MATHEMATICA
Policy Research

WORKING PAPER 46

Working PAPER

BY MOIRA MCCULLOUGH, STEPHEN LIPSCOMB, HANLEY CHIANG, AND BRIAN GILL

Do Principals' Professional Practice Ratings Reflect Their Contributions to Student Achievement?

Evidence from Pennsylvania's Framework for Leadership

June 2016

ABSTRACT

Although states and school districts have begun to evaluate school principals, there is little evidence on the validity of principal evaluation measures. To fill this gap, we examined Pennsylvania’s Framework for Leadership (FFL), a tool for measuring and evaluating principals’ professional practices. Using data on more than 300 principals, we find that FFL evaluation scores are significantly and positively correlated with estimates of principals’ contributions to student achievement. The strongest relationships are in the domains of (1) systems leadership and (2) professional and community leadership. Contributions to math achievement were more highly correlated than contributions to achievement in other subjects. The results are driven mainly by evaluations of principals who have led their schools for at least three years. This is the first study to find evidence that ratings of principals’ professional practice are correlated with credible measures of principals’ contributions to student achievement.

The authors are grateful to staff at the Pennsylvania Department of Education for their help throughout this study. The authors also appreciate thoughtful input on early drafts of the paper from Matt Johnson and programming assistance from Clare Wolfendale, Swaati Bangalore, and Irina Cheban. They also thank Cindy George and Alfreda Holmes for editorial and production support, respectively. This project has been funded at least in part with Federal funds from the U.S. Department of Education under contract number ED-IES-12-C-0006. The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

I. INTRODUCTION

States and districts across the country have been developing new measures to evaluate school principals, with the aims of providing richer information on performance, increasing differentiation among principals, and promoting improved practice. Development and implementation of new systems for evaluating principals' professional practices are motivated by state and district policies that focus on raising student achievement and improving struggling schools. Evidence from an emerging literature indicates that principals can have meaningful impacts on student achievement (Branch, Hanushek, & Rivkin, 2012; Chiang, Lipscomb, & Gill, 2016; Coelli & Green, 2012; Dhuey & Smith, 2013, 2014; Grissom, Kalogrides, & Loeb, 2015a). Yet researchers and policymakers have yet to validate a set of leadership practices as being associated with principals who make larger contributions to student achievement. Identifying such practices and aligning principal evaluations with them would enable states and districts to reward effective principals, support improvements in performance, and help students succeed.

A substantial challenge that states and districts face in revising systems for evaluating principals is that scant evidence exists on the accuracy of current principal evaluation tools. Nearly all tools lack any type of validity information (Condon & Clifford, 2012; Goldring et al., 2009). In particular, no evaluation tool has been shown to indicate principals' contributions to student achievement, even though improving student outcomes is a central task of school leaders. This lack of evidence for the validity of principal evaluation tools contrasts with several frameworks for rating teachers' professional practices, whose ratings have been shown to relate positively with teachers' contributions to student achievement (Kane & Staiger, 2012; Chaplin, Gill, Thompkins, & Miller, 2014). To inform the selection or development of valid principal evaluation tools, states and districts need more information on how to measure the quality of principals' leadership practices accurately.

This study makes three contributions to the research literature on principals' effectiveness at raising student outcomes. First, it provides evidence of a positive relationship between principals' ratings on an evaluation tool and their contributions to student achievement. Second, it provides this evidence in the context of a sample that has external validity, including principals from nearly 200 of the 500 school districts in Pennsylvania. Third, it describes an approach for estimating principals' contributions to student outcomes that imposes fewer restrictions than approaches used previously on the ability to compare effectiveness across a broad set of school principals.

We examined a measure called the Framework for Leadership (FFL) that was developed by the Pennsylvania Department of Education (PDE) as part of a mandated revision of the principal evaluation process under state legislation passed in 2012 (Act 82).[1] The FFL includes 20 leadership practices grouped into four domains that PDE considers important for school leaders to exhibit to raise student outcomes. We used data from principals who, during the 2013/14 school year, participated along with their supervisor in a pilot implementation of the FFL. The

[1] The FFL is available online on the PDE website (Pennsylvania Department of Education, 2014): https://static.pdesas.org/content/documents/Principal_Rubric.pdf.

pilot was a no-stakes trial of the evaluation tool, but in other respects it mimicked real-world conditions.

The study findings reveal small but statistically significant positive relationships between principals' FFL scores and estimates of their contributions to student achievement, particularly in the domains of systems leadership and professional and community leadership, and on math outcomes. Specifically, principals whose contributions to math achievement were one standard deviation above average tended to have scores on the FFL as a whole and in those two domains that were about 0.17 standard deviations above average, placing them at approximately the 56th percentile of professional practice scores. FFL scores in those two domains were also positively related to principals' combined contributions to student achievement across academic subjects. These relationships appear to be driven more by evaluations of principals with at least three years of experience leading their school than by those of principals with only one or two years of experience.

