

Running head: MEASURING PRINCIPAL Effectiveness

A Framework for Assessing Principal Effectiveness: Review of research and recommendations

Margaret Terry Orr
Bank Street College of Education
New York, New York
morr@bnkst.edu
September, 2011

Draft, not for citation without the author's permission

School principals have long been accountable for school success and student achievement (Glasman, 1984; Glasman & Biniaminov, 1981; Miskel, 1977; Sweeney, 1982). Moreover, recent research has firmly established that some principal practices significantly affect student learning, although indirectly by influencing a school's vision and direction, organizational effectiveness, and teacher effectiveness (Hallinger & Heck, 1996; Leithwood & Jantzi, 2008; Marzano, Waters, & McNulty, 2005; Robinson, Lloyd, & Rowe, 2008; Witziers, Bosker, & Kruger, 2003). Such findings shift the accountability focus for principals to the nature of school and staff conditions they are most likely to directly affect. Yet, while experts agree on the importance of school leaders, their evaluation has been neglected. According to Rotherham: "principals play a critical role in student learning, but they are evaluated almost as an afterthought" (Rotherham, 2010).

The federal No Child Left Behind Act established a strict accountability system for schools to improve student achievement. A 2007 report from the Commission on No Child Left Behind (CNCLB) concluded that while this national policy laid important groundwork for improving public schools and closing achievement gaps among student-population subgroups, further action—including principal evaluation—was needed to create "a high-achieving education system that succeeds for every student, in every school" (p. 13).

In recent years, based on these research and policy findings and the continued national priority to raise student achievement, policy analysts have encouraged state and local officials to further leverage their influence over leadership quality as a means of improving student achievement. They recommend several policies including establishing leadership standards and creating new principal evaluation systems (Augustine et al., 2009). Consequently, assessing and evaluating principals has taken on greater importance in the present accountability environment of the both NCLB and the newer Race to the Top (RTTT) program and other federal grant programs. Stimulated in large part by federal policy, there has been a recent explosion of principal evaluation proposals, methods and approaches. Many states have or are in the process of adopting statewide principal evaluation models based on student outcomes.

Consequently, there is a strong need for high-quality principal evaluation measures that are not ambiguous, offer clear performance expectations and differentiation, and reflect the chain of practices that links principal actions to improved teaching and learning, and student outcomes. As states and districts take steps to

design and implement principal evaluation systems for high stakes personnel decisions, it is critical to consider: (1) the current state of the principal evaluation field, (2) available research on appropriate approaches, and (3) evaluation design considerations, including validity and reliability issues (Brown-Sims, 2010). This article fills this gap by reviewing existing research on effective principal practice and current trends and practices in principal evaluation and assessment and proposes a framework for assessment system development.

Background

Recent and emerging developments in principal evaluation are based on current research about effective principal practice and the direct and indirect effects they have on their schools and students, prior research on principal evaluation, and the use of leadership standards to guide principal evaluation. This research and its use in principal evaluation developments are summarized below.

Leadership research

While much school effectiveness research has concluded that principals are critical to school improvement (Leithwood & Riehl, 2005), it has been more difficult to ascertain which principal practices are most essential and thus be the target for principal evaluation. Recent research has tried to unpack the relationship among specific leadership practices, teacher effectiveness and organizational conditions and school improvement and achievement outcomes (Leithwood & Jantzi, 2008; Robinson et al., 2008). Together, these findings suggests that effective principal

practices center on (a) vision and fostering coherence and persistence, (b) engaging in improvement of curriculum and instruction, (c) developing individual and collective capacity, (d) distributing leadership and shared responsibility, (e) using data to monitor progress, and (f) engaging family and community.

Some research has focused on the leadership and organizational conditions that are necessary to improving schools with challenging conditions, such as those in socio-economically disadvantaged areas. Muijs, Harris, Chapman, Stoll, and Ross (2004) identified several common themes in their review of research, particularly, having a focus on teaching and learning, effective leadership, an information-rich environment, positive school culture, and a learning community. Other researchers used survey analyses to distinguish breakthrough schools—low-performing schools that perform like high-performing ones (Glidden, 1999; Sebring, Allensworth, Bryk, Easton, & Luppescu, 2006; Watts et al., 2006; Williams, Kirst, Haertel, & et al, 2005). Their results confirmed the themes identified by Muijs and colleagues and expanded them by adding several leadership practices: implementing a coherent standards-based curriculum and instructional program, using assessment data to improve instruction, sticking with a reform over time, tailoring strategies to individual student needs, and ensuring the instructional resources as additional features of effective principals.

