess the observation materials. The rubric focuses on instructional leadership skills in two areas: (1) facilitating effective learning conversations, including assessing teacher’s needs, establishing a shared focus for the discussion, moving teacher practice forward, and promoting accountability and (2) instructional leadership equity, which includes awareness, effective action, analysis, and attitude change. Currently, the evaluators are leaders from the Reach Institute who receive training using the rubric and example videos and artifacts. They must reach
a standard score during training before they can formally assess observation evidence. The consortium plans to recruit principals outside of the network to score the video observations in the future. After scoring of the evidence, principals receive extensive formative feedback from a leadership coach. This occurs before their formal observation assessment, and principals also receive feedback from their evaluation following the summative review.

The above example illustrates different applications of multiple evidence sources in principal practice evaluation and, in particular, shows how principal observation is usable as a central source of practice measurement with related performance feedback. The use of multiple practice measures can help triangulate results from the practice evaluation. The practice-based rubric or other separate measures with different weights applied to different evidence sources based on important to district or state leadership priorities can inform assessments of the chosen practice measures. The next piece of the principal evaluation puzzle involves measuring school and student achievement outcomes.

Measures of Outcomes

In addition to practice measures, states and school districts receiving federal funding through RTT, SIG, and TIF must associate outcomes measures with school principal performance. An outcomes measure collects information about the degree to which a principal’s efforts have produced expected effects. Federal requirements indicate that states and districts accepting RTT, TIF, and SIG funding use multiple outcomes measures, with student growth accounting for a “significant” portion of a principals’ summative score. Our review of evaluation system designs suggests that outcomes measures fit into four broad categories: student growth, student in-school and post-school behaviors, instructional quality, and school climate. Outcomes measures include, but are not limited to:

- Student achievement measures:
  - Value-added models
  - Student achievement trends
  - Student learning objectives
  - Subject-specific test results
  - Graduation rates
- Instructional quality measures:
  - Teacher placement indicators (e.g., placement in subject area in which teachers are certified)
  - Teacher retention rates (especially for highly effective teachers)
  - Specific measures of instructional quality that are objectively determined
- School performance measures:
  - Student behavior measures (e.g., attendance, attrition, behavioral incidents)
  - School climate measures
  - Community participation, interaction, and satisfaction measures
  - Progress on school improvement plans
  - Progress on school fiscal management plans (as applicable)

Each of the school districts and states contacted to inform this paper incorporate a mix of outcomes measures for assessing principal performance. For example, the TAP system uses school-level value-added results as a large part of its evaluation approach. The Austin Unified School District uses an aggregated, school-level value-added measure for determining student learning gains and also includes measures of student behavior (incidents reports) and school climate surveys among other measures. The Reach-Arise project also includes a school achievement composite measure as part of its principal evaluation approach.

---

3 “Triangulate” is a validation method whereby multiple methods are used to gather data on the degree to which a principal meets a given standard (Clifford et al., 2012b).
The development and selection of principal outcomes measures should consider the:

- Strength and stability of the measure, which means that outcomes measures should be technically defensible (i.e., valid and reliable) and provide principals with actionable feedback.

- Application to leadership contexts that the principal can be reasonably expected to influence, given variations in principals’ authority across states and districts and the indirect nature of principals’ influence on some outcomes (e.g., instructional quality, student growth). Human and financial capacity needed to consistently measure and analyze data. Such capacity is essential to maintaining fidelity to the principal evaluation system (Clifford et al., 2012a).

Principal evaluation designers can address these challenges by selecting technically sound measures for the contexts of use and building options into evaluation systems to accommodate variation in principals’ work across contexts (Clifford et al., 2012b).

Valid principal evaluation systems must be careful to (a) assess principals on outcomes that they are responsible for attaining and (b) appropriately attribute outcomes to a principal’s work, and not to others’ work or factors beyond the school. Assessing principals on outcomes for which they are responsible means that a principal’s role, as written in job description or other documentation, includes the expected results. These outcomes may be as simple as the extent to which school finances are within budget or as complex as statistical models of student achievement or improved results on teacher professional community measures. In interviews for this paper, principal evaluation design teams noted variations in principals’ capacity to affect outcomes, which means that some outcomes measures may not be fairly applied to some principals. If school-value-added results are part of the outcomes measures, for example, it is possible that test data are not available to allow for calculation of value-added at the high school level. Other ways of assessing school achievement growth would be necessary.