The correlations between FFL scores and principals' contributions to student achievement were relatively small. Correlations for teacher evaluations often fall between 0.20 and 0.30 (Kane & Staiger, 2012; Lipscomb, Terziev, & Chaplin, 2015), whereas the correlations for the FFL scores fell at the low end of this range. Nevertheless, findings from the study constitute the first positive test of the concurrent validity of a principal evaluation tool—that is, its relationship with another measure of the same concepts—for contributions to student achievement outcomes.

To assess whether the FFL distinguished principals with larger and smaller contributions to student achievement, we needed to estimate the size of those contributions. Several studies have obtained such estimates using value-added models, which control for students' prior achievement and characteristics when comparing student achievement across schools or school leaders (Branch et al., 2012; Chiang et al., 2016; Coelli & Green, 2012; Dhuey & Smith, 2013, 2014; Grissom, Kalogrides, & Loeb, 2015a; Teh, Chiang, Lipscomb, & Gill, 2014). A key observation from this research is that a principal's value-added is not the same as his or her school's value-added, because the school's value-added may also reflect school-specific factors beyond the principal's control (Chiang et al., 2016).

Most of the previous literature on principal value-added has distinguished it from the influence of other school factors by comparing the same school's performance under different principals. The more effective principal is the one under whom the school fared better. This method can be applied only to schools with principal turnover during the period considered, a current limitation that will become less important as education data systems continue to develop and accumulate information across additional years. However, another limitation has made this method unsuitable for use in either this study or a large-scale evaluation system—in most cases, one can compare estimates only among principals who have served the same school.

In contrast, we estimated principal value-added using an approach that allowed us to compare estimates across a broad set of principals. The key component of our approach was the adjustment of the value-added of a principal's school for its value-added in the year before the principal began leading the school. Thus, we estimated principal value-added based on changes in school effectiveness that took place under a successor principal relative to the baseline established in the predecessor's final year. This method controls for school-specific influences such as lingering effects of the predecessor, while still yielding estimates that can be compared across principals leading different schools. We used these estimates as a benchmark to assess the FFL's concurrent validity.[2]

Four prior studies have examined the relationship between principal evaluation tools and the value-added of either schools or principals. In a study of two medium-size districts, principals' scores generally did not correlate with school value-added in reading and math, although in math the correlations were statistically significant in a minority of the analysis samples considered (Milanowski & Kimball, 2012). This study did not examine the relationship with the principals' own value-added. In Miami-Dade County Public Schools, principals' scores were positively associated with the value-added of their schools but not with the value-added of the principals themselves (Grissom et al., 2015a). A study of principals in Tennessee found that their ratings on the state professional practice rating tool were positively related to school value-added and to survey ratings of their performance by both teachers and assistant principals, but they were not related to teacher turnover rates (Grissom, Blisset, and Mitani, 2015b). Finally, a study using data from an earlier year of the same Pennsylvania pilot of the FFL examined in this study found no relationship with principal value-added (Teh, et al., 2014).

The rest of the relevant research literature has assessed the validity of principal evaluation tools through approaches other than examining relationships with value-added. For example, several studies have examined properties of the Vanderbilt Assessment of Leadership in Education (VAL-ED) (Porter et al., 2008). An assessment of its convergent validity—whether different measurement methods using the same tool produced similar scores—found that ratings of the same principal by different stakeholders (teachers, supervisors, and the principals themselves) had positive correlations in the range of 0.13 to 0.27 (Porter et al., 2010). An analysis of the concurrent validity of teachers' ratings of their principals as measured against a different rating tool, the Principal Instructional Management Rating Scale, found a positive correlation of 0.7 (Goldring, Cravens, Murphy, Porter, & Elliot, 2012). Although the findings on the VAL-ED's properties are promising, so far the extent to which this evaluation tool captures principals' value-added remains unassessed.

II. THE FRAMEWORK FOR LEADERSHIP

The FFL is an evaluation tool that PDE developed to measure the quality of school leaders' practices. It specifies 20 leadership practices, known as components, on which supervisors such as district superintendents rate school leaders. On each component, a school leader can receive a rating of distinguished (3 points), proficient (2 points), needs improvement (1 point), or failing (0 points). Ratings are based both on direct observation and on evidence submitted by school leaders. Supervisors are encouraged to rate school leaders on all components for which evidence exists to support a rating. The FFL is designed for use with both principals and assistant principals, although in the present study we focused on principal evaluations only.

[2] Similar to the method for estimating principal value-added based on school leadership transitions used in previous research, our approach depends on the availability of data for the year before principals began leading their current schools. In this study, we could estimate value-added for principals who had at most six years of tenure at their schools. This limitation will become less significant as data systems compile more years of records.

FFL components are grouped into the following four domains: strategic/cultural leadership (domain 1), systems leadership (domain 2), leadership for learning (domain 3), and professional and community leadership (domain 4). Table 1 describes the domains and lists their specific components.