State of the principal evaluation field

The current principal evaluation field is in its infancy, or, as Kuhn (Kuhn, 1970) might term, in a pre-paradigmatic stage conceptually and methodologically. Until recently, there have been little

agreement on what and how to evaluate principals. Few valid and reliable methods for principal evaluation exist and there is little agreement on what measures and information should be used as evidence of principal effectiveness (Porter, Goldring, Murphy, Elliot, & Cravens, 2006). Goldring and others (Goldring, Porter, Murphy, Elliot, & Cravens, 2007), in their content analysis of 66 districts' evaluations, found that evaluation instruments varied widely in breadth, scope and focus on instructional leadership. Most lacked depth and attention to quality organizational or classroom outcomes, behaviorally-oriented definitions for categories, or psychometric information.

Similarly, Knapp and colleagues (2006) analyzed existing practices in educational leadership assessments in another district sample. They identified three major uses of principal evaluations: (a) evaluating leaders' performance, (b) providing formative feedback for leadership development, and (c) investigating how to improve schools. They found that as districts' assumptions about leadership and school improvement have evolved, so also has the role and nature of their leadership assessments. Among the changes identified are movement (a) from traits and dispositions to behaviors and actions, reflecting an emphasis on outcomes; (b) toward a professionalized basis for leaders' work building on adoption and use of national and state leadership standards; (c) in increased focus on the centrality of leadership to improve student learning; (d) for greater use of assessments for leadership development purposes; and (e) for increased understanding of the influences of organizational and district contexts affecting leadership (Knapp, Copland, Plecki, & Portin, 2006).

Finally, Cantano and Stronge (2006) analyzed the content of one

state's principal evaluation instruments. They found that districts most often focused on principals' instructional leadership, organizational management, and community relations.

Leadership standards and principal evaluation

The greatest progress in principal evaluation appears to be the practice of adopting local and statewide standards for assessment purposes. Drawing from this and other research as well as expert opinion and best practices, local, state and national groups have developed leadership standards to be used to guide policy and practice to improve principal effectiveness. The most widely used are the Interstate School Leadership Licensure Consortium (ISLLC) *Standards for School Leaders* (Council of Chief State School Leaders [CCSSO], 1996) which were recently reviewed and revised (Council of Chief State School Officers, 2008).

State and local adoption of research-based leadership standards has been accelerating in recent years, particularly as a means of focusing expectations for principals' work and their evaluation (Toye, Blank, Sanders, & Williams, 2007). There is evidence that state adoption of leadership standards has led to changes in district evaluation of school leaders. For example, Cantano and Stronge (2006) found, in their study of local principal evaluation systems in one state, that districts were using common expectations that were congruent with state and professional standards (N. Catano & J. H. Stronge, 2006). They found however, as earlier research had shown (Glasman & Martens, 1993), that districts varied considerably in their use of leadership standards to frame principal evaluation systems. Similarly, in their review of district leadership assessment instruments, Goldring and others found that about half the districts

used local, state or national leadership standards, while others lacked reference to the basis for their leadership expectations (Goldring et al., 2007)

Framing principal evaluation

While leadership standards provide a means for framing expectations of principal practice, principal evaluation systems, as currently being developed, are shifting the focus away from assessing the nature and quality of principal practice toward assessing expected outcomes (Orr, 2010). Given the newness of the principal evaluation field and debates over criteria for evaluating principal effectiveness, it is critical to establish a framework that encompasses both expectations for principal practice and appropriate outcomes, along with relevant measures and methods.

Several experts propose a framework of “ingredients” for leadership assessment. These ingredients can be grouped into seven categories, as shown in Table 1: the purpose of the evaluation, what is assessed and measures used, sources of evidence used, who is assessed, who provides feedback, when assessment occurs and how assessment is conducted, and the psychometric qualities of the assessments (Brown-Sims, 2010; Condon & Clifford, 2010; Portin, Feldman, & Knapp, 2006). To this list is added the relationship between the evaluation and follow up support. These elements are clustered into two areas—content and organization and implementation of evaluation—and discussed below.

Table 1:
Key elements of a principal evaluation framework

<i>Content-related qualities</i>
Purposes of assessment
Who is assessed
What is assessed and measured
Sources of evidence
<i>Organizational and implementation qualities</i>
How assessments are conducted
How evidence is valued
Psychometric qualities (including reliability and validity)
How the assessment system is implemented and operates

Content-related qualities of principal evaluation

The content of a principal assessment system depends upon the intended purposes, the target of the assessment, the behaviors, practices and outcomes being assessed, and the sources of evidence being used. Current practice and expert recommendations are summarized below.

Purposes of assessment

Commonly, there are three primary purposes for principal assessment: summative, formative and organizational change. Each has different implications for the content of the assessment. The first type of use—and most typical-- is summative, for personnel management in making consequential decisions pertaining to continued employment, tenure, promotion and supplemental

compensation (New Leaders for New Schools, 2010; Portin et al., 2006)(Milanowski & Schuermann, 2009). The second use is for principal development, as a means of fostering continuous improvement, addressing areas of weakness or gaps in practice or guiding professional learning (Portin et al., 2006). A third use is as a lever for broader organizational change by providing clarity and fostering coherence in expectations for principals’ work (Portin et al., 2006).

