Research-to-Practice Partnerships to Strengthen Research-Based Principal Evaluation Systems

Integrative Statement

In 2011-12, many states revised their laws and regulations to support new educator evaluation systems as part of Race to the Top and ESEA flexibility. Principal evaluation systems were being designed with very little research evidence to guide them. States therefore designed principal evaluation systems using many elements of teacher evaluation systems, which had comparatively more research to guide development. To inform future modifications to the principal evaluation system, New Jersey and Pennsylvania asked the Regional Educational Laboratories (REL) - MidAtlantic to partner in studying the measures to assess validity and reliability of the evaluation systems.

Two papers in this symposium will present the findings from the principal evaluation studies in New Jersey and Pennsylvania. In Pennsylvania, the study focused on the reliability and validity of a principal practice measure, the Framework for Leadership (FFL), developed by the state and used in pilot districts (600 schools). In New Jersey, the study focused on the validity and reliability of components of the evaluation system, including student achievement growth and principal practice measures selected by districts. Information from these studies will inform future efforts to improve the principal evaluation systems and the guidance districts receive.

The third paper will present the perspective of a state education agency staff member who partnered in the REL study since its inception and used the findings to inform policy. The REL partnerships entailed sharing principal evaluation data and student data, conducting rigorous analysis of the reliability and validity of component measures of the principal evaluation systems taking into account limitations of the available administrative data, and communicating the findings and policy implications through reports and presentations that were understandable to a variety of education stakeholders in the states. The tension between rigor and relevance in education research partnerships through the REL is most evident in the lengthy process for peer review and publication of REL research papers, which constrains the use of emerging findings in policy discussions. Ideas for improving rigor and relevance for REL studies will be discussed.

A discussant who is conducting research on research-to-practice partnerships will comment on the partnership approach and the potential for strengthening state-district-researcher partnerships through the REL. As the new REL contracts are starting in January 2017, this session will be particularly timely.
Pilot Study of Pennsylvania’s Framework for Leadership: Abstract

Authors: Moira McCullough, Stephen Lipscomb, Hanley Chiang, and Brian Gill

Background / Context

Across the country, encouraged by federal policy, states are developing and implementing new systems for evaluating school principals. They face a substantial challenge: there is scant evidence on the reliability and validity of current principal evaluation tools (Goldring et al., 2009). Notably, although improving student outcomes is a central task of school leaders, no evaluation tool has been shown to indicate principals’ contributions to student achievement. To inform the selection or development of valid and reliable principal evaluation tools, states and districts need more information on ways to accurately measure principal performance.

Purpose / Objective / Research Questions

Our study was conducted in partnership with the Pennsylvania Department of Education to examine the statistical characteristics of Pennsylvania’s newly developed Framework for Leadership (FFL) rubric, including score variation, internal consistency, and score stability. Most importantly, this is one of the first studies to examine the relationship between a measure of principals’ professional practice and their contributions to student achievement growth. Study findings will help inform implementation and refinement of the FFL in Pennsylvania, and will provide valuable information for educators, policymakers, and researchers across the country by helping to fill a considerable information gap on principal evaluation tools.

The FFL specifies 20 leadership practices, known as components, on which principals are rated by their supervisors. FFL components are grouped into four domains: strategic/cultural leadership, systems leadership, leadership for learning, and professional and community leadership. Ratings are based on direct observation and on evidence submitted by principals.

Setting / Population / Participants / Data Collection / Analysis

The study examines the FFL using data for more than 600 principals across Pennsylvania who were rated during two pilot years prior to statewide implementation. Specifically, the pilot data are used to provide information on four key FFL properties:

- First, we describe the distributions of scores on each component, each domain, and the overall FFL rating, thereby assessing whether the FFL distinguishes high and low performers.
- Second, we examine the internal consistency of the FFL by calculating Cronbach’s α for the full FFL and for each domain. The leadership qualities captured by different parts of the FFL should be related in order to reflect overall leadership ability.
- Third, we analyze score stability for participants in both pilot years by calculating the strength of linear association between scores in each year, providing evidence about whether the FFL is a reliable measure of performance.
- Finally, we assess the concurrent validity of the FFL by examining its correlation to principals’ contributions to student achievement growth—that is, their value added. Each principal’s value added is estimated by examining how much the school’s value added deviates from its predicted school value added based on to its performance under the preceding principal.
Findings / Results / Conclusions