Table 1. Framework for Leadership domains and components

Domain 1. Strategic/Cultural Leadership: *Principals/school leaders systemically and collaboratively develop a positive culture to promote student growth and staff development. They articulate and model a clear vision of the school's culture that involves students, families, and staff.* **Components** 1a. Creates an organizational vision 1b. Uses data for informed decisionmaking 1c. Builds a collaborative and empowering work environment 1d. Leads change efforts for continuous improvement 1e. Celebrates accomplishments and acknowledges failures
Domain 2. Systems Leadership: *Principals/school leaders ensure that there are processes and systems in place for budgeting, staffing, problem solving, communicating expectations and scheduling that result in organizing the work routines in the building. They must manage efficiently, effectively and safely to foster student achievement.* **Components** 2a. Leverages human and financial resources 2b. Ensures school safety 2c. Complies with federal, state, and local education agency mandates 2d. Establishes and implements expectations for students and staff 2e. Communicates effectively and strategically 2f. Manages conflict constructively 2g. Ensures a high-quality, high-performing staff
Domain 3. Leadership for Learning: *Principals/school leaders ensure that a Standards Aligned System is in place to address the linkage of curriculum, instruction, assessment, data on student learning and teacher effectiveness based on research and best practices.* **Components** 3a. Leads school improvement initiatives 3b. Aligns curricula, instruction, and assessments 3c. Implements high-quality instruction 3d. Sets high expectations for all students 3e. Maximizes instructional time
Domain 4. Professional and Community Leadership: *Principals/school leaders promote the success of all students, the positive interactions among building stakeholders and the professional growth of staff by acting with integrity, fairness and in an ethical manner.* **Components** 4a. Maximizes professional responsibilities through parent involvement and community engagement 4b. Shows professionalism 4c. Supports professional growth

Supervisors are instructed to judge the preponderance of evidence from the components in each domain and assign a summary score, known as a domain score, using the same rating scale as for the component scores (3, 2, 1, or 0 points). Because our data include only component scores (see section 4), we calculated the domain scores as the equal-weighted average of scores from the components on which the leader was evaluated in that domain. PDE regards the four domains as equally weighted elements of a school leader's annual evaluation rating, so the study defined a school leader's full FFL score as the equal-weighted average of the four domain scores.

PDE designed the FFL to have a structure consistent with the Framework for Teaching (FFT), a well-known classroom observation tool that many states and school districts—including those in Pennsylvania—use to measure teachers' professional practices. In particular, PDE considers the FFL and the FFT to be aligned in eight areas: vision, common standards, high expectations for all, instruction, assessment, collaboration, safety and security, and professionalism.[3] However, the FFL was developed independently from the FFT and is not associated with its developer, the Danielson Group.

III. DATA SOURCES AND DESCRIPTIVE STATISTICS

PDE piloted the FFL in nearly 200 of the state's 500 districts in 2013/14, before introducing it statewide during the 2014/15 school year. The 2013/14 pilot sought to obtain information on the reliability and validity of FFL scores, and followed a smaller pilot undertaken in 2012/13 with similar aims (Teh et al., 2014).[4] Participants in the 2013/14 pilot included 517 school principals whose local education agency or school participated for one of three reasons: (1) they were receiving Race to the Top funds, (2) they were receiving School Improvement Grants to implement a transformation model of improvement, or (3) they volunteered. Participating principals agreed to be rated by their supervisor—typically a superintendent or assistant superintendent—and to have those scores collected as part of the pilot. Ratings were not used for any formal evaluation.

Although the pilot ratings had no consequences for participating principals, pilot conditions were designed to mimic actual implementation of the FFL statewide. In particular, PDE provided a one-day training on the FFL to supervisors, which included a discussion of concrete examples of the evidence that would merit a proficient score for every FFL component. PDE also instructed supervisors to rate principals on all FFL components for which they had evidence to support a rating, which is how the FFL is intended to be used in practice.

This study relies on FFL scores submitted by local education agencies to the Pennsylvania Training and Technical Assistance Network, an agency within PDE. These data included supervisor-assigned principal performance ratings at the component level. As described in section 2, we computed a principal's domain score as the equal-weighted average of scores from the components in the domain on which the principal was evaluated and the full FFL score as the equal-weighted average of the four domain scores. Our analyses use full FFL scores and the four domain scores because domain scores reflect the rater's overall assessments of a principal's performance in the four key areas of leadership quality identified by PDE, and the full FFL score is the rating ultimately intended to be incorporated as 50 percent of a principal's overall evaluation.

[3] See http://www.education.pa.gov/Documents/Teachers-Administrators/Educator%20Effectiveness/Principals%20and%20CTC%20Directors/Principal%20Effectiveness%20Program%20Brochure.pdf.

[4] This analysis focused exclusively on the 2013/14 pilot. The sample of participating principals was substantially larger in the 2013/14 pilot, and training sessions for raters were adapted to reflect feedback from the 2012/13 pilot year, including discussing concrete examples of the evidence that would merit a proficient score for each FFL component.

We used data on student achievement scores and background characteristics, and on principals' job assignments, to estimate school leaders' contributions to student achievement growth (see section 4). These administrative records came from databases maintained by agencies at PDE. The student achievement data included statewide assessment scores for all students in the state who were administered assessments from 2006/07 to 2013/14. During this period, Pennsylvania administered end-of-grade assessments from the Pennsylvania System of School Assessment (PSSA) in the following subjects and grades: reading and math in grades 3–8 and grade 11; science in grades 4, 8, and 11; and writing in grades 5, 8, and 11. It also administered modified PSSA tests to students with disabilities who were eligible for those assessments and end-of-course assessments, called the Keystone Exams, which were given statewide starting in 2012/13 in algebra I, biology, and literature.