The design and use of a principal evaluation system will vary based on whether it serves one or more of these purposes. Until recently, only about half of all assessments provide principals with clear feedback linked to a development plan to improve practice (Goldring et al., 2007) and little information exists on how much these are used for organizational direction setting and change.

Whatever purposes are used, should, as Brown-Sims (2010) proposed, be clearly articulated as goals and expectations for the assessment system.

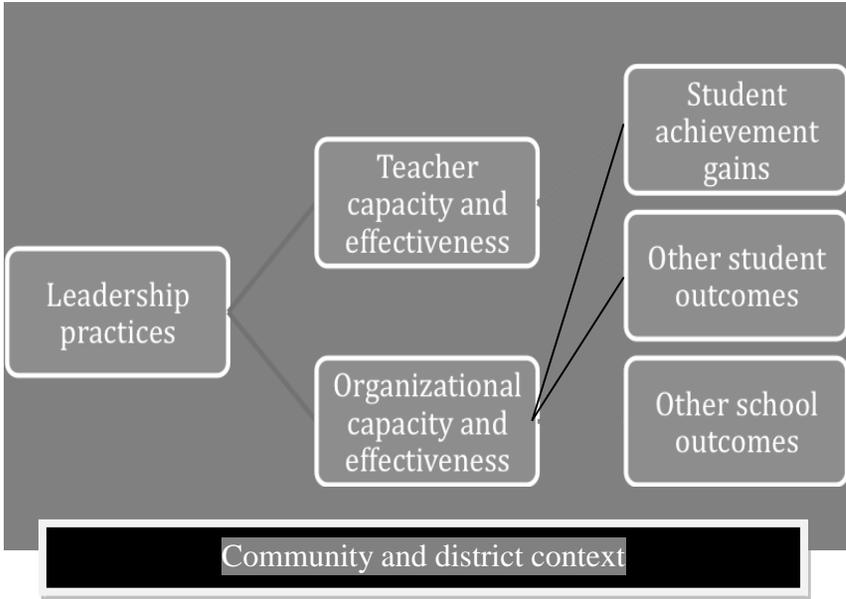
Who is assessed

The assessment system should clarify who is the target of the evaluation. By definition, principal evaluation systems are designed primarily for principals. Some states extend their evaluation systems to other school and district leaders or develop general evaluation systems that are applicable to all educators (Orr, 2010). Defining who is assessed includes whether to differentiate criteria based on principal experience and employment status as new, tenured or experienced, and by level and scope of responsibility.

What is assessed and measured

There is considerable debate over what should be measured in a principal evaluation. A review of research and commentary shows that at the center of these debates are three kinds of factors which mirrors the relationship between leadership and school outcomes: principal practices, teaching and organizational effects, and student outcomes, as shown in Figure 1.

Figure 1
Leadership pathway of influence on school outcomes



(Orr, 2011).

This set of relationships mirrors Reeves (2004) proposed principal evaluation framework that sorts assessment content into influences

over which a principal has 1) direct control, 2) direct influence, 3) indirect influence, or 4) neither control or influence. He acknowledges that the nature of principal control and influence in each domain varies by local policies and conditions, reflecting the need to incorporate context consideration into principal evaluation.

Leadership practices. There is increasing agreement that principal practice efficacy be part of principal evaluation (Milanowski & Schuermann, 2009), although some debate still remains about which practices to use. Knapp and colleagues (2006) found there to be an increased focus on principal behaviors and actions while Catano and Stronge (2006) found a common focus on instructional leadership, organizational management and community relations.

As states and local districts adopt national leadership standards (Toye et al., 2007), they increasingly share common expectations on areas of leadership practice that are critical to evaluation. Other educational experts and professional associations have advocated that principal evaluation be based on other practices, using research evidence about how principals influence student achievement and some groups have developed their own leadership standards and expectations (N. Catano & J. H. Stronge, 2006). Reeves, for example, argued that principal evaluation be based on proficiency in domains of leadership related to improved student achievement; he proposed 10 potential leadership domains, such as communication, decision-making, resilience, faculty development and learning (Reeves, 2004).

Finally, a review of six states' RTTT proposals shows that all planned to incorporate assessments of principal practices as part of their

evaluation (Orr, 2010). Most define these practices based on state or national standards.

Teacher effectiveness and organizational conditions. Increasingly, principal evaluation is focusing on measuring their effectiveness based on the outcomes they most directly influence: teachers' instructional effectiveness and organizational conditions for improved student learning. Milanowski (2009) proposes that principal evaluation systems incorporate measures of intermediate outcomes.

