MEASURING PRINCIPAL EFFECTIVENESS

Daniel Player

March 2018

EdPolicyWorks
University of Virginia

This paper has been prepared for the Virginia Department of Education under contract 7.7-PR8703018- 2019.

MEASURING PRINCIPAL EFFECTIVENESS

Principals play a critical role in school organization and culture, and as with most jobs, it is uncontroversial to assert that some principals are likely more effective in their roles than others. A substantial and growing body of research has confirmed this hypothesis by showing that principals have meaningful influence on student achievement. This suggests that a well-executed principal evaluation system could have meaningful effects on student achievement. Much of the recent evolution in principal evaluation strategy has focused on more precisely drawing the connection between principal behavior and student achievement.

While we know that principals matter in terms of student success, what has been difficult to measure is what precisely a principal does that makes one principal more effective than another. Likewise, research has demonstrated that the types of student achievement measures included in many current state evaluation systems are biased; they tend to reward principals in advantaged schools and penalize principals in disadvantaged schools. This limits the ability to use principal evaluation for high-stakes decisions such as dismissal or promotion. Nonetheless, principal evaluation can still be used as a development tool and help principals understand how to focus their actions on the behaviors that have the greatest likelihood of translating to greater student achievement.

In what follows, we provide a brief background around principal effects and the history of principal evaluation. We provide a rationale for a principal evaluation system, summarize what is known about current evaluation strategies and provide some recommendations for principal evaluation to improve practice. We conclude with detailed descriptions of three statewide principal evaluation systems that show some of the greatest promise in promoting effective principal practice.

Background

The role of teacher effectiveness on student achievement is intuitively appealing and is well established in empirical literature (for example, Chetty, Friedman, & Rockoff, 2014; Konstantopoulos, 2014). Because principals are more distal to student outcomes, the direct connection between principal effectiveness and student achievement has proved harder to identify conceptually and quantify empirically. Nonetheless, a growing body of evidence suggests that principals do have a causal effect on student achievement. In what follows, we briefly summarize the literature that has examined this relationship.

Principals and student achievement

The literature on principal effects is fairly well established. One of the most influential papers on the topic asserts that principals explain up to 25 percent of total student achievement (Leithwood, Seashore Louis, Anderson, & Wahlstrom, 2004). Several recent studies examining changes in student achievement that correspond with principals lend even more empirical support to this connection (Branch, Hanushek, & Rivkin, 2012; Coelli & Green, 2012; Dhuey & Smith, 2014; Grissom, Kalogrides, & Loeb, 2015). Therefore, it is clear from the literature that principals influence student achievement.

While it is well established that principal leadership makes a difference in student achievement, it has proved remarkably difficult to identify through empirical research precisely what it is about leadership that makes a difference for students. The job of a school principal is complex. One obvious way through which principals affect student achievement, albeit indirectly, is through their influence on the hiring, development, and retention of teachers (Béteille, Kalogrides, & Loeb, 2012; Grissom & Loeb, 2011; Leithwood et al., 2004; Loeb, Kalogrides, & Béteille, 2012). Several studies point to the critical role school leadership plays in a teacher's voluntary retention decision. For example, a study of North Carolina teachers found that those who expressed intent to depart their schools most commonly listed school leadership as a key motivation for leaving (Ladd, 2011). Other large-scale research confirms that teachers' retention decisions are sensitive to school leadership (Boyd et al., 2011; B. Fuller, Waite, & Torres Irribarra, 2016; Kraft, Marinell, & Yee, 2016).

While the research shows that a principal is key to setting the tone for a school and influencing whether high quality teachers leave or stay, the day-to-day responsibilities of a principal that comprise that influence are varied (Horng, Klasik, & Loeb, 2010). They often include tasks that range from basic facilities management to communicating with parents to providing teacher support. Collectively, all of the tasks a principal conducts day to day help to establish school culture, positive or negative, although the relative importance of some tasks is likely higher than others. Naturally, some principals might be quite adept at some aspects of the job and struggle with others. If some aspects of the job have a greater influence on student achievement, then we should seek principals with relative strengths in the attributes that matter most and, once hired, support and develop principals in those areas. For example, if providing useful feedback to teachers is ultimately more valuable to student progress than managing community relations, principals should be incentivized to provide better teacher feedback and less toward managing community relations. Unfortunately, a key question for which a concrete answer has eluded scholars is what principal behaviors and attributes are causally linked, directly and indirectly, to greater student achievement.

Over the last several decades, many researchers have wrestled with this problem and a number of correlational studies have identified leadership practices that are correlated with greater student achievement. These best practices have been classified into two broad categories: Instructional leadership (i.e., the degree to which the principal sets the direction of the school as it relates to the quality of instruction and the way in which instruction occurs) and transformational leadership (i.e., the degree to which the principal sets the overall vision and mission for the school and maintains the morale necessary to accomplish the vision). There is a long debate in education about the relative importance of instructional leadership versus transformational leadership. There is theoretical and some empirical support that both aspects of leadership are important for any school leader. Currently, however, instructional leadership is often viewed as the most important of the two when it comes to improving student outcomes. For example, one influential meta-analysis of more than 25 studies found that the relationship between instructional leadership and student achievement was up to five times greater than the relationship between transformational leadership practices and student achievement (Robinson, Lloyd, & Rowe, 2008). Many efforts to understand principal leadership have therefore focused on instructional leadership as the primary way by which principals influence outcomes. But, the studies on which these conclusions were based did not typically use rigorous measures of the principals' contributions to student achievement growth, meaning that the correlations between leadership practice and levels of student achievement were not necessarily causal.

Not all empirical research suggests that the tasks reflective of instructional leadership are absolutely more important than tasks not commonly associated with instructional leadership. Taking an agnostic approach to the question of instructional versus transformational leadership, two studies of Miami principals attempted to identify the specific activities that most closely correlated with positive school outcomes. In the first, the authors assessed principals on a comprehensive inventory of 42 practices and found that only one, a principal's organization and management skills, was correlated with student achievement growth (Grissom & Loeb, 2011). Importantly, their research found no connection between traditional measures of instructional management (e.g., using data to inform instruction, evaluating curriculum, coaching and observing teachers) and student achievement growth. The second used observational data and climate surveys to match principal activities to outcomes such as student achievement and teacher perceptions (Horng et al., 2010). They likewise found no relationships between traditional measures of instructional leadership and student achievement, and possible negative relationships between instructional activities and parent and teacher assessments of school climate.

While the research has been mixed, many principals are expected to be both transformational leaders and instructional leaders while also managing the day-to-day operations of what are increasingly becoming complex educational systems. Without a clear sense of their most important roles, some district and state leaders err on the side of expecting principals to do a little bit of everything. This has been characterized by some as expecting principals to attain status as "superprincipals" (Copland, 2001), a very lofty, and likely unrealistic, expectation. Principal evaluation has the potential to clarify the role of principals if it is designed to incent the most important behaviors and minimize those that take the principal away from their most important roles. Unfortunately, we do not yet have conclusive evidence of which activities fall into which categories and further research is warranted.