Additional administrative records on students and school leaders came from the state's longitudinal data system. These data covered all students who were enrolled in the state's public schools and all principals who worked in those schools at any time from 2007/08 to 2013/14. For each student in each year, the data indicated the schools in which the student was enrolled and information on the student's gender, age, race/ethnicity, free and reduced-price lunch status, English learner status, and disability status. Data on principals indicated which schools they led and provided information on their gender, education degrees, race/ethnicity, and total work experience in prekindergarten through grade 12 (PK-12) education.

This study focuses on the 305 of the 517 principals in the pilot sample for whom the available administrative data allow us to estimate their school's value-added in the year before they began at their schools, which is a key component of our approach to estimating principals' value-added. Students in schools participating in the pilot were lower achieving at baseline relative to all students in the state, which may be expected given that many local education agencies and schools were receiving Race to the Top funds or School Improvement Grants (Table 2). Relative to the statewide average, students in schools led by principals in the pilot also tended to have more disadvantaged background characteristics. For example, they were significantly more likely to be eligible for free or reduced-price lunch or to change schools during the year and significantly less likely to be white, although these differences were not substantively large.

Table 2. Characteristics of students in 2013/14 from all Pennsylvania schools and from schools led by principals in the analysis sample

	Grades 4 to 5		Grades 6 to 8		Grades 9 to 12	
Student characteristic (percentages unless otherwise noted)	Statewide	Analysis sample	Statewide	Analysis sample	Statewide	Analysis sample
Baseline math score (average z-score)a	0.02	-0.12	0.02	-0.06	-0.05	-0.14
Baseline reading score (average z-score)a	0.02	-0.11	0.02	-0.05	-0.04	-0.12
Receives free lunch	40.2	46.2	37.7	40.1	34.4	35.8
Receives reduced-price lunch	5.0	5.7	5.2	5.8	5.4	6.3
English language learner	2.3	2.0	2.1	1.5	1.7	1.6
Has a disability	18.4	19.2	17.3	18.1	16.6	16.0
Moved schools during school year	3.5	4.0	3.7	4.6	11.7	12.5
Grade repeater	0.2	0.3	0.7	0.9	4.2	6.0
Overage for grade	0.2	0.2	0.3	0.4	0.9	1.1
Age (average years)	10.1	10.1	12.6	12.7	15.6	15.6
Female	49.2	49.2	49.0	48.5	49.3	50.4
Race and ethnicity						
Asian or Pacific Islander	3.7	2.5	3.4	2.3	2.8	1.8
Black, non-Hispanic	14.3	19.1	14.1	16.0	14.1	13.3
Hispanic	9.3	10.2	8.3	7.8	7.3	8.9
White, non-Hispanic	69.3	65.0	70.6	69.7	71.2	70.0
Other race/ethnicity	2.7	2.5	1.3	1.6	0.6	0.6
Number of students	247,286	27,068	375,847	46,849	282,629	22,553

Bold indicates statistical significance compared with the statewide mean at p = .05.

Notes: The analysis sample column summarizes the characteristics of students in the schools led by the 305 principals included in the regressions reported in Table 6.

[a] For students in grades 4–8, baseline scores come from the previous year. For students in grades 9–12, baseline scores come from grade 8.

Source: Authors' calculations based on student achievement and background data provided by the Pennsylvania Department of Education.

Despite differences in the characteristics of their students, principals in our analysis sample and principals statewide had similar observable characteristics (Table 3). Relative to the state mean, there were no statistically significant differences in terms of degree attainment, race and ethnicity, and gender. One exception is that principals in the pilot sample had less total experience in PK-12 education than the average principal in the state (16.2 versus 18.2 years).[5] Moreover, using the method described more fully in the next section, the average value-added of pilot participants was statistically indistinguishable from the average value-added of all school leaders statewide for whom value-added could be estimated using the available data, whether measured in each subject area or combined across subjects (Table 4).[6]

Table 3. Characteristics of principals in 2013/14 from all Pennsylvania schools and from schools led by principals in the analysis sample

Principal characteristic (percentages unless otherwise noted)	Statewide	Analysis sample
Total experience in PK-12 education (average years)	18.2	16.2
Highest degree attained		
Bachelor’s	14.7	14.2
Master’s	75.7	77.4
Doctorate	9.6	8.4
Race and ethnicity		
Black, non-Hispanic	10.3	11.6
White, non-Hispanic	87.7	87.5
Other race/ethnicity	2.0	1.0
Gender		
Female	44.5	40.9
Male	55.5	59.1

PK–12 is prekindergarten to grade 12.

Bold indicates statistical significance compared with the statewide mean at p = .05.

Notes: The analysis sample column summarizes the characteristics of the 305 principals included in the regressions reported in Table 6. Percentages may not sum to 100 because of rounding.