Very little work exists on the identification of these areas of evidence and how to measure these, beyond expert recommendations and leadership effectiveness research, noted above. Some field related examples exist. Brown-Sims, in proposing measures of principal effectiveness, suggested assessing teacher working conditions (Brown-Sims, 2010). New York City includes in principal evaluations teacher, student and parent assessments of the school's learning environment as well as a Quality Review of a school's organization and use of data. Finally, Rhode Island's *Building Administrator Professional Practice Framework* (2010) is designed to use evidence from a variety of sources, including "professional practices that are linked to student outcomes" (p. 2). Such evidence, while used to rate standards-based leadership practices, is drawn from observations of the school, classroom and teachers, such as the mission statement is evident and understood by the school community and classroom visits show that lessons are conducted based on lesson objectives aligned to meet student learning goals (Rhode Island Department of Education, November 9, 2010). Taken together, these illustrate the variety of intermediate

outcomes related to teacher and organizational effects that could be considered as intermediate or direct outcomes of principal practice, for purposes of principal evaluation.

Student outcomes. Educational experts and policy makers are now linking evaluation of principal performance to improved student outcomes. Since 2002, through NCLB, schools have been held accountable for improved student achieving using state assessment data. Now, federal policy and educational experts are pushing for using the same (and other assessment data) to evaluate teacher and principal performance. Such use for teacher evaluation is hotly debated, primarily on validity and measurement grounds (Glazerman et al., 2010; Harris, 2010). Little attention has been given to the applicability their use in principal evaluation and whether the same issues apply, although some early criticism exists.

Reeves (2004) argued against using student test scores as measures of principal performance because these do not illuminate the leader decisions and actions that influence student learning. As he uses a black box analogy to illustrate:

“The presumption of any model, whether using mean test scores or value-added models, is that the relationship between leader, teacher, and student is an imponderable black box and thus only the output can be examined. Any other discipline in the natural or social sciences that took such a facile approach would be subject to well-deserved ridicule.” (p. 29 (Reeves, 2004)).

For Reeves, using test scores to evaluate principals’ performance

raises both methodological and moral problems, although suggesting that test scores could be used as one strategic piece of data in evaluating performance.

Context. Among the types of evidence to be included in an evaluation system is the nature of the context of principals’ work. Portin and others (2006) stress the importance of including evidence about context—national, state, local and school-level-- in developing and using principal assessment systems. According to them, different contexts interact to influence local leadership standards, conceptions of leaders’ roles, and the supports provided school leaders that enable them to be effective. As Portin and others (2006) explain further, state and national policies, particularly for school accountability, interact with local priorities, beliefs and expectations about schools and leaders’ roles.

In addition, assessment policies emerge from district personnel practices for hiring, assignment, compensation, supervision and related personnel issues, as well as union contract negotiations. Local efforts to improve schools and staff effectiveness create additional expectations and some district may use assessments for professional growth and development purposes. Portin and others (2006) point to other local conditions which have bearing on assessment practices, including the district’s reform need and climate of the district; size, wealth and diversity; data capacity, and organizing culture. Other local context considerations includes, according to Reeves (2004), principals’ span of control and authority, challenging nature of the contexts, and resources that influence their accountability.

Finally, Portin and others (2006) stress that each school context should be taken into account because its attributes—including parent, staff and student expectations, school resources, and reform climate and trust—cannot be easily disentangled from principal practices as influences on organizational and student outcomes.

Sources of evidence and their measurements

Coupled with the discussion of the types of evidence are the sources of evidence and their measurements, which, by their nature, add to or detract from the evaluation validity of different types of evidence. Recent analysis of six states' RTTT proposals (Orr, 2010) and recommendations for principal evaluation (Brown-Sims, 2010; Milanowski & Schuermann, 2009), suggest that common sources of evidence and their measurement include judgment ratings of principal performance, ratings of first-hand observations and document reviews, and ratings of principal prepared portfolios of documented performance evidence. In practice, some states are also recommending multiple sources of evidence and measures. For example, the Rhode Island principal assessment model recommends using three types of school practices data upon which to make judgments about principal proficiency: school visits to observe practices and operations; document review; and feedback from stakeholders through surveys, interviews and discussions (Rhode Island Department of Education, January 2011).

Judgment ratings. A source of evaluation evidence is solicited feedback from one or more stakeholders about a principal's performance as a school leader. Such assessments vary on the practices that are assessed, how performance is evaluated or rated,

and who provides the rating feedback. Assessment instruments such as Vanderbilt Assessment of Leadership in Education (VAL-Ed) (<http://valed.com/>), Reeves assessment system (2004) and McRel's principal evaluation system (<http://www.mcrel.org/product/392>) are feedback or rating systems designed around their review of research on effective principal practices and related assessment development work. All include a list of indicators of principal performance to be rated.

Feedback is typically measured through Likert-like scale ratings, and various instruments differ in the degree of specificity for each principal practice being assessed. Most typically, judgments are made by the principal's immediate supervisor and the principal him or herself (as a self-assessment). Other ratings can be provided by peers, teachers and other professional staff, students and parents.

Judgments can also be solicited through interviews and discussions, but not typically because of the time required and administration difficulty to collect the information.