Principal Development and Improvement

Knowing that principal effectiveness varies and it makes a difference in student outcomes, most states require principals to have received training in a teacher leadership program and many states and districts offer opportunities for principals to receive formal and informal professional development. Despite the ubiquity of these opportunities, few studies have attempted to identify the causal effects of strategies designed to improve principal effectiveness (Hallinger & Bridges, 2017). Few researchers have used experimental designs to identify whether principals improve when they are provided with professional development opportunities. Stated more directly, "…Rigorous evaluations of programs designed to provide professional development to help train and retain school leaders are virtually nonexistent" (Jacob, Goddard, Kim, Miller, & Goddard, 2015, p. 314). One recent exception is an experimental study of balanced leadership approach in a sample of Michigan schools (Jacob et al., 2015). In this study, principals were provided with a commercially available McREL Balanced Leadership Professional Development program, a comprehensive professional development program that includes many of the principal practices that have been correlated with student achievement growth in prior research (Waters, Marzano, & McNulty, 2003). The authors of the experimental evaluation found that principals who participated in the program expressed greater efficacy at the conclusion of the three-year study, but teachers in the participants' schools noted no changes in the instructional climate of the school nor were there

any demonstrated changes in student achievement as a result of the professional development within the three years of receiving the professional development. The authors categorized the findings as "disappointing…given the widespread use of the program" (p. 328).

Taken together, prior research convincingly establishes the critical role of principals in the effective functioning of a school. This fact establishes firm grounds for evaluating principal effectiveness; good principals should be recognized and retained while ineffective principals should be remediated or dismissed. Unfortunately, knowing that some principals are more effective than others does not yield particularly useful insight as to what principals do to make them more or less effective. In short, principal effectiveness is critical but undefined. This poses an important limitation to principal evaluation systems.

History of Principal Evaluation

Principal evaluation has existed in one form or other for decades, and concern about effective principal evaluation is not new (see Stufflebeam & Nevo, 1991). The most significant development over the last several years is the move toward rigorous state-level systematic principal evaluation. Prior to the Race to the Top (RTT) grant competition, funded by the American Recovery and Reinvestment Act of 2009, principal evaluation was often mandated at the state level, but was largely pro forma and often left at the discretion of districts (Jacques, Clifford, & Hornung, 2012; Superville, 2014). In order for states to receive RTT funding, they had to implement "rigorous, transparent, and fair evaluation systems for teachers and principals" (US Department of Education, 2009, p. 59844). Among other requirements, states needed to design evaluation systems that differentiated principal quality by levels and included student growth as a "significant factor" in their evaluation (p. 59844). The availability of this initial funding and subsequent rounds of principal evaluation spurred states to implement statewide principal evaluation programs that met the rigorous requirements described in the notice. Others adopted principal evaluation as required by No Child Left Behind (NCLB) waivers granted during the period between the expiration of NCLB and the reauthorization of the Every Student Succeeds Act (ESSA). Currently, every state and Washington D.C. have some type of statewide principal evaluation plan in place (Fuller, Hollingworth, & Liu, 2015).

The RTT requirements for principal evaluation were a large departure from what states had been previously using to evaluate principals, and it was largely unfamiliar territory. While teacher evaluation had undergone extensive research and evaluation, such as in the large Gates Foundation funded Measures of Effective Teaching study (Kane, Mccaffrey, Miller, & Staiger, 2013), principal evaluation was One education researcher described principal evaluation as "the stepchild of teacher evaluation" (Superville, 2014) because it lagged so far behind teacher evaluation in terms of the volume of research on the topic and the professional scrutiny of the evaluation instruments and processes.

Current Status of Principal Evaluation

Much of the research noted above has focused on academic studies of particular attributes and practices of principals and the degree to which they are correlated with student achievement. Here we briefly summarize the research about existing principal evaluation tools and strategies.

Because principal evaluation continues to evolve, particularly in the use of measures of student achievement, we prioritize recent literature.

States and districts use a wide variety of principal evaluation strategies and tools. Currently, all states mandate some form of principal evaluation (Fuller et al., 2015). In their comprehensive review of state regulations and guidelines, Fuller et al. (2015) conclude that more than 75% of states listed professional growth and improvement as one of the purposes of their evaluation systems, and more than half indicated one of purposes was to improve student outcomes. Other purposes included improving instruction, identifying effective leaders, and making personnel decisions such as contract renewal or salary. Overall, nearly 70% (35 of 51) stated that principal evaluation results are intended to be linked to high-stakes decisions such as compensation, promotions, or continued employment. It is worth noting this is likely a lower bound; these were the states that listed high stakes decisions explicitly in their regulations, but it does not necessarily mean that the remaining states do not use them in high-stakes decisions.

Despite the high stakes nature of many principal evaluation approaches, very few are backed by research. In a comprehensive review of the literature around principal evaluation, Davis (2011) concluded the research base around principal evaluation "lacks volume and depth" (p. 2) and is "extremely thin" (p. 6). This lack of research is no doubt partly due to the fact that there is still a basic lack of consensus around the qualities and practices of an effective leader.

There are two ways of thinking about the validity of principal evaluation approaches, and both are necessary for a robust evaluation system. The first is that the evaluation strategy must be a valid measurement tool in that it consistently measures what it purports to measure. In order to meet this standard, the evaluation should consistently produce a result that accurately reflects the actual practices or behaviors of the principal. Consistent with this idea, the evaluation tool must include detailed protocols for how the evaluation is to be completed such as the number of observations a supervisor is to make, the minimum survey response rates from teachers, or other detailed implementation protocols. Many principal evaluation tools fail to meet this basic requirement. For example, in a study of 65 principal practice instruments, Goldring et al. (2009) found only two that included any information about the measurement properties of the instrument.

A second way in which evaluation systems must be valid reflects whether the evaluation tool measures what we want it to measure. This is commonly done by establishing what is often called content validity, or showing that the principal evaluation tool focuses on measures that are hypothesized to be connected with the behaviors or practices of effective principals. The Professional Standards for Educational Leaders (formerly known as the ISLLC standards) form the set of professional standards which are generally regarded by researchers and practitioners as the set of standards that reflect effective practice (National Policy Board for Educational Administration, 2015). In their 2009 review of principal assessment tools, Goldring et al. found substantial variability in the types of leadership behaviors and practices that were reflected, suggesting what they viewed as a troubling lack of agreement across tools. For example, according to their coding they found that the number of items measured ranged from 10 to 180, and only 25 of 65 tools that included all six core components of the ISLLC standards (p. 30).

The lack of content validity is consistent with the general lack of agreement about effective leadership practice. A stronger way of establishing validity is by connecting the evaluation tool directly with the outcomes that are indicative of effective leaders using causal

research methods. As highlighted above, there is very little evidence that current principal evaluation strategies are connected with growth in student achievement. We note two exceptions below, Pennsylvania and Tennessee.