Evaluation designers must select measures that assess principals on outcomes that they are responsible for attaining. For example, the Hillsborough County School District establishes decision rules for associating outcomes with principal performance due to mid-year job assignment changes. According to the Hillsborough principal evaluation decision rules, outcomes attained by a previous principal should not be included in a new principal’s summative evaluation, particularly if the new principal began leading the school at mid-year. Statistical adjustments could address the partial attribution of assessment results, but multiple years of achievement data would still be needed to improve reliability.

Principal evaluation systems designers may also wish to build differentiated evaluation systems that accommodate differences in principal job descriptions to ensure system fairness as well as validity. According to the National Association of Elementary School Principals and National Association of Secondary School Principals (forthcoming), principal evaluation systems might be tailored to (a) organizational level (e.g., high school, middle school, elementary school), (b) school performance history (e.g., turnaround), and (c) other factors.

**Combining Practice and Outcome Measures**

Principal evaluation processes are intended to yield a set of practice and outcome measures that form the basis for evaluation decisions. States and districts must determine the relative weight of the practice and outcome measures. As noted above, the federal RTT and TIF initiatives require student growth measures to constitute a “significant” factor in principal effectiveness determinations. States, districts, and other agencies have defined “significant” in
different ways, but most have interpreted the requirement to mean that student growth measures will determine about 30 percent to 50 percent of the effectiveness rating. The other 50 percent to 70 percent typically consists of a mix of practice measures that include evaluation of school improvement planning, principal observation, and school climate improvement.

Decisions about measure weight should reflect state, district, or other organizations’ definition of principal effectiveness, evaluation goals, and variation in principal responsibility (Clifford et al., 2012b). If, for example, the goal of evaluation is primarily formative, then the evaluation system might place greater emphasis on performance and observational data. As another example, turnaround leader evaluations might emphasize accomplishment of near-term school improvement benchmarks because good turnaround leaders achieve “quick wins” in schools, and these accomplishments carry more weight than other evaluation measures for turnaround leaders.

After deciding on the relative weight of practice and outcome measures, the next step is to determine whether the measures will stand alone (e.g., a performance score card) or form an overall effectiveness rating (e.g., an aggregate measure).

The private sector has used performance scorecards as a way to rate managers on the primary behaviors they should demonstrate and organizational results they produce. Some school districts have also used performance scorecards that include a mix of practice and outcome measures. Table 4 represents an adaptation of the New York City Principal Performance Report using principal practice and school outcomes. The illustration shows the use of the performance report as a score card.

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal’s Goals and Objectives</td>
<td>0–31 points, based on whether the principal met up to five academic goals and objectives set in collaboration with supervisor.</td>
</tr>
<tr>
<td>School’s Progress Report</td>
<td>0–32 points based on Progress Report grade, which includes measures of student growth (60%), student performance (25%), and school environment (15%).</td>
</tr>
<tr>
<td>School’s Quality Review</td>
<td>A 2- to 3-day school visit by superintendent or other trained reviewer, which includes: (1) principal interview, (2) classroom visits, (3) meeting with parents, (4) meeting with students, (5) meeting with teacher teams. The visit results in an overall score: Well Developed (WD), Developed (D), Proficient (P), or Underdeveloped (UD). Principals receive written feedback on strengths and areas for growth.</td>
</tr>
<tr>
<td>Compliance with legal mandates and NYCDOE policies</td>
<td>Principals assessed based on checklist developed by Compliance Office in partnership with program offices. The checklist consists of approximately 50 items across a variety of topics (budget, school transportation, English language learners [ELL], special education [SPED]).</td>
</tr>
<tr>
<td>Attention to populations with high needs</td>
<td>11 items on the compliance checklist directly link to related services for ELL and SPED students. Scores on these items are double-weighted.</td>
</tr>
</tbody>
</table>

Source: Webinar presentation by David Weiner, Deputy Chancellor, Division of Talent, Labor & Innovation, January 31, 2012, for Teacher Quality Center
Although this system is not a score card per se, each of the main elements could generate separate scores that result in different consequences. For example, if a principal did not meet all of the set goals and objectives, the principal could receive targeted professional development for the component needing support. If the schools test-based progress report was at a consistently high level, the principal may receive a performance bonus. There could be separate targets that trigger or offset awards. If the school quality review was consistently low, for example, the principal may not receive a bonus. The idea behind a scorecard is that the various performance measures have some independence from the others, potentially adding clarity to performance reports and providing actionable feedback on each element. However, as shown below, aggregating the different scores from this performance report is a possibility.