Observations and documentation of school practices. Some districts and states are collecting observable evidence of principal practice, based on school visits and document review and assessing principal effectiveness using a standardized rating rubric. For example, the Rhode Island model for principal performance assessment lists wide variety of sources of evidence to observe in school visits (in classrooms, halls, and meetings) or review from documents (such as agendas, strategic plans and reports) to rating specific principal skills and competencies (Rhode Island Department of Education, January 2011). The model lists possible sources of evidence for each

of 13 leadership domains, which are somewhat aligned to the ISLLC standards. The intent is for district officials to rate their principals (and other school leaders) based on the relevant and aligned sources of evidence.

Portfolios. Some principal evaluation systems require principals to compile documentation on their accomplishments, as a portfolio, for independent assessment. Typically, these portfolios are standards-based and require principals to collect data, compile plans and reports, and provide reflections. For example, this method is the basis of Alabama’s principal evaluation system (Alabama professional education personnel evaluation program, 2002) and the recently piloted National Board Certification for Principals (<http://www.nbpts.org/>). Key to these evaluation systems is the rubrics used to evaluate the documentation that principals compile and the training and reliability of evaluation raters.

Organizational and implementation qualities of principal evaluation

Key to the validity and reliability of a principal assessment system are the organizational and implementation features. These include the how the assessment is conducted, how evidence is valued, psychometric considerations, and assessment system implementation and use. Key considerations are summarized below.

Conduct of the assessment

How the assessment system is conducted has bearing on the quality and reliability of the data collected and its validity in being used to make evaluative judgments about principals’ performance. These considerations include the frequency and timing of evaluation, and

the use of multiple measures.

Frequency and timing. Until recently, according to Portin and others (2006), principal evaluation typically occurs infrequently, often no more than once a year, and usually only early in a principal’s career. Many of the RTTT proposed principal evaluation systems are requiring that all principals be evaluated annually, regardless of experience or tenure (Orr, 2010).

Use of multiple measures. Experts recommend that assessment of principals’ practice and performance be based on multiple measures ((Brown-Sims, 2010; Milanowski & Schuermann, 2009). Encompassed in multiple measures are different sources of data (survey, interview, observations, document analysis and school indicators), different stakeholders (such as supervisors and staff providing feedback), and different outcomes of principal performance (both practice and direct and indirect outcomes). Commonly, RTTT states require that multiple measures be used, combining both types of measures, types of data and different stakeholders’ input. Most typically in the states’ RTTT were to combine multiple ways of documenting and rating (using different stakeholders) a variety of principal practices, with student outcome measures.

How evidence is valued

Given the multiple forms of data used and different types of ratings, a principal evaluation system entails several kinds of value determinations. Based on a review of six RTTT proposals and principal evaluation studies cited above, four types of value determinations were identified:

- establishing a standard of expectation for performance, proficiency or effectiveness;
- assigning value that relates to this standard for each source or type of data,
- weighting similar or different types of evidence to create a final score for this standard
- determining a final score in relation to this standard of expectation.

Many commonly cited principal evaluation systems are designed to collect value judgments about principals' practices and accomplishments from one or more types of assessors. This is done by incorporating rubric-based rating scales into assessment instruments in order to differentiate principal actions, practices or outcomes for one or more behaviors along a continuum of proficiency or effectiveness, typically on a 3- to 5-point scale (Goldring et al., 2007). The VAL-Ed instrument, for example, focuses on six components of school performance (e.g. high standards, rigorous curriculum and quality instruction) and six leadership processes (e.g. planning, implementing, and monitoring). Similarly, McREL's principal evaluation system, based on its Balanced Leadership Framework, combines "three sets of formative rubrics—purposeful community, managing change and focus of leadership-- that emphasize 21 research-based leadership responsibilities associated with improving student achievement." (p. 1)(McREL, 2010).

Increasingly, principal evaluation systems use weights for different sources of data (ranging from no weighting, equal weighting, or differential weighting). For example, New Leaders for New Schools

argues that given that principals improve student achievement by increasing the effectiveness of teachers and taking leadership actions, their evaluation should consist of and be heavily weighted toward student achievement and teacher effectiveness outcomes, and recommend that only 30% of an evaluation assess "demonstration of effective practices and Leadership actions." (New Leaders for New Schools, 2010, p. 1). They define student outcomes as measures of student growth and proficiency attainment, while measuring teacher effectiveness based on teacher-based gains in student achievement outcomes, the retention of effective teachers, and successful "exiting poor performers." They outline six domains of leadership action (such as vision for results and equity and learning and teaching), but give these little weight in their proposed evaluation.

In contrast, Milanowski (2009) proposes a principal score card that includes four types of evidence, each of which is rated on a five point scale and then ascribed different weights when being totaled into a composite score. These types of evidence and their weights are: development (goal setting) (20%), principal practices (20%), intermediate outcomes (30%) and student outcomes (30%). The weights and ratings are combined mathematically to create a single composite score.

Principal evaluation systems designed to inform personnel decisions will typically yield a single final score. Historically, this had yielded a simple dichotomy of effective/not effective. Now, the shift is to a categorical rating (as stipulated by RTTT guidelines) (with four-point score ranges from ineffective to highly effective or from needs improvement to exemplary) or other kind of metric or description.