Theory of Change: Principal Evaluation

We begin by assuming that the primary purpose of a school is to increase student achievement, and therefore the effectiveness of a school is directly related to student achievement. A principal can influence a school's effectiveness through a variety of mechanisms. For example, a principal can have direct affects through interactions with students; the anecdotal principal who greets her students by name each morning could have a lasting mentoring effect on students. Likewise, the principal plays a role in establishing the culture of the school which may influence student performance. The principal also has influence over teacher hiring, development, and retention, and influences student achievement indirectly through that channel. In reality, the principal's job is complex and involves a variety of activities. Some activities, such as teacher observation, might have a very intuitive connection to student achievement. Others, such as managing school schedules, might not be as clear.

Principal evaluation could serve two purposes, which are not mutually exclusive. The first is that evaluation serves as a means of providing feedback to encourage the growth and development of leadership skills, ultimately improving principal effectiveness. The second is that evaluation serves as an accountability device to provide the proper incentive for principals to exert effort toward the most meaningful efforts. An effective evaluation system could accomplish both tasks.

Evaluation to improve effectiveness. Because of the complexity of a job, it is unlikely that one person will be adept at every task that is required. For example, a principal who is quite capable at mentoring new teachers might struggle managing budgets. Evaluation could provide a principal targeted feedback that allows the principal to work on areas of weakness. Likewise, a well-designed evaluation process will give the principal clear guidance on the relative importance of the various tasks. With regular feedback through a transparent evaluation process, a principal will have clarity on how to spend his or her time in a way that is consistent with improving student performance.

Evaluation as accountability. Theory on social behavior suggests that managers, such as principals, might have incentives that are inconsistent with the ultimate goal of the organization. For example, a principal might have the incentive to try to satisfy teachers rather than engaging in difficult conversations or decisions that might ultimately lead to better student outcomes. If that is the case, the evaluation process provides supervisor with an opportunity to more directly observe the principal behavior and properly align incentives. Likewise, accountability evaluation provides supervisors the opportunity to recognize principals who are unable or unwilling to engage in the behaviors that are best for the school and remove them as necessary. An evaluation system that is designed strictly as an accountability device has the potential advantage that it can be based only outcomes (i.e., growth in student achievement) and does not need to include any measure of the principal's behaviors. Such an evaluation system would be less costly to implement if accurate measures of student achievement were readily available through existing data.

A well-designed evaluation system could accomplish both purposes. In order to do so, the evaluation must accurately identify the practices and behaviors that are connected to student achievement and accurately measure the extent to which the principal engages in those practices. Likewise, the evaluation system would need to be able to accurately measure the principal's contribution to student achievement to fulfill the second purpose. In the next section, we examine these two aspects of principal evaluation.

Components of Principal Evaluation

As with teacher evaluation, most systems of principal evaluation include multiple components and are not based on a single measure. A principal evaluation system could include many components, but in the next section we summarize the most common elements that comprise most principal evaluation systems grouped into two general categories: inputs and outcomes. The body of inputs reflect what the principal is doing in the school and includes all evaluation strategies that focus on measuring principal behaviors and practices. Outcomes, in contrast, focus on measuring the *effects* of a principal's actions. Because of the requirements of RTT and NCLB waivers, most states have adopted principal evaluation systems that include both input measures and student achievement measures.

Evaluating Principals on Inputs

School leadership plays a key role in setting the tone of a school in terms of the social aspects of an organization such as morale, trust, and general teacher satisfaction, but also in terms of the orientation of the school toward high quality instruction. The first class of evaluation strategies focuses on measuring a principal's actions that are hypothesized to contribute to these dimensions of the school. Principals can be evaluated on leadership inputs either observationally, through site visits and document reviews, or by way of surveys of stakeholders. Both are common and have advantages. Examples of each are provided below.

Survey Measures of Principal Effectiveness. Two of the best-known survey tools designed to assess a principal's leadership are the Principal Instructional Management Rating Scale (PIMRS) and the Vanderbilt Assessment of Leadership in Education (VAL-ED). Both are 360° survey assessments in which feedback from teachers, principal supervisors, and the principals themselves are incorporated to assess the degree to which the principal practices the targeted behaviors at the school. Both are designed to provide feedback to the principal from multiple sources.

The PIMRS was developed by Philip Hallinger and has been around since the 1980s. It is a 360° survey assessment The PIMRS focuses 10 instructional leadership areas that fall under three primary dimensions: Defining the school mission, managing the instructional program, and developing the school learning climate program (Hallinger & Murphy, 1985). The PIMRS has undergone many studies of its content and construct validity and interrater reliability and has been generally found to be a reliable instrument to measure principal practice, although the teacher reports were generally more closely related to other measures of principal practice than were principal or supervisor reports (Hallinger, 2008)

The VAL-ED measures six components and six processes of instructional leaders. The components include high standards for student learning, rigorous curriculum, quality instruction,

culture of learning and professional behavior, connections to external communities, and performance accountability. Processes are planning, implementing, supporting, communicating, monitoring, and advocating (see Goldring, Porter, Murphy, Elliott, & Cravens, 2009). Each component is based on one of the Interstate School Leaders Licensure Consortium (ISLLC) standards and are consistent with the updated version of the ISLLC standards, the Professional Standards for Educational Leaders (PSEL). The VAL-ED is also administered as a 360° assessment, where teachers, supervisors, and principals themselves are required to provide specific evidence of each of the behaviors. The VAL-ED has been shown to be a reliable measure with well-documented psychometric properties (Porter et al., 2010), but to date has not been shown to be predictive of student achievement. In other words, the VAL-ED is likely a reliable measure of whether the principals are engaged in practices that are hypothesized to be connected with student achievement, but we still do not have evidence that a principal who scores higher on the VAL-ED leads to greater student success than a principal who scores lower.

Other surveys also exist to measure leadership. Some, such as the Comprehensive Assessment of Leadership for Learning (CALL) are available for states to adopt and adapt to local needs. Others have been developed by states and even districts based on the ISLLC/PSEL standards. Many of these are very similar in construction. For example, a recent working paper found high correlations between the CALL and the VAL-ED components (Goff, Salisbury, & Blitz, 2015).

Observational Measures of Principal Effectiveness. A second approach to measuring principal instructional leadership is through direct observation. In this approach, an external rater assesses the principal on predetermined criteria. The observation can consist of a formal observation protocol, in which the rater is physically present in the school to observe the principal in day-to-day activities according to a guided rubric, or the observation assessment rely on informal observation and/or document review, typically by the principal's supervisor. In the case of informal observation, the evaluator is often expected to apply a rating more holistically on the sum of all interactions with the principal throughout the year. These types of assessment often require principals to provide evaluators with portfolios illustrating work the principal has done throughout the year to demonstrate the key practices.