Source: Authors’ calculations based on job assignment and background data on school leaders provided by the Pennsylvania Department of Education.

[5] The findings about how the characteristics of principals and their students compare to statewide averages are the same using the full sample of 517 principals in the pilot (Authors, Year).

[6] We standardized the value-added estimates of all school leaders statewide to have a mean of zero and an error-adjusted standard deviation of one. As a result, the extent to which the average value-added of pilot participants differed from zero indicated how dissimilar they were relative to all leaders statewide in their contributions to student achievement.

Table 4. Mean and standard deviation of the value-added estimates for principals in the analysis sample relative to the statewide distribution of principals' value-added estimates

Outcome subject	Mean relative to statewide average (principal standard deviations)	Error-adjusted standard deviation (principal standard deviations)
All combined	0.05	0.94
Math	-0.01	0.96
English language arts	0.10	1.08
Science	0.08	0.94

Bold indicates statistical significance of mean differences in principal value-added between the analysis sample and statewide at p = .05.

Note: The sample size for this table includes 305 principals.

Source: Authors' calculations based on student achievement and background data and school leaders' job assignment data provided by the Pennsylvania Department of Education.

Principals tended to score well on the FFL during the 2013/14 pilot. As shown in Table 5, most principals in our analysis sample received component ratings of proficient or distinguished. The most common performance rating category was proficient (ranging from 59 percent to 81 percent of principals across components). On all but one component (3b—aligns curricula and instruction), less than 10 percent of principals received the needs improvement rating, and on every component, less than one percent of principals received the failing rating. Consistent with the prevalence of high component score ratings, school leaders were overwhelmingly likely to receive domain scores of 2.0 or above on the 0–3 point scale. In every domain, 95 percent or more of principals scored at least 2.0. Full FFL scores likewise fell within the top third of the rating scale. As shown in McCullough et al. (2016), the distributions of FFL scores using the full sample of principals in the 2013/14 pilot were very similar.

IV. METHODS

Does the FFL measure the leadership qualities and practices that enable principals to improve student achievement? To address this question, we examined the extent to which principals with higher value-added—those who made larger contributions to student achievement—earned higher FFL scores than principals with lower value-added. The two key steps in the analysis, described in this section, were to (1) estimate the value-added of each principal and (2) estimate the relationship between principals' value-added and their FFL scores.

4a. Estimating principals' value-added

To estimate principals' value-added, we began by estimating the value-added of the schools they led. Each school's value-added bundled together the principal's contribution with the influence of school-specific factors beyond the principal's control, such as neighborhood quality and teacher personnel decisions made by previous principals. Therefore, we subsequently adjusted the school value-added estimates to distinguish principals' contributions from these other school-specific factors.

Estimating school value-added

To obtain estimates of schools' value-added, we used a regression model similar to school value-added models used in prior work (see, for example, Chiang et al., 2016; Deutsch 2014; Lipscomb, Chiang, & Gill, 2012; Rotz, Johnson, & Gill, 2014). We estimated the following regression model for the test score, A_{istg}, of student i in school s during year t in grade g:

$$(1)\quad A_{istg} = \boldsymbol{A}_{i,t-1}\boldsymbol{\beta} + \boldsymbol{X}_{\boldsymbol{istg}}\boldsymbol{\delta} + \alpha_{stg} + \varepsilon_{istg},$$

where $\boldsymbol{A}_{i,t-1}$ was a vector of the student's prior-year test scores in math, reading, and, when available, science and writing; $\boldsymbol{X}_{\boldsymbol{istg}}$ was a vector of student background characteristics; α_{stg} was a school-by-year-by-grade fixed effect; and ε_{istg} was a random error term. We estimated equation (1) separately by year, grade, and subject for all school years from 2007/08 through 2013/14. For high school outcomes, we took baseline test scores from eighth grade, the most recent prior year in which students were assessed.

Table 5. Summary statistics on the distribution of Framework for Leadership component, domain, and full FFL scores for principals in the analysis sample, 2013/14 school year

	Principal performance category (percentages)					
Component or Domain	Failing	Needs Improvement	Proficient	Distinguished	Mean	Std. dev.
1a (Strategic goals)	0.8	5.3	74.3	19.6	2.1	0.5
1b (Data for decision making)	0.0	7.6	69.8	22.7	2.2	0.5
1c (Empowering work environment)	0.4	7.2	65.6	26.9	2.2	0.6
1d (Continuous improvement)	0.0	4.9	68.2	26.9	2.2	0.5
1e (Lessons from accomplishments and failures)	0.0	2.1	71.2	26.7	2.2	0.5
2a (Leverages resources)	0.4	5.1	79.1	15.4	2.1	0.5
2b (School safety)	0.4	1.9	73.5	24.3	2.2	0.5
2c (Complies with mandates)	0.4	3.9	81.3	14.3	2.1	0.4
2d (Clear expectations for students and staff)	0.0	3.0	76.8	20.2	2.2	0.5
2e (Communicates effectively)	0.4	5.1	71.0	23.5	2.2	0.5
2f (Manages conflict)	0.4	8.0	71.9	19.7	2.1	0.5
3a (School improvement initiatives)	0.4	4.9	77.1	17.6	2.1	0.5
3b (Aligns curricula and instruction)	0.8	10.1	69.6	19.4	2.1	0.6
3c (High quality instruction)	0.7	7.6	73.6	18.1	2.1	0.5
3d (High expectations for students)	0.4	4.7	76.6	18.3	2.1	0.5
3e (Maximizes instructional time)	0.4	0.8	72.2	26.6	2.2	0.5
4a (Parent and community involvement)	0.4	7.4	64.9	27.3	2.2	0.6
4b (Professionalism)	0.8	2.4	58.9	37.9	2.3	0.6