An aggregate approach distills various measures of principal practice and school outcomes into an overall rating. Many systems are using this method to compile ratings for evaluation decisions. Below is an illustration that shows the aggregate scoring system for the principal performance assessment developed by the State of Tennessee.

Source: Tennessee First to the Top program
In Tennessee, 35 percent of the principal rating relies on aggregated results from the practice assessment; 15 percent relies on the quality of his/her teacher evaluations; 35 percent relies on a schoolwide growth score; and 15 percent relies on an additional achievement measure (e.g., graduation, student attendance). Each has different metrics. First, the calculation and weighting of scores from each area takes place. Next, the weighted scores combine to form an overall effectiveness rating. The total score is compared to score bands set for each effectiveness level (i.e., < 200 is a 1, the lowest rating; 425–500 is a 5, the top rating).

Creating an aggregate measure is complicated. In an aggregate approach, scores on different measures are not interchangeable. The process of developing this type of scoring approach is not as simple as adding the results of four different measures and dividing by 4. For example, a score of a “2” on a practice-focused rubric is not necessarily equivalent to a “2” on an aggregated set of outcome measures. Practice evaluation measures are criterion-referenced (relative to set standards or criterion) while school value-added measures are norm-referenced (relative to a comparison group). The different measures have differences in variance and other measurement properties and are all subject to measurement error (e.g., factors that affect scores beyond true performance). Measurement experts should consider a complex set of decision rules as the scoring system is developed and refined.

**Process Guide**

Both evaluators and principals have to know what the expectations are for the principal evaluation process. A guide should spell out the timing and nature of each step in the evaluation process, how often formal and informal evaluations occur, types and descriptions of evidence collected and who collects the evidence, formative features, and the consequences related to summative decisions.

---

Most districts and states recognize the need for process manuals, but the depth and breadth of the process descriptions may vary. Colorado (see [http://www.cde.state.co.us/EducatorEffectiveness/PartnerSCEE.asp](http://www.cde.state.co.us/EducatorEffectiveness/PartnerSCEE.asp)) and North Carolina (see [http://www.dpi.state.nc.us/profdev/training/online-evaluation/](http://www.dpi.state.nc.us/profdev/training/online-evaluation/)) are two states that have developed or are currently working on comprehensive descriptions for educator effectiveness evaluation systems. Each state provides details on the development of the systems; standards, rubrics and evidence sources; and information on the use of evaluation results.


**Stakeholder Engagement and Communication**

Designing a comprehensive principal evaluation system presents policymakers with a series of complex and often politically sensitive choices and decisions. If educators do not have a role in the process, they may be predisposed to ignoring the system’s intent, and implementation will likely suffer. Engaging stakeholders throughout the process provides knowledge and diverse perspectives that can improve decisionmaking and, when used with transparent processes, lend legitimacy to principal evaluation systems.

Thoughtful consideration of who should be included and how to best involve stakeholders throughout the design, implementation, and revision process is well worth the time invested. The process should try to ensure that: (a) all perspectives are accounted for during the design and review of tools and processes; (b) there are multiple and varied opportunities for meaningful participation and feedback; (c) multiple

---

Presentation of Tony Milanowski at Teacher Incentive Fund papering, March 31, 2011, at US Department of Education.
strategies are used to garner and sustain buy-in through a comprehensive communication plan; and (d) changes are made to the system based on constituent feedback (Weber, 2012). To address these elements, states and districts should consider the following four strategies for promoting successful communication and stakeholder engagement.