Historically, many principal assessments were strictly holistic observations in which supervisors rated principals according to checklists of principal practices. These types of observational often give great latitude to the observer in their interpretation of a principal's actions, and are by nature more subjective. More recent evolutions of the observational approach rely on detailed rubrics that require observers to rate principals on various categories of behavior and provide specific examples are indicative of different levels that behavior. In the state examples below, we provide three examples of rubrics that are used in the state. Each was developed by the respective states in consultation with national experts and stakeholders in the state.

To date, we are aware of only one such observational system that has been empirically connected to student achievement. Researchers recently examined Pennsylvania's Framework for Leadership (FFL) and found that principals with high FFL scores were associated with greater effectiveness in improving student achievement (McCullough, Lipscomb, Chiang, & Gill, 2016). The correlations were relatively small, but the authors were able to conclude they reflected real differences that were not just due to sampling error.

Grissom et al. (2015) evaluated an observational system and found that it was correlated with measures of school value-added and measures of student demographics. Unfortunately, as described in more detail below, school value-added is a biased measure of principal effectiveness. This finding suggests that observational measures of principal evaluation could be similarly biased. This could happen if, for examples, observers tend to rate principals as being more effective on observational measures if students are higher achieving, even if the achievement has little to do with the principal's behavior.

Student Achievement Measures of Principal Effectiveness. As we discuss in more detail in our discussion of teacher evaluation, teacher evaluation has moved steadily in the direction of holding teachers directly accountable for student learning, as measured by standardized tests. Recently, many states have adopted similar requirements for principal evaluation. Currently all states include some form of student performance as part of their official principal evaluation system (E. Fuller et al., 2015). Virginia's *Guidelines for Uniform Performance Standards and Evaluation Criteria for Principals* (2015) indicate student academic progress should constitute 40 percent of a principal's evaluation.

As noted above, the academic literature on principal effectiveness has begun to amass an impressive body of evidence that principals have meaningful effects on student achievement. In many of the recent quantitative studies that have documented a persistent principal effect, the researchers used a statistical technique in which they examined student achievement before and after a principal changed schools and attributed those changes, positive and negative, with the principals who had changed jobs. They conclude from these studies that principals do have a meaningful causal effect on student achievement.

Knowing that principals vary in their effectiveness in terms of student achievement does not necessarily indicate that we can easily measure that influence. Except in the case of principals changing jobs, there is not a clear way of isolating the effect of the principal from other factors beyond a principal's control. Analogous to the challenge of measuring teacher effects, external factors beyond a principal's control also affect student achievement and must be accounted for when attempting to attribute achievement to principals. For many of the same reasons already described for teachers, it is inappropriate to conclude that *levels* of student achievement reflect the contribution of the principal. Such a strategy inappropriately attributes all of the factors, positive and negative, that affect student achievement to the principal. Any fair assessment of the principal's contribution to student learning should ultimately compare what changes have occurred in student achievement with a given principal compared with what would have happened with a different principal. A principal evaluation system that rewards principals based on levels of performance will systematically overestimate the effectiveness of principals in affluent schools, which typically have more resources and greater family support, and underestimate the effectiveness of principals in schools in low-income neighborhoods.

One solution that many states have proposed to overcome this challenge is to use value added modeling (VAM) in an analogous way as it is used to evaluate teachers. This approach makes some intuitive sense at first blush, but several researchers have urged caution in implementing VAM as a strategy to evaluate principals because principals are likely subject to persistent conditions beyond their control that are not accounted for with VAM (Chiang, Lipscomb, & Gill, 2016; E. Fuller et al., 2015; Grissom et al., 2015). For example, while many principals have some discretion over teacher hiring, most do not have complete control over the composition of their teacher workforce due to collective bargaining agreements, district transfer

policies, or even the relative attractiveness of the neighborhood in which the school is located. If school-level VAM does not accurately account for the portion of teacher quality over which a principal has no control, it will inappropriately credit or penalize principals for factors beyond their control. This is an important distinction between teacher-level VAM and school-level VAM. Teacher level VAM is more likely to identify the effectiveness of a teacher because we can directly account for factors beyond the teacher's control through the use of student controls. School-level VAM would have to likewise account for all factors beyond a principal's control, but those controls are much more difficult to include in models.

Two recent rigorous principal effectiveness measurement studies found that school-level VAM has very little or no correlation with the types of principal effectiveness measures that have been identified by examining changes in principal leadership within schools (Chiang et al., 2016; Grissom et al., 2015). Both papers urge policy makers to use extreme caution in the ways in which student achievement is incorporated into principal evaluation. In a measured recommendation, Grissom et al. (2015) suggest that student achievement could be included in formative principal evaluations to emphasize to the principal the importance of improving student outcomes, but it should be acknowledged that the measures could inherently be unfair to some principals: "[An evaluation system] that uses [student achievement] to highlight the value of student learning while balancing them with other measures and understanding of their shortcomings may have positive effects even if the measures are imprecise or biased" (p. 25).

The fact that VAM methods to identify principal effects is currently unreliable does not mean it will always be that way. This is a topic of great interest among policy makers and researchers and will undoubtedly be the subject of methodological improvements over time. A recent paper suggests an alternative approach to measuring principal VAM that could prove promising subject to greater scrutiny by experts in measurement and evaluation (McCullough et al., 2016).

Conclusion

Research on principal evaluation, in its current evolution that includes formal measures of principal inputs and student achievement outcomes, is still nascent. Unfortunately, the best available research suggests that the inclusion of student achievement measures should be interpreted with caution. There is convincing evidence that the measures currently available are biased; they inappropriately favor principals in schools with more advantaged students and punish principals in schools with less advantaged students. A high stakes evaluation environment in which principal job security depends on these measures could create incentives for principals to seek positions in the most advantaged schools and spurn positions in disadvantaged schools. This has obvious negative implications for already disadvantaged students.

The state of the research on the observation measures included in many principal evaluation systems have yet to show conclusively that they are connected to student achievement, although a few notable exceptions have begun to emerge. With time and additional research, there is reason to believe these measures will be refined and improved. In the meantime, the observation rubrics often reflect the best professional practice in the field. They represent an opportunity for principals to focus on particular behaviors and could serve as a professional development opportunity for principals who struggle with the overwhelming complexity of the job.

What is a state to do if we are unsure of the alignment between evaluation and "true" principal effectiveness? As suggested by Grissom et al. (2015), including measures of student achievement in an evaluation could have the positive benefit of focusing principals on their ultimate responsibility to improve student outcomes, even if they are biased and/or imprecise measures of actual principal effectiveness. They also have the potential to focus principals on the activities that we feel most confidently lead toward that end. The best use of evaluation at this point may be to use evaluation primarily as a leadership development tool. If it is to be used high stakes decisions such as principal dismissal, caution should be exercised to be certain that the behaviors of the principal are inconsistent with effective school management.