Whatever the categories, how they differentiate gradations of performance should be clearly defined. Reeves (2004), for example, recommends using four categories to differentiate principal performance—exemplary, proficient, progressing and not meeting standards—and that each domain have clearly described performance expectation in each category to distinguish performance levels. Each performance category would have a numeric range for interpreting composite scores.

Psychometric considerations

Psychometrically, principal evaluation measures and methods are still in their early stages of development and lack psychometric evidence. Most commonly, principal assessment instruments are developed and evaluated only for content validity, such as based on prior research or available leadership standards. There is no evidence of their concurrent validity (with other evaluation measures) or predictive validity on organizational outcomes, such as school measures or student achievement results except as extrapolated from the research upon which they are based (Claudet, 2002; Oyinlade, 2006).

Establishing high levels of validity and reliability are critical to the integrity of a principal evaluation system, particularly for their use in making consequential decisions. To clarify further Condon and Clifford (2010) define validity and reliability in their assessment of effective principal evaluation instruments: “assessments are considered *valid* when they measure what they are intended to measure.” (p. 3). Yet, they only stress the important of two types: content and construct validity, and define construct validity reflects the “degree to which test items measures a “construct,” which is

the element that the items purpose to assess.” ((Condon & Clifford, 2010), p. 3). They explain that reliability is a measure of consistency and stability in what is being measured and how it is being measured. The goal is for an evaluation to yield relatively similar results if there are multiple administrations of the assessment or multiple assessors or raters.

Condon and Clifford (2010) then used these psychometric attributes to compare eight principal assessment tools, selected for their detail on specific practices and qualities to be used for formative assessment and available evidence of content and construct validity. Among the eight, only six have been published and only two within the last ten years: Leadership practices Inventory (Kouzes & Posner, 2002) and the Vanderbilt Assessment of Leadership in Education (VAL-ED) (Porter et al., 2006). All eight instruments have good content validity, established through research reviews and expert feedback. Most had construct validation through confirmatory factor analysis or inter-item correlation. Only two had concurrent validation, based on comparisons to other evaluation measures or by correlating different raters’ feedback. The instruments range in their reliability from poor (less than 0.80 alpha coefficient) to high (near or equal to 1.00). The VAL-ED instrument, designed to assess “learning-centered leadership,” was most highly rated for its content validity and reliability (Goldring et al., 2007; Murphy, Elliott, Goldring, & Porter, 2006; Porter et al., 2006).

These results demonstrate that much work is needed to develop and evaluate high quality measures and methods for principal evaluation.

Assessment system implementation and use

Given the limited standardized, field-tested principal evaluation models and methods, there is little information on best practices around their implementation and use in improving principals' performance. Available guidance comes from psychometric protocols, RTTT plans, and expert opinion.

Field testing measures, methods and systems. A standard practice in the development of assessments is piloting the measures and field testing their use in practice (Louden & Wildy, 1999; Oyinlade, 2006; Porter et al., 2010). How assessment tools are applied and the system of administration contributes to the methods' validity and reliability. Before being implemented for field use, they must be evaluated to ascertain whether the methods and systems for evaluation, as well as the measures themselves, meet appropriate psychometric standards.

Most RTTT states adopted this recommended practice, as evidenced by their proposals, with plans to either pilot or pilot and field-test (with volunteer or designated—usually low-performing--districts) their principal evaluation measures, methods and evaluation systems before requiring their use statewide (Orr, 2010).

Implementation and operations of the assessment system. While much attention has focused on selecting assessment methods to fit assessment purposes and measure principal performance, experts also draw attention to operations issues. As Brown-Sims (2010) pointed out, "when selecting an assessment system, district must consider the time needed to administer the instrument, the cost and ease of use or implementation."(p. 6)

Evaluating the assessment system. As Brown-Sims (2010) underscored, a principal assessment system should be evaluated for its fidelity in use. A few RTTT states have proposed to evaluate their assessment system, once implemented, for continued quality control (Orr, 2010). Since receiving RTTT funding, Delaware has annually solicited feedback from principals and their evaluators on their educator evaluation methods, instruments, and assessment processes (including the time involved). As their evaluator described the purpose of the annual evaluation:

"The majority of the findings center on the practices and processes of DPAS II. The practices provide an understanding of the quality of training, manuals, forms, and general deployment. The processes stem from fundamental policies and underlying theory about performance appraisal." (p. 1 (Beers, 2010)).

In addition, some RTTT states proposed assessment outcome goals to be used in evaluating their assessment system's effectiveness. Such goals are stated in terms of the percentage of effective and highly effective principals, and the reduction over time in the percentage of ineffective principals (Orr, 2010).