	Principal performance category (percentages)					
Component or Domain	Failing	Needs Improvement	Proficient	Distinguished	Mean	Std. dev.
4c (Supports professional growth)	0.7	2.7	71.5	25.2	2.2	0.5
All components	0.4	5.0	71.8	22.7	2.2	0.5
Domain 1 (Strategic/Cultural Leadership)	0.0	2.6	76.4	21.0	2.2	0.4
Domain 2 (Systems Leadership)	0.3	1.3	81.6	15.1	2.1	0.3
Domain 3 (Leadership for Learning)	0.7	3.3	82.2	13.8	2.1	0.4
Domain 4 (Professional and Community Leadership)	0.0	2.0	61.3	30.1	2.2	0.4
Full FFL	0.0	1.6	83.0	15.4	2.2	0.3

Notes: FFL=Framework for Leadership. The sample size for this table includes 305 principals. Percentages may not sum to 100 across performance categories because of rounding.

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14 provided by the Pennsylvania Department of Education.

When estimating equation (1), we accounted for measurement error in the pretest scores and the presence of students in multiple schools (see McCullough et al. 2016 for details). The background characteristics ($\boldsymbol{X_{istg}}$) included the student's age in years and a vector of dummy variables for free lunch recipients, reduced-price lunch recipients, English language learners, students with particular disabilities, students who moved during the year, grade repeaters, students who were overage for their grade, students who took a modified test, males, and students of various racial/ethnic groups.

The estimated fixed effect, $\hat{\alpha}_{stg}$, was the estimated value-added of school s in year t for grade g. Within each subject, year, and grade, we standardized the school value-added estimates for schools across the state to have mean zero and standard deviation one. Because the objective was to measure a school's contribution to student achievement across all grades, we averaged each school's value-added estimates in each subject across the grades it served, weighting grades by the number of students. We also averaged value-added estimates in reading and writing together, again weighting by the number of students contributing to each estimate. For each school and year, this produced three subject-specific value-added estimates—in math ($\hat{\alpha}_{st}^{math}$), English language arts ($\hat{\alpha}_{st}^{ela}$), and science ($\hat{\alpha}_{st}^{sci}$). In some analyses, we used a school-by-year value-added estimate that averaged across the three subjects ($\hat{\alpha}_{st}^{all}$), weighting subjects by the number of students.

Adjusting school value-added estimates to isolate principals' contributions

As discussed earlier, the school value-added estimates were imperfect measures of principals' effectiveness because they also reflected school-specific influences on student achievement that were beyond principals' control (Chiang et al., 2016). These school-specific influences include persistent school characteristics and decisions made by previous principals that are hard to alter in the short run. For example, schools in safer neighborhoods might consistently attract more effective teachers who generate higher school value-added. Similarly, a previous principal might have hired less effective teachers whom the current principal cannot easily dismiss.

Accounting for school-specific influences is therefore important for avoiding bias in estimating principals' value-added. Prior research has sought to account for persistent school-specific factors by comparing the value-added of the same school under different principals (Branch et al., 2012; Chiang et al., 2016; Coelli & Green, 2012; Dhuey & Smith, 2013, 2014; Grissom et al., 2015a). This method is intended only to estimate a principal's effectiveness relative to that of his or her predecessor or successor at the same school (or, in some cases, relative to the average principal in a small network of schools connected by principal transfers). Yet for real-world evaluations as well as for the purposes of this study, comparing only principals who have served at the same school was too restrictive. FFL scores—with which the value-added estimates would be compared—came from a single year, so we typically did not have FFL scores for both a predecessor and a successor at the same school. Even if we did, comparisons only within the same school would still have had limited policy relevance. Instead, we would like to know whether principals who tend to have higher value-added relative to others in the state also tend to earn higher FFL scores.

To account for school-specific factors in a way that still permits comparisons of value-added across a broad population of principals, we assumed a parametric form for the effects of these school factors. A school's value-added before the arrival of the current principal—referred to as the school's baseline value-added—was assumed to encapsulate school-specific factors beyond the current principal's control. The relationship between schools' baseline and current value-added captured the degree to which these school-specific factors persisted over time to influence the school's current value-added. Given this degree of persistence, we predicted each school's current value-added based on its baseline value-added. The difference between this prediction and the school's actual value-added represented the principal's contribution to student achievement.