Statewide Principal Evaluation in Practice: Three Illustrative Examples

Principal evaluation has been around for decades, but state-driven principal evaluation systems are a relatively new phenomenon. Below we examine three state systems that reflect current practice in state-sponsored principal evaluation. These systems were chosen for closer examination because they reflect two state systems that have been identified with some research to be connected to student outcomes (Tennessee and Pennsylvania) or, in the case of Washington D.C., grow out of a system in which the teacher evaluation has been linked to higher student outcomes (Dee & Wyckoff, 2015).

Across the three systems, all reflect a multi-dimensional approach to principal evaluation. All include supervisor observation as half the evaluation score with various measures of student achievement representing the other half. The observation protocols vary somewhat, but many familiar themes emerge with a focus on instructional leadership. The three systems all use state-developed observation protocols that are based on existing instructional leadership research. Ultimately, the three are quite similar in terms of content and structure. All three states collect data for observational scores based on direct observation as well as other artifacts to support the ratings, and all three require evaluators to make holistic judgments and rate principals on multiple dimensions according to a discrete 1 to 4 (or 5) scale.

Each of the systems include measures of student achievement that make up half of the principals' ratings. The states use different metrics for student achievement, but in each case they include components that reflect schoolwide measures of student achievement that are consistent with the state's school accountability measures. Each state's evaluation includes measures of student achievement that are customized to the level of the school (elementary, middle, or high school). All provide some local flexibility for goals and other metrics to be developed by the principal and/or the district. Below we describe each of the state systems in greater detail.

The greatest difference are in the consequences, at least as defined according to official state policy. For example, Tennessee provides little information about what the consequences of evaluation should be except to state that they are to be used for human resource decisions and that principals in the lowest category are subject to additional training or monitoring. In contrast, Washington D.C. provides fairly concrete guidelines about the consequences of principals scoring at each of the three performance levels. They are also the only system to provide monetary rewards for principals scoring in the top category. We cannot rule out that those types

of rewards are provided to effective principals on a local basis in other states, but the state systems do not mandate any explicit reward structure.

Below we provide more information on each of the state systems including the components, the consequences, and any available research about the effectiveness of the evaluation systems.

Tennessee

Overview

While principal evaluation has been in place for decades, Tennessee was one of the first states to adopt a rigorous statewide principal evaluation system that included specific measures of student achievement (Grissom, Blissett, & Mitani, 2017), called the Tennessee Educator Acceleration Model (TEAM). Each principal's evaluation score is based 50% on student achievement and 50% on supervisor evaluations of principal performance using the state-created TEAM rubric.

Student achievement

Half of the administrator's evaluation score is based on student achievement. A total of 35% of the administrator's total evaluation score is based on the one-year school-wide growth as measured by the Tennessee Value-Added Assessment System (TVAAS) and 15% is based on meeting other achievement goals that the administrator sets in collaboration with the administrator's supervisor. The TVAAS is calculated by a third-party vendor, SAS, and is the measure the state uses for school accountability reporting.

TEAM Observation

Each principal is rated by a supervisor according to the TEAM rubric. The rubric was developed collaboratively with superintendents', supervisors', and principals' councils and is informed by research around instructional leadership (see Tennessee Department of Education, 2017). The rubric is publicly available[1] and is included as an appendix to this report.

The TEAM observation rubric rates principals based on four standards:

- Instructional Leadership for Continuous Improvement
- Culture for Teaching & Learning
- Professional Growth and Learning
- Resource Management

Within each standard, the supervisor rates principals based on multiple indicators. In total, the four standards include 17 indicators (see Table 1). Each of the 17 indicators is described in the rubric in detail at level 5 ("Significantly Above Expectations"), level 3 ("Meets Expectations"), and level 1 ("Significantly Below Expectations"). The rubric cites specific examples of practices

[1] See http://team-tn.org/wp-content/uploads/2013/08/TEAM-Admin-Evaluation-Rubric-20161.pdf

that are consistent with each of the levels and provides guidance to the supervisor as to sources of evidence that can be used to determine the principal's level.

In addition to direct observation by the principal's supervisor, TEAM also requires supervisors to conduct teacher evaluations of principals in the form of a teacher perception survey. Evaluators are allowed to use a survey instrument of their choice, but the state provides one option for use. The sample survey asks teachers to weigh in on the extent to which they agree that the principal engages in practices that are consistent with effective instructional leadership. The structure is designed to be parallel to the four standards and the 17 indicators that form the observation rubric. The scores from the surveys are not explicitly included as a percentage of the principal's evaluation score, but they are intended to be used to form the supervisor ratings of the principals.

Table 1. Tennessee TEAM Rubric Standards and Indicators

Standards	Indicators
Instructional Leadership for Continuous Improvement	• Capacity building • Data analysis & use • Interventions • Progress monitoring
Culture for Teaching & Learning	• Leveraging educator strengths • Environment • Family involvement • Ownership • Recognition & celebration
Professional Growth and Learning	• Evaluation • Differentiated professional learning • Induction, support, retention, & growth • Teacher leaders • Self-practice
Resource Management	• Community resources • Diversity • Employee & fiscal management

Evaluation Process

Principals are to be evaluated throughout the year on a continuous basis. Evaluators are instructed to visit schools at least twice, at least once during the first half of the year and at least once during the second half of the year. One of the visits must be unannounced. Evaluators are instructed to use the evaluation rubric not as "a checklist, but should be used to weigh the preponderance of evidence over time" (p. 6). The state provides some leeway as to how the evaluators should interact with principals, but provides some guidance for best practices:

- Conduct observations as a team
- Use multiple observers, but have at least one evaluator consistent across both observations
- New principals are always to have the superintendent be part of the observation team

At the conclusion of the year, the student achievement scores and the TEAM evaluation scores are compiled. Each principal then receives an overall score category from 1 ("significantly below expectations") to 5 ("significantly above expectations").

Consequences

By state statute, the principals' evaluation scores are to be used "to inform human capital decisions, including, but not limited to individual and group professional development plans, hiring, assignment and promotion, tenure and dismissal, and compensation" (Tennessee State Board of Education, 2017). The effectiveness ratings are published by the state, and those that fall below an acceptable range are "subject to additional training and monitoring" (p. 1).

Conformity with Best Practices

The TEAM system relies heavily on school value added, which has been shown to be a biased measure of principal performance. If the system is used primarily for feedback and improvement, this might not pose a problem. The TEAM observation rubric is closely aligned with what the field has designated as best practices for effective leadership. The use of a teacher evaluation survey is an innovative approach of the TEAM system and is consistent with research showing principals have a meaningful impact on teachers' perceptions of the work environment.

Research on TEAM

Tennessee has conducted annual evaluations of the TEAM system. In the most recent available review (2014-15), administrators reported very favorable about the evaluation system. More than 80% of principals felt that the evaluation system led to better leadership practices and ultimately better student achievement. A significant majority (81%) reported that the system helps them identify areas for improvement and 95% felt they had improved as a leader over the previous year. The results from 2014-15 suggest that a large majority (70%) principals exceed expectations as rated by TEAM while only 45% exceed expectations when considering only student achievement.