**Strategy 1. Thoughtfully plan for stakeholder representation on steering committees and design teams**

Two general groups are involved in the principal evaluation design process: state- or district-level steering committees and design teams. These groups should include a broad representation from key stakeholder groups with well-defined tasks. For a list of key stakeholder groups often represented on design teams, see *A Practical Guide to Designing Comprehensive Principal Evaluation Systems, Component 2* (Clifford et al., 2012b). In determining the scope of work for steering committees and design teams, authorizers should clearly define whether the work will be authoritative or advisory. If the bulk of the committee’s work will involve creating advisory documents, recommendations, or general guidance, it is important that the state or district make clear from the start exactly how the committee’s advice will be incorporated into final policy decisions.

When selecting membership, planners should carefully consider which stakeholder groups have more of an oversight or guidance role versus which groups have more of a design role. For example, it might make sense to have school board members or legislative staff on an oversight committee, but have content and measurement experts on the design teams. Obtaining principals, assistant principals, and district leaders with diverse experiences in different contexts (e.g., urban/rural, low/high poverty, school levels) and expertise are other important considerations when selecting design team members.

**Strategy 2. Use targeted outreach opportunities to create a common starting point, communicate key decisions, and gather regular feedback from stakeholders**

As the design process moves forward, targeted opportunities to both communicate key decisions and/or deliverables and to garner broader stakeholder feedback should take place. Examples include face-to-face or online events, informational meetings, forums, and webinars for principals, teachers, and community members. Organizers must clearly communicate the purpose for these events in advance. Supporting material should be available at events, and, ideally, moderators should be skilled in facilitating a mixture of connecting, learning, and discussion activities. In addition, the event facilitators need to be clear from the start about the planned use of feedback and data gathered (Weber, 2012).

Targeted feedback is excellent for informing key design decisions around principal performance standards, leadership frameworks and/or performance rubrics, and supporting forms and tools. One method to improve the quality of feedback is to provide suggested questions or specific suggestions to focus stakeholders’ efforts along with the draft documents. These more informal, targeted requests for feedback can be rich sources of information for improving final products, but they may not allow for an assessment of the representativeness of the feedback. More formal means of deriving feedback, particularly on politically sensitive or complex topics, may be available through surveys or focus groups, (see, for example, Koppich, Prince, Guthrie, & Schuermann, 2009).

**Strategy 3. Develop a communication plan early in the process**

A good communication plan helps ensure the development of a common vision for principal evaluation among all stakeholders and provides a focus for engagement and communication efforts. It can also help foster understanding of objectives and key messages among target audiences and provide...
multiple avenues for stakeholder feedback (Benson-Glanz, 2010). The heart of developing a communication plan requires creating core messages about the principal evaluation system that convey the big ideas and key points about the project and should address questions, concerns, or goals of target audiences (Weber, 2012). The CECR stakeholder engagement and communication guide provides a good format for thinking about developing communication plans (http://cecr.ed.gov/development/engagement/). For examples of communication plans with strategies and tactics developed by states engaged in evaluation reform, see Colorado’s (http://www.cde.state.co.us/cdegen/downloads/SB%20191/Template_StayingInformed111010.doc) and Rhode Island’s (http://www.ride.ri.gov/commissioner/RaceToTheTop/docs/RI%20RTTT%20Communications%20Plan.pdf) publically available documents.

As the primary stakeholder in any principal evaluation reform process, work styles, schedules, and principals’ needs should be important when selecting communication methods as part of a broader communication plan. Several states have developed stand-alone websites for their evaluation systems designed explicitly for busy educators that reflect the format and features of popular blogging sites rather than conventional websites. Washington’s Teacher and Principal Evaluation Pilot (TPEP) project maintains an extensive website (www.tpep-wa.org) that provides information on each of the state’s nine pilots, evaluation system plans, video and print resources for the pilot districts, and all communication updates linked to both RSS and Twitter feeds. New York created “EngageNY” (www.engageny.org), which includes an entire section designed specifically for administrators with access to training modules, webinars, and other resources for both teacher and principal evaluation reforms.

5 The TPEP project has also partnered with Apple to provide access to all podcasts and videos related to the TPEP project through iTunes. To access the podcasts and videos, go to: http://itunes.apple.com/podcast/washingtons-teacher-principal/id426838641.