To formalize this method, in each subject k, let $\hat{V}^{k}_{p(s),t} \equiv \hat{\alpha}^{k}_{st}$ be the estimated value-added of the school s led by principal p in year t, obtained using the steps described earlier. Let $\hat{B}^{k}_{p(s)}$ be the estimated value-added of school s in the "baseline" year—the year immediately before principal p arrived at school s. Separately for each $k \in \{math, ela, sci, all\ subjects\}$, we estimated the regression

$$(2)\quad \hat{V}^{k}_{p(s),t} = \delta_0 + \delta_1\hat{B}^{math}_{p(s)} + \delta_2\hat{B}^{ela}_{p(s)} + \delta_3\hat{B}^{sci}_{p(s)} + \delta_4 high_{p(s)} + \delta_5(high_{p(s)} * \hat{B}^{math}_{p(s)}) + \delta_6(high_{p(s)} * \hat{B}^{ela}_{p(s)}) + \delta_7(high_{p(s)} * \hat{B}^{sci}_{p(s)}) + \theta_t + u_{pt},$$

where $high_{p(s)}$ was an indicator for high schools, θ_t was a year fixed effect, and u_{pt} was an error term.

For each principal and year, the residual from equation (2), $\hat{u}_{pt}$, was the estimate of the principal's value-added in that year. The principal's value-added estimate captured the degree to which her school's value-added in the current year exceeded or fell short of a prediction based on the same school's value-added under the previous principal. It answered the question, "How different is the school's contribution to student achievement this year compared to the contribution it would have had if an average principal had led the school after the previous principal left?" Because principals of different tenures in their school had different amounts of time to shape their schools' effectiveness, we standardized $\hat{u}_{pt}$ to have mean zero and standard deviation one separately for principals with different tenure lengths in their school.

The validity of these principal value-added estimates hinges on two assumptions. First, because each principal is effectively being compared to other principals whose schools had similar baseline value-added, a key assumption is that true principal value-added is uncorrelated with schools' baseline value-added. In contrast, if effective principals were, for example, disproportionately assigned to schools with low baseline value-added, then the value-added of these principals would be estimated relative to an unfair comparison group—other effective principals—rather than relative to the average principal in the state. Second, time-varying influences on student achievement beyond principals' control—which the schools' baseline value-added does not capture—must be uncorrelated with principals' true value-added.

Several features of equation (2) were intended to minimize bias in the principal value-added estimates. First, we included interaction terms between baseline value-added and the high school indicator to account for the possibility that the lingering effects of previous school leaders may have been stronger in high schools than in other schools. For high schools, school value-added estimates controlled for students' grade 8 (not prior-year) scores, so the current value-added of high schools could have reflected, in part, growth that students experienced under the current principals' predecessors if the current principals began their positions after the students had already completed one or more years of high school. Second, we accounted for measurement error in the schools' baseline value-added estimates. Because those covariates were estimates, estimation error in those covariates could bias the estimated coefficients on those variables toward zero unless addressed. To account for measurement error, each baseline value-added variable was adjusted by an empirical Bayes "shrinkage" procedure before being used in equation (2), so that the regression coefficient on the adjusted variable would no longer be attenuated (Morris, 1983; see McCullough et al. 2016 for details on this adjustment).

The estimation sample for equation (2) included all principals in Pennsylvania with valid estimates of their schools' current and baseline value-added. To increase the precision of the estimated coefficients, the regressions pooled together principal-year observations from all available years of principal value-added estimates (2008/09–2013/14). We then extracted the value-added estimates from 2013/14 ($\hat{u}_{p,2014}$) for the principals in the FFL pilot to assess relationships with the principals' 2013/14 FFL scores, as described next.

4b. Estimating the relationship between principals' value-added and FFL scores

After generating estimates of principals' value-added, we then assessed the extent to which principals with higher value-added earned higher FFL scores. Our basic approach was to estimate a regression model in which FFL scores were the dependent variable and principals' value-added estimates were the key independent variable. Because the value-added estimates, $\hat{u}_{p,2014}$, had estimation error, we again applied empirical Bayes "shrinkage" to adjust for this error so that the estimated coefficient on this variable in the regression model would not be biased toward zero. The final regression model had the following form:

$$(3)\quad FFL_{p,2014} = \lambda_0 + \lambda_1 \tilde{u}_{p,2014} + \sum_{k=2}^{6} \gamma_k T_{p,2014}^{(k)} + v_{p,2014},$$

where $FFL_{p,2014}$ was the FFL score of principal p in 2013/14, $\tilde{u}_{p,2014}$ was the empirical Bayes estimate of the principal's value-added in 2013/14, $T_{p,2014}^{(k)}$ was a dummy variable indicating the principal had a tenure of k years at his or her current school, and $v_{p,2014}$ was a random error term.

The key coefficient of interest, λ_1, captured the average change in the FFL score (measured in points on the FFL) for a one-unit change in principal value-added (measured in standard deviations of principal value-added). To make comparisons with prior research, we also expressed λ_1 in standard deviations of FFL scores (by dividing λ_1 by the standard deviation of FFL scores), which effectively produced a correlation coefficient between value-added estimates and FFL scores. We estimated equation (3) separately for each type of FFL score (full FFL score, each domain score, and each component score) and for each academic subject of the value-added estimate (math, English language arts, science, and all subjects combined).