We are aware of only one paper to examine the properties of the TEAM observation evaluation component. A working paper by Grissom et al. (2017) examined the relationship between principal TEAM ratings and other measures of principal effectiveness. Their primary conclusions are as follows:

1. TEAM ratings are internally consistent and stable over time, although there is some evidence that the TEAM rubric may not accurately differentiate between different leadership domains.
2. TEAM ratings are correlated with principal characteristics. Some of these are as expected (e.g., more experienced principals have higher ratings, on average) while others point to potential bias in the instrument (e.g., female principals receive higher ratings, on average, while Black principals receive lower ratings, on average).
3. TEAM ratings are associated with school characteristics. Some point to TEAM as being a significant predictor of effective leadership (e.g., higher TEAM ratings are

associated with higher student achievement results) while others suggest potential bias (e.g., schools with fewer free/reduced lunch recipients have higher scoring principals).

The authors are cautious to connect the measures on the TEAM causally to student achievement. Likewise, the study does not attempt to determine how leadership improved as a result of the introduction of the evaluation system. They call for greater research to explore these questions.

Pennsylvania

Overview

The Pennsylvania Act 82 of 2012 established a statewide principal evaluation system, although the first statewide implementation was not completed until 2014-15. In this system, all principals and assistant principals receive annual evaluation scores made up of measures of student performance (50%) and observational practice ratings (50%).

Student Achievement

The student performance data portion of the principal's evaluation is made up of three components. Fifteen percent of the evaluation is based on building level achievement data. The state uses a formula to rate the academic performance of the school based on achievement levels, progress towards closing achievement gaps, and overall growth in achievement. The overall school rating ranges from 0-107, which is then converted to a 0-3 scale based on published cutoff scores. The second component, teacher-level measures, also receives 15% weight in the final principal evaluation. Teacher level measures are aggregated measures of teacher effectiveness connecting teacher performance evaluation to student achievement. The final student performance component, elective data, receives 20% weight in the final principal evaluation. Elective data are to be determined by the LEA and principal based on approved metrics from the state including district tests, industry certification exams, and student projects or portfolios. The state provides significant leeway to the LEA in determining what elective data to include. Each of the latter two components are scaled from 0-3. The principal's overall score is the weighted average of the three components.

Framework for Learning Observation

Principals are evaluated holistically by their supervisors according to a rubric called the Framework for Leadership (FFL). The rubric was developed by Pennsylvania state officials in consultation with professional experts, and the structure is meant to mirror the Danielson Framework for Teaching that is used to evaluate teachers in the state. The rubric for the FFL is publicly available[2] and is included as an appendix to this report.

The FFL includes 4 domains that include a total of 20 components (see Table 2):

- Strategic/Cultural Leadership
- Systems leadership
- Leadership for learning
- Professional and community leadership

[2] See http://static.pdesas.org/content/documents/Principal_Rubric.pdf

Supervisors are instructed to evaluate each domain based on the preponderance of evidence collected through site visits, observations, and evidence provided school leaders (McCullough et al., 2016). Each component is rated on a 0 (=failing) to 3 (=distinguished) scale and the supervisor then provides an overall rating for the domain on the same scale (i.e., 0-3). The rubric cites specific examples of practices that are that should be observed for each component at each level (failing, needs improvement, proficient, and distinguished). Each of the four domains receives equal weight in the principal's final evaluation such that the FFL score for the principal is the average of the four domain ratings. Thus, a principal's overall FFL rating ranges from 0-3.

Table 2. Pennsylvania's Framework for Leadership Domains and Components

Domains	Components
Strategic/Cultural Leadership	• Creates an organizational vision, mission, and strategic goals • Uses data for informed decision making • Builds a collaborative and empowering work environment • Leads change efforts for continuous improvement • Celebrates accomplishments and acknowledges failures
Systems Leadership	• Leverages human and financial resources • Ensures a high quality, high performing staff • Complies with federal, state, and LEA mandates • Establishes and implements expectations for students and staff • Communicates effectively and strategically • Manages conflict constructively • Ensures school safety
Leadership for Learning	• Leads school improvement initiatives • Aligns curricula, instruction, and assessment • Implements high quality instruction • Sets high expectations for all students • Maximizes instructional time
Professional and Community Leadership	• Maximizes professional responsibilities through parental involvement and community engagement • Shows professionalism • Supports professional growth

Evaluation Process

Principals are evaluated by their supervisor holistically and receive regular feedback on their performance. Principals must be evaluated at least once annually, and state regulations require principals to be given feedback within 7 days of any observations.

Consequences

Principals ultimately receive one of four ratings: Failing (score= 0.00-0.49), Needs Improvement (score=0.50-1.49), Proficient (score=1.50-2.49) or Distinguished (score=2.50-3.00). Principals receiving a "Failing" rating are required to participate in a performance

improvement plan, which includes an observation schedule with "intense supervision", and they are considered to be performing unsatisfactorily. Principals who receive a "Needs Improvement" rating are considered to be performing satisfactorily unless it is their second such rating within a 10-year period. If it is their second "Needs Improvement" rating, they are considered unsatisfactory. The consequences for an unsatisfactory rating are not described in detail in state regulations.

Conformity with Best Practices

The Pennsylvania system relies on building-level achievement data and school-level value added data, which have been shown to be biased measures of a principal's performance. Twenty percent of the principal's evaluation is based on meeting goals, which could be a better indicator of principal performance as long as the goals are consistent with the objectives of the principal. The FFT is closely aligned with what the field has designated as best practices for effective leadership.

Research on Pennsylvania's System

Two studies of the Pennsylvania evaluation system have examined the relationship between FFL and student achievement. Teh, Chiang, Lipscomb, & Gill (2014) examined an early pilot of the program to determine the degree to which there was meaningful variation in principal evaluation scores and the degree to which evaluation scores were correlated with student achievement. They found that there was not much variation within the principals (i.e. principals who scored high on one component tended to score high on other components) or across principals (i.e., only 5 percent of principals were rated as needing improvement). They also found no evidence that principals' FFT scores were correlated with the principals' estimated contributions to student achievement.

A second study examined a second pilot of the FFT and found different results. McCullough et al. (2016) confirmed the findings from Teh et al. (2014) that a large majority of principals were rated as proficient or distinguished. But, in contrast, they found evidence that principals who scored higher on the FFT were correlated with greater contributions to student achievement, although the correlations were modest. These results suggest that the FFL could identify important practices that are related to the principals' contributions to student achievement.

The conflicting findings of these studies are not necessarily incongruous. Both were pilot studies, which were relatively small. For a variety of reasons, the findings from the initial study could be an artifact of a small sample size and/or the first year of implementation of the pilot.

Washington D.C.