Providing feedback and support. While some states' RTTT proposal outline steps for improvement plan and follow up support for principals rated as ineffective, only a few build in on-going feedback and support for all principals, underscoring their formative evaluation purposes. Vitcov and Bloom (2010), based on their work with school districts, outlined several recommendations for effective principal supervision and evaluation. Among these are that districts must "ensure the process is consistent with knowledge of

adult learning and professional development best practices, including collaboration and a sense of shared ownership” (p. 2010). Thus, they do not separate on-going support and development from assessment, stressing instead that the assessment focus should be on improving principal performance, not documenting whether principals met set performance expectations,

Conclusion

Taken together, principal assessment and related evaluation systems are still in their infancy. Recent research, expert opinion and state plans provide a significant foundation for their future development and use. Table 2 provides a summary of the key evaluation elements and suggested considerations based on the above review of research and expert opinion.

In the rush to develop and implement state wide principal evaluation systems, care must be taken to map the measures and content of the evaluation systems to reasonable and appropriate expectations of principals’ work and its influence, as noted above, and to take psychometric issues into consideration when designing and using rating systems and instruments. Given the infancy of the field, it is critical that all new evaluation systems themselves be evaluated for concurrent and predictive validity and reliability and for unintended consequences in their use.

Table 2:
Summary of key evaluation elements and considerations

<i>Elements</i>	<i>Considerations</i>
The purposes of assessment	<ul style="list-style-type: none"> • Personnel management to make consequential decisions • Leadership development for growth and improved practice • Organizational change
Who is assessed	<ul style="list-style-type: none"> • Principals only, or includes other school and district leaders • Differentiation based on years of experience, level and responsibilities • Differentiated based on context
What is assessed	<ul style="list-style-type: none"> • Leadership practices • Teacher effectiveness and organizational conditions • Student outcomes • Context
What sources of evidence are used	<ul style="list-style-type: none"> • Judgments (through surveys, interviews or focus groups) • Observations of principal • classroom visits and site visits • Documents and other evidence • Portfolios and artifacts
How the assessment is conducted	<ul style="list-style-type: none"> • Frequency and timing • Use of multiple measures

<p>How evidence is valued</p>	<ul style="list-style-type: none"> • Using leadership standards against which to make judgments • Rating of individual sources of evidence • Weighting each source of evidence when combining them into a total score • Generating a total score that discriminates principals as proficient or effective
<p>What psychometric qualities are maintained</p>	<ul style="list-style-type: none"> • Content and construct validity • Concurrent validity • Predictive validity • Reliability
<p>How the assessment system is implemented and operates</p>	<ul style="list-style-type: none"> • Field testing the assessment system before implementation • Implementation and operations of the assessment system • Evaluation of the assessment system’s qualities, implementation and use • Feedback and support mechanisms

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Alabama professional education personnel evaluation program. (2002). *Principal evaluation system: Manuals, forms and other materials*.

Augustine, C. H., Gonzalez, G., Ikemoto, G. S., Russell, J., Zellman, G. L., Constant, L., et al. (2009). *Improving school leadership: The promise of cohesive leadership systems*. Santa Monica, CA: Rand.

Beers, D. E. (2010). *Delaware Performance Appraisal System Second Edition (DPAS II) Year 3 Report*. Chicago, Ill: Progress Education Corporation.

Brown-Sims, M. (2010). *Evaluating School Principals. Tips & Tools*. Washington, DC: National Comprehensive Center for Teacher Quality.

Catano, N., & Stronge, J. H. (2006). What are principals expected to do? congruence between principal evaluation and performance standards. *NASSP Bulletin*, 90(3), 221-237.

Catano, N., & Stronge, J. H. (2006). What are principals expected to do? Congruence between principal evaluation and performance standards. *NASSP Bulletin*, 90(3), 221-237.

Claudet, J. (2002). Integrating School Leadership Knowledge and Practice Using Multimedia Technology: Linking National Standards, Assessment, and Professional Development. *Journal of Personnel Evaluation in Education*, 16(1), 29.

Condon, C., & Clifford, M. (2010). *Measuring principal performance: How rigorous are commonly used principal performance assessment instruments?* Naperville, Il: Learning Point Associates.

Council of Chief State School Officers. (2008). *Educational leadership policy standards: ISLLC 2008, as adopted by the National Policy Board for Educational Administration*. Washington, DC: author.

Glasman, N. S., & Martens, P. A. (1993). Personnel evaluation standards: The use in principal assessment systems. *Peabody Journal of Education*, 68(2), 47-63.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating teachers: the important role of value added*. Washington, DC: Brookings Institution.