Strategy 4. Encourage deeper opportunities for educator participation that capitalize on stakeholder-driven processes to build local capacity.

States and districts have the option of capitalizing on stakeholder engagement as an opportunity to build the capacity of principals and superintendents. One example of a stakeholder-driven process designed to both garner stakeholder engagement and build capacity is the TPEP project. In 2010, the Washington Legislature mandated a two-year, collaborative pilot process to develop options for a statewide evaluation model(s). The state allocated between $100,000 and $180,000 annually for each of the nine pilot sites (eight districts and one consortium of rural districts) to design and implement their own teacher and principal evaluation systems. All pilot sites were required to include and engage stakeholders from both the school and community in the design process. Preliminary interview and survey data suggest that stakeholder engagement efforts at the district-level and training on the new principal evaluation models have had a positive impact on pilot participants’ perceptions, level of buy-in, and growth experiences from participating in the process.

Stakeholder-driven processes, however, present their own set of challenges and tradeoffs. Interview and focus group data from the TPEP project suggest that the focus on principal evaluation planning, design, and implementation has often taken a backseat to the teacher evaluation efforts in the pilot districts. Although interviewees see them as similarly important, pilot leadership often found balancing the needs of both projects daunting. Without adequate time and staff, a simultaneous design approach could result in inadequate time and resources being devoted to principal evaluation development. Second, stakeholder-driven processes require collaborative partnerships with the legislature and governor’s office to ensure consistent planning, communication, and messaging about state-reform processes. In Washington, the state legislature passed...
an evaluation reform bill four months prior to the completion of the pilot and included mandates that were contrary to some of the models created through the pilot process. Strong stakeholder-driven approaches to principal evaluation work best when state-level decisions are made through clear and transparent methods, and decisionmakers can directly demonstrate how they have incorporated or responded to stakeholder views and ideas.

Evaluator Training and Pilot Testing

Evaluator Training

The lynchpin of a strong principal evaluation system is a rigorous and ongoing training program for principal evaluators. Well-trained and properly supported evaluators help to ensure adherence to evaluation procedures, proper collection and interpretation of evaluation data, and usefulness of evaluation feedback and results in encouraging a principal’s professional growth and improved leadership practice. States and districts are responsible for providing a training and certification program that not only provides thorough initial training, but also monitors fidelity of implementation of the evaluation process, inter-rater reliability, and whether evaluator feedback has a positive impact on principal performance.

Although the exact content of the evaluator training program will depend on the goals of the evaluation system and its specific measures, a high-quality evaluator training program should include the following features:

- The evaluation system’s overarching purposes, process, and cycle;
- Relevant leadership standards and frameworks;
- Instruction on each selected measure and development of skills required for interpreting evidence collected through each measure;
- Using all evaluation system tools (e.g., observation tools, online data systems, summative feedback forms, scoring rubrics);
- Proper procedures and techniques for determining final summative ratings (if applicable);
- Techniques and skills for providing formative and summative feedback, as well as ongoing coaching using evaluation results; and
- Calibration of ratings to set standards to certify each rater before he/she evaluates performance.


In designing the evaluator training material, states and districts should ensure thoughtful tailoring of the training to the types of measures included in the principal evaluation system. Value-added measures of student growth, for example, will require information on both the technical features of the data system as well as development of skills to accurately interpret the data. Measures of principal practice, such as observation or portfolio review, require training that emphasizes techniques for promoting inter-rater reliability and training in self-reflection tools and portfolio assembly processes for principals themselves. Usage of teacher, parent, or student surveys will require local staff to receive additional training in survey administration and principal evaluators to receive training on proper interpretation of survey data, including assessing response rates and dealing with outliers in the data.

One current example of a principal evaluator training program is Iowa’s iEvaluate and Assessing Rigor courses. Although Iowa does not have a statewide educator evaluation system, school districts are required to have principal and teacher evaluation systems in place. The broader administrator certification requirements in Iowa include evaluator training.
The School Administrators of Iowa (SAI) and the Iowa Department of Education, drawing on grant funding from the Wallace Foundation, developed both courses in cooperation with each other.