Washington D.C. Public Schools (DCPS) has designed School Leader IMPACT (SL-IMPACT) to assess all assistant principals and principals in the district. The evaluation parallels their Teacher IMPACT that rates all teachers based on effectiveness. The SL-IMPACT score is made up of two components: Student outcome goals and Leadership Framework Standards, or observations of the principals' leadership practices. Each component comprises 50% of the principal's overall evaluation score, although the exact composition of the student outcome goals is different for elementary, middle, and high school teachers. The process is described in detail in

a regularly updated guidebook made publicly available[3], which is attached as an appendix to this report.

Student Outcome Goals

Each year every principal sets student outcome goals in consultation the teachers in their school and in collaboration with their supervisors. One set of goals must be related to increasing levels of math and literacy proficiency and decreasing the percentage of students scoring in the lowest proficiency category as measured by the PARCC exam, which is the standardized test DCPS uses for accountability purposes. These PARCC performance goals are weighted equally in their evaluation.

In addition to the academic proficiency goals, principals must also set goals that are related to other school performance measures. Principals have greater flexibility in determining these goals. They can include aspects of student experience like school culture or other measures of student learning. These goals are flexible, but they must be approved by supervisors and must use a district-approved metric to allow for clear measurement of progress. At the middle and high school levels, principals are also required to set goals related to specific metrics including promotion rates, school climate, and graduation rates. The weight of these goals in their evaluation depends on the level (middle or high) of the school in which they are working.

Leadership Framework

The other half of SL-IMPACT is determined through the DCPS Leadership Framework (LF) observations. The LF identifies 20 themes embedded within 6 standards: Instruction, Talent, School Culture, Operations, Family and Community, and Personal Leadership (see Table 3). The observation rubric includes specific examples of behaviors that correspond with each level of performance on every theme, ranging from Level 1 (lowest) to Level 4 (highest). The rubric also provides guidance to the observers on sources for gathering information on each of the standards.

With the exception of Instruction, which receives 25% weight, the components are weighted equally at 15%. The framework was developed by DCPS after consulting with experts in the field and combining several existing leadership evaluation tools or strategies being used in other places.

Table 3. Washington D.C.'s Leadership Framework Standards and Themes

Standards	Themes
Instruction	• Develops the school's instructional vision and goals • Oversees effective school- and classroom-level planning • Ensures effective classroom instruction • Establishes a culture of data-driven instruction
Talent	• Identifies and strategically places outstanding talent • Evaluates staff members, provides support, and removes low performers • Retains key staff and builds leadership capacity
School Culture	• Creates a positive, student-centered environment

[3] Guidebook available here: https://dcps.dc.gov/publication/dcps-school-leader-impact-guidebook-principals

	• Ensures students meet high academic and behavioral expectations • Implements effective interventions that support student success
Leadership Actions	• Efficiently manages school operations • Maximizes impact of limited resources • Fulfills all legal and policy requirements
Family and Community	• Builds relationships with families and community members • Efficiently responds to families' inquiries and concerns • Shares information with families to support their children's success
Personal Leadership	• Engages in continuous self-improvement • Communicates effectively • Demonstrates cultural competence • Perseveres in the face of obstacles

The rubric includes descriptions for each action that range from Level 4 (highest) to Level 1 (lowest). This rubric is published and provided to school leaders prior to the start of the year so it is clear what they will be evaluated on as part of the LF. The LF also assesses "on-track" indicators of the school data such as the percentage of students on track, the percentage of students who took the SAT, etc. Principals are expected to provide artifacts to demonstrate their progress toward some of the key indicators.

Evaluation Process

Principals set goals in the summer and early school year (August-October) in consultation with their area superintendent. The goals must be measurable and related to their overall comprehensive school plans. Beginning in November, principals have their first LF assessment with the instructional superintendent. This includes collecting artifacts and on-track indicators from the school, the principal conducting a self-assessment, and the instructional superintendent making judgment based on observations. After the instructional superintendent discusses the evaluation results with the superintendent, the instructional superintendent hosts a one-on-one meeting with the principal to discuss the results of the performance evaluation. If their evaluation score is below a certain level (2.6 out of 4), they receive a second performance evaluation by April and all others receive it by June. The two LF scores are both included in the principal's overall score, but the first LF is weighted at 20% while the second is weighted at 30% to generate the overall LF for the year.

After LF scores have been compiled and PARCC assessments have been conducted, the final SL-IMPACT scores are calculated for each principal based on the weights relevant for their level. The principal's LF score is worth a total of 200 points and the goal scores are worth a combined total of 200 points. Principals who score below 300 points are rated "Minimally Effective". Those above 300 but below 350 are rated "Effective" and those scoring at or above 350 are rated "Highly Effective".

Consequences

Principals who are rated Highly Effective are eligible for performance bonuses of up to $30,000, with higher bonuses going to those in high needs schools, and can be considered for district leadership opportunities and awards. If they are rated Highly Effective and they score at least 3.0 on the LF, principals are also eligible for a 3-year contract instead of the typical 1-year contract. Those who receive an "Effective" rating are considered to be performing as expected. Principals who receive two consecutive Effective ratings with an average LF score of 3.0 are also eligible for a 3-year contract. Those who are rated Minimally Effective are flagged for improvement. They do not receive their salary step for the subsequent year and may be considered for non-renewal of their contract, although non-renewal is not automatically an outcome for minimally effective principals.

Conformity with Best Practices

The SL-IMPACT system does not use value added measurement to evaluate principals and instead uses the principals' success in meeting goals designed around changes in performance. While the field has not yet identified whether these types of measurement strategies are unbiased measures of principal performance, the avoidance of value added measures and overall achievement levels is consistent with best practices. The system also requires principals to set goals related to other achievement metrics, which is likely helpful in broadening the principals' focus beyond achievement test scores. The LF system is consistent with best practices in the literature. The process of having two observation minimizes the measurement error that could plague a single observation metric. Of the three systems highlighted in this brief, the SL-IMPACT system has the most clearly described incentive structure. It is not yet clear whether this is effective at improving principal performance.

Research

We are unaware of any research on the effects of SL-IMPACT. But, a study of the teacher IMPACT system has identified causal evidence that the performance categories induced teachers to either improve their performance or to depart (Dee & Wyckoff, 2015), providing some hope that it may have similar effects for principals. There are sufficient differences between the two systems, particularly in the consequences that accompany the performance ratings, that we cannot necessarily conclude that it will have identical effects on principals. Research is currently under way to measure this.

References

Béteille, T., Kalogrides, D., & Loeb, S. (2012). Stepping stones: Principal career paths and school outcomes. *Social Science Research, 41*(4), 904–919.

Boyd, D., Grossman, P., Ing, M., Lankford, H., Loeb, S., & Wyckoff, J. (2011). The Influence of School Administrators on Teacher Retention Decisions. *American Educational Research Journal*, *48*(2), 303–333.