- Glidden, H. G. (1999). Breakthrough schools: Characteristics of low-income schools that perform as though they are high-income schools. *ERS Spectrum*, 17(2), 21-26.
- Goldring, E., Porter, A. C., Murphy, J., Elliot, S. N., & Cravens, X. (2007). *Assessing learner-centered leadership: Connections to research, professional standards and current practices*. Nashville, TN: Vanderbilt University.
- Hallinger, P., & Heck, R. (1996). Reassessing the principal's role in school effectiveness: A review of empirical research, 1980-1995. *Educational Administration Quarterly*, 32(1), 5-44.
- Harris, D. N. (2010). Value-added measures of education performance: Clearing away the smoke and mirrors. *PACE Brief*, 10(4).
- Knapp, M. S., Copland, M. A., Plecki, M. L., & Portin, B. S. (2006). *Leading, Learning and Leadership Support*. Seattle, WA: Center for the Study of Teaching and Policy, University of Washington.
- Kouzes, J., & Posner, B. (2002). The leadership practices inventory: Theory and evidence behind the five practices of exemplary leaders.
- Kuhn, T. (1970). *The Structure of Scientific Revolutions*. Chicago, Ill: University of Chicago Press.
- Leithwood, K., & Jantzi, D. (2008). Linking leadership to student learning: The contributions of leader efficacy. *Educational administration quarterly*, 44(4), 496-528.
- Leithwood, K., & Riehl, C. (2005). What we know about successful school leadership. In W. Firestone & C. Riehl (Eds.), *A New Agenda: Directions for Research on Educational Leadership* (pp. 22-47). New York: Teachers College Press.
- Louden, W., & Wildy, H. (1999). Short shrift to long lists An alternative approach to the development of performance standards for school principals. *Journal of Educational Administration*, 37(2), 99.
- Marzano, R. J., Waters, T., & McNulty, B. A. (2005). *School leadership that works: From research to results*. Alexandria, VA: Association for Supervision and Curriculum Development.
- McREL. (2010). McREL's Principal Evaluation System.
- Milanowski, A. T., & Schuermann, P. (2009). Principal evaluation (powerpoint slides), *Teacher Incentive Fund Grantee Meeting*. Bethesda, MD: Center for Educator Compensation Reform.
- Murphy, J., Elliott, S. N., Goldring, E., & Porter, A. C. (2006). *Learning-centered leadership: A conceptual foundation*. Nashville, TN: Vanderbilt University.
- New Leaders for New Schools. (2010). Evaluating principals: Balancing accountability with professional growth. Executive Summary: author.
- Oyinlade, A. O. (2006). A Method of Assessing Leadership Effectiveness: Introducing the Essential Behavioral Leadership Qualities Approach. *Performance Improvement Quarterly*, 19(1), 25.
- Porter, A. C., Goldring, E., Murphy, J., Elliot, S. N., & Cravens, X. (2006). *A framework for the assessment of learning-centered leadership*. Nashville, TN: Vanderbilt University.
- Porter, A. C., Polikoff, M., Goldring, E., Murphy, J., Elliott, S. N., & May, H. (2010). Developing a psychometrically sound assessment of school leadership: The VAL-ED as a case study. *Educational administration quarterly*, 46(2), 135-173.
- Portin, B., Feldman, S., & Knapp, M. S. (2006). *Purposes, Uses, and Practices of Leadership Assessment in Education* Seattle, WA: Center for the Study of Teaching and Policy, University of Washington.
- Reeves, D. B. (2004). *Assessing educational leaders*. Thousand Oaks, CA.: Corwin Press.

- Rhode Island Department of Education. (January 2011). Rhode Island Model Educator Evaluation System. Working draft. Providence, RI: Rhode Island Department of Education.
- Rhode Island Department of Education. (November 9, 2010). Working draft. Rhode Island Model. building administrator professional practice framework. Providence, RI: Rhode Island Department of Education.
- Robinson, V. M. J., Lloyd, C. A., & Rowe, K. J. (2008). The impact of leadership on student outcomes: An analysis of the differential effects of leadership types. *Educational administration quarterly*, 44(5), 635-674.
- Rotherham, A. J. (2010). Paging Principal Skinner: Evaluating school leaders. *Time U.S.*
<http://www.time.com/time/nation/article/0,8599,2026632,00.html>. downloaded September 19, 2011.
- Sebring, P. B., Allensworth, E., Bryk, A. S., Easton, J. Q., & Luppescu, S. (2006). *The essential supports for school improvement*. Chicago, IL: University of Chicago, Consortium on Chicago School Research at the University of Chicago.
- Toye, C., Blank, R., Sanders, N. M., & Williams, A. (2007). *Key state education policies on P-12 Education: 2006. Results of a 50 state survey*. Washington, D.C.: Council of Chief State School Officers.
- Watts, M. J., Campell, H. E., Gau, H., Jacobs, E., Rex, T., & Hess, R. K. (2006). *Why some schools with Latino children beat the odds and others don't*. Tempe, AZ: Arizona State University.
- Williams, T., Kirst, M., Haertel, E., & et al. (2005). *Similar students, different results: Why do some schools do better? A longitudinal survey of California elementary schools serving low-income students*. Mountain View, CA: EdSource.
- Witziers, B., Bosker, R. J., & Kruger, M. L. (2003). Educational leadership and student achievement: The elusive search for an association. *Educational Administration Quarterly*, 39(3), 398-425.