The iEvaluate course for principal evaluators is a two-credit, online course that familiarizes superintendents with the state’s teacher and principal standards, the research-base and best practices for improved principal performance and superintendent support, critical components of effective evaluation, effective coaching (questioning skills and providing feedback), and developing professional development plans. Because Iowa does not have a state-mandated model with selected measures, the course focuses on developing the skills and knowledge any principal evaluator would need, regardless of the measures included in the school district’s particular model. The online course includes both periodic quizzes and on-site assignments, including a final assignment that requires evaluators to videotape themselves conducting a formative evaluation conference. The course participant submits the video and a structured self-reflection to the instructor who then assesses the participant’s skills in conferencing and provides detailed information on strengths and areas for improvement.

The Assessing Rigor course provides the content needed for the coaching and evaluation skills learned in iEvaluate. The four-day course is designed to prepare administrators and school leadership teams to “thoroughly and accurately measure, monitor, and encourage rigor” at both the classroom and school levels and includes information on the Iowa Common Core. Participants apply the Assessing Rigor content by engaging in and reflecting on a coaching conversation with an administrator in their school or district. Administrators also practice developing specific rigor-related goals as part of their individual professional development plans and learn how to develop a Rigor Dashboard for use in their schools.

According to SAI, the evaluator training program went through three iterations of development over the course of several years before reaching its present form. One important lesson learned was the importance of involving stakeholders from the very beginning as well as keeping them regularly involved. SAI reported that establishing the state advisory committee to bring more perspectives to the planning process significantly improved the third round of evaluator trainings in comparison with the first two rounds. A second lesson learned was recognizing early in the process that principals and superintendents struggled with writing and developing appropriate professional development plans. To address this challenge, SAI developed a module to train superintendents to assist principals in the individual professional development plan process. SAI indicates developing clearer mandates for assessing fidelity of implementation and inter-rater reliability on a periodic basis could further improve both principal evaluation and evaluator training in Iowa.

Pilot Testing

Pilot testing is an indispensable part of the principal evaluation design process. A rigorous pilot test is needed to better understand how evaluators implement the system and what impact the evaluations have on practice and important outcomes, as well as to answer questions about validity and reliability. At a minimum, pilot evaluations should consider the following questions:

- Do evaluators correctly conduct and complete evaluations within the required time?
- Do district leaders and principals understand what to evaluate, how evaluations occur, and proper application of results?
- Do district leaders and principals accept the evaluation process and results? Do they perceive the process and outcomes as being fair?
• Do evaluations occur consistently and with evidence of inter-rater reliability?
• How has the evaluation process affected practice?
• What is the validity of evaluation results? (Exploring distributions of scores on various components and whether other important organizational outcomes relate to results.)

Although it is possible to carry out pilot evaluations using internal evaluators, to maximize credibility and impartiality, there should be consideration of external evaluation experts. The higher the stakes to the evaluation decisions (i.e., termination, compensation), the more important a rigorous pilot evaluation is. Plans for on-going evaluation of system implementation are also necessary. The CECR program evaluation guide includes detailed steps for designing evaluation strategies [see: http://cecr.ed.gov/pdfs/CECR_ProgramEval_Guidebook.pdf]. Although the guide focuses on evaluation of compensation reforms, it can also help in the design of program evaluations for principal evaluation systems.

**Conclusion**

Developing or revising principal evaluation systems involves a number of interconnected parts. These systems start with standards for school leadership selected to match the leadership needed to enact school, district, or state educational priorities. Performance on leadership standards can be assessed by multiple measures of leadership practice. Outcomes from leadership practice can also be assessed using multiple measures. The results, whether aggregated into an overall effectiveness measure or reported separately, can be used to improve performance and as a basis for decisions related to position renewal or performance awards. Given these important uses, principals must be involved in the design, and the systems require thorough testing.

Principal evaluation system design and implementation is clearly a complicated endeavor. The process is further complicated given the lack of research or evaluation studies on new approaches to principal evaluation. However, states and districts are planning and deploying new measures of principal practice and school outcomes as the importance of developing better systems of principal evaluation to help principals improve and recognize effective performance has compelled action. Although the policy is moving ahead of the research, this paper, along with other resources referenced in this review, can help reformers begin the process of developing new evaluation approaches or refine their principal evaluation system.

**References**