Branch, G. F., Hanushek, E. A., & Rivkin, S. G. (2012). *Estimating the effect of leaders on public sector productivity: The case of school principals*. National Bureau of Economic Research.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, *104*(9), 2633–2679.

Chiang, H., Lipscomb, S., & Gill, B. (2016). Is school value added indicative of principal quality? *Education Finance and Policy*, *11*(3), 283–309.

Coelli, M., & Green, D. A. (2012). Leadership effects: School principals and student outcomes. *Economics of Education Review*, *31*(1), 92–109.

Copland, M. A. (2001). The myth of the superprincipal. *Phi Delta Kappan*, *82*(7), 528–533.

Davis, S., Kearney, K., Sanders, N., Thomas, C., & Leon, R. (2011). *The policies and practices of principal evaluation: A review of the literature*. San Francisco, CA: WestEd.

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267–297.

Dhuey, E., & Smith, J. (2014). How important are school principals in the production of student achievement? *The Canadian Journal of Economics / Revue Canadienne d'Economique*, *47*(2), 634–663.

Fuller, B., Waite, A., & Torres Irribarra, D. (2016). Explaining Teacher Turnover: School Cohesion and Intrinsic Motivation in Los Angeles. *American Journal of Education*, *122*(4), 537–567.

Fuller, E., Hollingworth, L., & Liu, J. (2015). Evaluating state principal evaluation plans across the United States. *Journal of Research on Leadership Education*, *10*(3), 164–192.

Goff, P., Salisbury, J., & Blitz, M. (2015). Comparing CALL and VAL-ED: An Illustrative Application of a Decision Matrix for Selecting Among Leadership Feedback Instruments.

Goldring, E., Cravens, X. C., Murphy, J., Porter, A. C., Elliott, S. N., & Carson, B. (2009). The Evaluation of Principals: What and How Do States and Urban Districts Assess Leadership? *The Elementary School Journal*, *110*(1), 19–39.

Goldring, E., Porter, A., Murphy, J., Elliott, S. N., & Cravens, X. (2009). Assessing learning-centered leadership: Connections to research, professional standards, and current practices. *Leadership and Policy in Schools*, *8*(1), 1–36.

Grissom, J., Blissett, R., & Mitani, H. (2017). *Evaluating School Principals: Supervisor Ratings of Principal Practice and Principal Job Performance*. Vanderbilt University Working Paper.

Grissom, J., Kalogrides, D., & Loeb, S. (2015). Using student test scores to measure principal performance. *Educational Evaluation and Policy Analysis*, *37*(1), 3–28.

Grissom, J., & Loeb, S. (2011). Triangulating Principal Effectiveness How Perspectives of Parents, Teachers, and Assistant Principals Identify the Central Importance of Managerial Skills. *American Educational Research Journal*, *48*(5), 1091–1123.

Hallinger, P., & Bridges, E. M. (2017). A Systematic Review of Research on the Use of Problem-Based Learning in the Preparation and Development of School Leaders. *Educational Administration Quarterly*, *53*(2), 255–288.

Hallinger, P., & Murphy, J. (1985). Assessing the instructional management behavior of principals. *The Elementary School Journal*, *86*(2), 217–247.

Horng, E. L., Klasik, D., & Loeb, S. (2010). Principal's time use and school effectiveness. *American Journal of Education*, *116*(4), 491–523.

Jacob, R., Goddard, R., Kim, M., Miller, R., & Goddard, Y. (2015). Exploring the Causal Impact of the McREL Balanced Leadership Program on Leadership, Principal Efficacy, Instructional Climate, Educator Turnover, and Student Achievement. *Educational Evaluation and Policy Analysis*, *37*(3), 314–332.

Jacques, C., Clifford, M., & Hornung, K. (2012). State policies on principal evaluation: Trends in a changing landscape. *Washington DC: National Comprehensive Center for Teacher Quality*.

Kane, T. J., Mccaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating Measures of Effective Teaching using Random Assignment. In *Research Paper. MET Project. Bill & Melinda Gates Foundation*.

Konstantopoulos, S. (2014). Teacher Effects, Value-Added Models, and Accountability. *Teachers College Record*, *116*(1).

Kraft, M. A., Marinell, W. H., & Yee, D. S.-W. (2016). School Organizational Contexts, Teacher Turnover, and Student Achievement: Evidence From Panel Data. *American Educational Research Journal*, *53*(5), 1411–1449.

Ladd, H. F. (2011). Teachers' Perceptions of Their Working Conditions How Predictive of Planned and Actual Teacher Movement? *Educational Evaluation and Policy Analysis*, *33*(2), 235–261.

Leithwood, K., Seashore Louis, K., Anderson, S., & Wahlstrom, K. (2004). *How leadership influences student learning* (pp. 1–87). New York, NY: The Wallace Foundation.

Loeb, S., Kalogrides, D., & Béteille, T. (2012). Effective Schools: Teacher Hiring, Assignment, Development, and Retention. *Education Finance and Policy*, *7*(3), 269–304.

McCullough, M., Lipscomb, S., Chiang, H., & Gill, B. (2016). Do Principals' Professional Practice Ratings Reflect Their Contributions to Student Achievement? Evidence from

Pennsylvania's Framework for Leadership. Working Paper 46. *Mathematica Policy Research, Inc.*

National Policy Board for Educational Administration. (2015). *Professional Standards for Educational Leaders 2015*. Reston, VA.

Porter, A. C., Polikoff, M. S., Goldring, E. B., Murphy, J., Elliott, S. N., & May, H. (2010). Investigating the Validity and Reliability of the Vanderbilt Assessment of Leadership in Education. *The Elementary School Journal*, *111*(2), 282–313.

Robinson, V. M., 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.

Stufflebeam, D., & Nevo, D. (1991). Principal evaluation: New directions for improvement. *Peabody Journal of Education*, *68*(2), 24–46.

Superville, D. R. (2014, May 21). States Forge Ahead on Principal Evaluation - Education Week. *Education Week*. Retrieved from https://www.edweek.org/ew/articles/2014/05/21/32principals_ep.h33.html

Teh, B., Chiang, H., Lipscomb, S., & Gill, B. (2014). Measuring School Leaders' Effectiveness: An Interim Report from a Multiyear Pilot of Pennsylvania's Framework for Leadership. REL 2015-058. *Regional Educational Laboratory Mid-Atlantic*.

Tennessee Department of Education. (2017). *Tennessee Educator Acceleration Model (TEAM) Administrator Evaluation Evaluator Handbook*. Retrieved from http://team-tn.org/wp-content/uploads/2013/08/TEAM-Administrator-Evaluator-Handbook-2017-18.pdf

US Department of Education. (2009). Race to the Top Fund. *Federal Register*, *74*(221), 59688–59872.

Waters, T., Marzano, R. J., & McNulty, B. (2003). *Balanced Leadership: What 30 Years of Research Tells Us about the Effect of Leadership on Student Achievement.* Aurora, CO: Mid-Continent Research for Education and Learning.