The study finds that the FFL is a reliable measure, with good internal consistency and a moderate level of year-to-year stability in scores. The study also finds evidence of the FFL’s concurrent validity: principals with higher scores on the FFL, on average, make larger estimated contributions to student achievement growth. Higher total FFL scores and scores in two of the four FFL domains are significantly or marginally significantly associated with both value added in all subjects combined and value added in math specifically. This evidence of the validity of the FFL sets it apart from other principal evaluation tools—no other measures of principals’ professional practice have been shown to be related to principals’ effects on student achievement growth. However, in both pilot years, variation in scores was limited, with most school leaders scoring in the upper third of the rating scale. As the FFL is implemented statewide, continued examination of evidence on its statistical properties, especially the variation in scores, is important.
Measures of Principals’ Effectiveness: Findings from New Jersey’s Principal Evaluation Study

Abstract

Authors: Mariesa Herrmann and Christine Ross

Background / Context

States and districts across the country are implementing new principal evaluation systems that include measures of the quality of principals’ school leadership practices and measures of student achievement growth. Because these evaluation systems will be used for high-stakes decisions, their component measures must fairly and accurately differentiate between effective and ineffective principals. Thus, the measures need to be reliable (consistent across raters and observations) and valid (correctly measure principal performance). But the research base on the reliability and validity of principal evaluation measures is thin compared with research on teacher evaluation measures. A review of principal practice instruments found that only 2 of 65 instruments documented reliability or validity (Goldring et al., 2009), whereas recent studies documented the reliability and validity of 4 widely used teacher practice instruments (Kane & Staiger, 2012; Kane, McCaffrey, Miller, & Staiger, 2013).

Program / Policy

New Jersey piloted a new principal evaluation system in 2012/13 in 14 school districts and implemented it statewide in 2013/14. The principal evaluation system calls for half of a principal’s summative rating to be composed of two measures of practice: a principal practice instrument selected or developed by each school district and an evaluation leadership instrument developed by the New Jersey Department of Education. The other half of the summative rating is composed of measures of student achievement. One measure of student achievement — a rating based on the school’s median student growth percentile — is only available for schools with grades 4–8. The two other measures of student achievement — a rating based on attainment of principal goals for student achievement and the average of teachers’ student growth objective ratings (measuring teachers’ success in achieving their student growth objectives) — are available for all principals. (Some principal evaluation measures used in the pilot year were replaced or revised for the statewide year.)

Purpose / Objective / Research Questions

The study examined several statistical properties of the component measures used to evaluate principals in New Jersey: the variation in ratings across principals, the year-to-year stability, the associations between component ratings and the characteristics of students in the schools, and the correlations among the ratings. Study findings will inform efforts to improve the principal evaluation system and revise the guidance districts receive.

Setting / Population / Participants / Data Collection / Analysis

The data for the study included information collected by the New Jersey Department of Education on principal evaluation ratings, principals’ job assignments, the principal practice instruments selected by districts, school-level student achievement growth (school median student growth percentiles in math and English language arts), and student background characteristics at the school level. Data from the
pilot year (2012/13) included up to 14 school districts (238 schools) and data from the statewide implementation year (2013/14) included up to 2,500 schools.

The distribution of overall ratings and ratings on each component measure were characterized by the percentage of principals rated in different intervals on the 1–4 point scale. The intervals corresponded to the performance categories associated with intervals of the overall rating: ineffective (1–1.84), partially effective (1.85–2.64), effective (2.65–3.49), and highly effective (3.50–4). Analyses also described the stability of principal practice instrument ratings, school median student growth percentiles, and school median student growth percentile ratings across two or three years for principals who were in the same school for those years and for whom the measures were available. Stability was measured using a Pearson correlation coefficient. The study also examined the relationship between principal evaluation ratings and two measures of student disadvantage: the percentages of economically disadvantaged students and English learner students in the school. These relationships were measured using a Pearson correlation coefficient. Finally, the study examined the relationships among the component measure ratings using a Pearson correlation coefficient.

Findings / Results / Conclusions

The paper will present findings drawn from the two-year study, including the pilot year and the statewide implementation year. Key findings include:

- Many commonly-used principal practice instruments need further development to demonstrate reliability and validity. The developers of six principal practice instruments used by most school districts in New Jersey provided partial information about their instruments’ reliability and validity.

- Statewide, nearly all principals received effective or highly effective overall ratings. Variation in the overall ratings was limited, with 99 percent of principals rated as effective or highly effective overall ratings.

- The percentage of principals who received highly effective overall ratings was lower for principals who were evaluated on school median student growth percentiles than for principals who were not evaluated on this measure. When school median student growth percentiles were not available, principal goals factored more heavily into the overall rating, and most principals received higher ratings on principal goals than on school median student growth percentiles.

- Principal practice instrument ratings and school median student growth percentiles had moderate to high year-to-year stability. But school median student growth percentiles changed more across years in smaller schools than in larger ones.

- School median student growth percentiles, which measure student achievement growth during the school year, exhibit year-to-year stability even when the school changes principals. This may reflect persistent school characteristics, suggesting a need to investigate whether other measures could more closely gauge principals’ contributions to student achievement growth.

- Several component measure ratings—school median student growth percentile ratings, teachers’ student growth objective ratings, and principal practice instrument ratings—as well as the overall rating, had low, negative correlations with student socioeconomic disadvantage. This
suggests that these ratings are biased against principals of schools with more disadvantaged students or that less effective principals are serving schools with more disadvantaged students.

- **Principals’ ratings on component measures had low to moderate positive correlations with each other.** This suggests that the components measure distinct dimensions of overall principal performance.
In 2012-13, to better align educator evaluation with practices to improve student outcomes, New Jersey piloted a new principal evaluation system in 14 school districts and after making adjustments, implemented it statewide in 2013/14. Believing that it was important to understand how New Jersey’s implementation of principal evaluation fit into the national picture, the Department requested an independent evaluation of the pilot and statewide principal evaluation systems to inform future decisions about the system and guidance to districts.

The study was conducted through the Regional Educational Laboratories – MidAtlantic. Several findings from the study were key to informing principal evaluation policy, including regulations and guidance to districts:

- Most principals received higher ratings on principal goals than on school median student growth percentiles, and the highest ratings were on the average of teachers’ student growth objective goal attainment. The findings suggested that principals and teachers were setting attainable but not ambitious goals.

- Measures of student outcomes include year-to-year random variation and the amount of year-to-year change depends in part on the measure and the baseline value. Both of these findings need to be considered when setting student outcome goals for evaluating principals and teachers.

The department developed revised guidance on setting goals that includes a requirement for including information on the baseline outcome when goals for the future are set. More importantly, the rigorous research conducted by the REL helped the Department avoid acting on anecdotal evidence when making policy decisions. Particularly as it relates to the goals, many administrators analogously believed in the value of goals that the research suggested are not, on average appropriate. Absent this research, the Department’s guidance, and perhaps its policies, would be less effective.

- The developers of six principal practice instruments used by most school districts in New Jersey provided partial information about their instruments’ reliability and validity.

The department developed guidance for school districts to strengthen their processes for training observers so that principals and teachers are rated reliably. A major part of the guidance was a focus on ensuring that instruments were implemented in a way that made sense for the local context.

Participating in a REL study had several advantages for the Department. A careful study of the component measures of the principal evaluation system yielded important information about the strengths and weaknesses of the measures and ways to improve the system. The finding that principals’ ratings on component measures had low to moderate positive correlations with each other suggested that the components measure distinct dimensions of overall principal performance, providing some confidence in the system as a whole. At the same time, the study suggested some areas of weakness that merit more investigation and policy development, such as areas in which educators develop their goals and the possibility that some measures, such as the practice measure and SGP, are partly influenced by persistent school-specific factors.

The REL partnership also helped the Department to use its data in new ways. At the end of the study, the researchers developed a data mapping program that will allow the department to track and analyze district-level data more effectively, given the large number of districts in the state. Broadly, the analysis provided as part of this partnership showed the Department how to examine its newly collected
evaluation data. Analyses that were developed for use on principal evaluation data have been used for internal analyses in other areas, and some analyses will be continued in the area of principal evaluation.

Nevertheless, the REL partnership had shortcomings also. The report review process is incredibly long, and as a result, the department was not able to make as much of the partnership as we would have liked—for example, presenting the findings to the state board when there were relevant pending regulations would have been very useful, but by the time the paper was cleared for release, the moment had long passed. This diminished the department’s ability to demonstrate the thoughtfulness with which it developed its policy and guidance, and it means that stakeholders throughout the state are unaware of the effective work conducted as part of this partnership.

As the REL begins its next 5-year cycle, a rapid review process for key findings that would enable them to be used in real time by policymakers would be helpful. It would allow the department to demonstrate its commitment to rigorous evaluation projects, and it would allow the public to understand the important work conducted by the REL.