V. RESULTS

This section presents the findings from estimating equation 3 for different combinations of FFL measures and value-added measures. We first discuss results for the full sample of 305 principals. We then discuss findings for separate subgroups of principals defined by their tenure leading their school and their school's grade span.

5a. Overall findings

Principals' scores on several parts of the FFL, despite having a limited range, were positively and significantly related to their value-added estimates. As shown in Table 6, estimated regression coefficients were statistically significant ($p < .05$) for the full FFL and in both domains 2 (systems leadership) and 4 (professional and community leadership). Most of these statistically significant relationships were with principal value-added in math. For example, a principal whose math value-added was one standard deviation above average is expected to have a full FFL score that is 0.05 points above average (on a 0 to 3 scale). Dividing the 0.05 by the standard deviation of full FFL scores from Table 5 converts the estimate to 0.17 standard deviations above average, placing the principal at approximately the 56th percentile of professional practice scores. The correlation coefficient between full FFL scores and value-added estimates in math, which is also 0.17[7], is at the low end of the 0.20 to 0.30 range that is often found in the context of teacher evaluations (Kane & Staiger, 2012; Lipscomb et al., 2015). Nevertheless, it constitutes a positive test of the FFL's concurrent validity for contributions to student achievement outcomes.

[7] Regression coefficients can be converted to correlation coefficients by multiplying them by $\delta_{VAM}/\delta_{FFL}$. Because the value-added estimates are calculated in z-score units, their standard deviation is one.

Table 6. Associations between the Framework for Leadership scores in 2013/14 and principal value-added estimates in 2013/14

		Predicted difference in FFL scores between principals at the 84th and 50th percentiles of value-added		
Outcome	Value-added measure	Estimate	*p* value	Correlation coefficient
Score on the full FFL	All subjects	0.04	.070	0.13
	Math	0.05	.017	0.17
	English language arts	0.02	.412	0.07
	Science	0.04	.060	0.13
Score on domain 1 (Strategic/cultural leadership)	All subjects	0.03	.242	0.08
	Math	0.05	.056	0.13
	English language arts	-0.00	.857	-0.00
	Science	0.04	.138	0.10
Score on domain 2 (Systems leadership)	All subjects	0.05	.037	0.17
	Math	0.05	.008	0.17
	English language arts	0.02	.287	0.07
	Science	0.05	.045	0.17
Score on domain 3 (Leadership for learning)	All subjects	0.03	.221	0.08
	Math	0.03	.214	0.08
	English language arts	0.02	.426	0.05
	Science	0.04	.121	0.10
Score on domain 4 (Professional and community leadership)	All subjects	0.06	.048	0.15
	Math	0.07	.011	0.18
	English language arts	0.03	.216	0.08
	Science	0.05	.089	0.13

Bold indicates statistical significance at p = .05.

Notes: FFL=Framework for Leadership. Correlation coefficients are obtained by dividing the value in the "Estimate" column by the respective full or domain-level standard deviation of FFL scores from Table 5. The sample size for this table includes 305 principals.

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

Table 6 reveals several other positive and statistically significant relationships as well. In particular, professional practice scores in domain 2 (systems leadership) were significantly related to principal value-added overall and in math and science. Scores in domain 4 (professional and community leadership) were significantly related to principal value-added overall and in math.[8] The magnitudes of these relationships convert to correlation coefficients between 0.15 and 0.18. Several marginally significant relationships (p < .10), such as between

[8] McCullough et al. (2016) estimate relationships between individual FFL components and value-added measures. Component 4b (shows professionalism) has the largest relationship of any individual component with principal value-added across all subjects combined and in math, and is likely driving the relationship between domain 4 scores and estimated value-added.

full FFL scores and value-added both overall and in science, were similarly sized and provide suggestive evidence that larger principal samples would detect additional statistically significant relationships. We found no statistically significant evidence of a relationship between principal value-added and FFL scores in domains 1 (strategic and cultural leadership) or 3 (leadership for learning).[9] We also found no evidence of a relationship between FFL scores and principal value-added in English language arts.

5b. Subgroup findings based on school tenure length and grade span

The main findings about the FFL's concurrent validity could vary based on the characteristics of principals or the schools they lead. We examine this issue by estimating separate relationships between FFL scores and value-added estimates for separate subgroups of principals based on their tenure length as their school's leader and their school's grade span. The subgroup findings provide greater context for understanding which groups of principals' professional practice scores are most strongly related with their estimated value-added.

School tenure may affect the size of relationships between FFL scores and value-added estimates. Prior research studies (e.g., Coelli & Green, 2012) suggest that principals need several years to have an impact on the schools they lead. Principals in their first years at a school may exhibit effective leadership practices, but their current estimated value-added may not accurately represent their practices and future value-added. The value-added of longer-serving principals, in contrast, may better align with their leadership practices.

We explored this possibility by separating the full principal sample (n=305) into those with one to two years of school tenure (n=154) and three to six years of school tenure (n=151), and estimating separate relationships in each group.[10] The findings, reported in Table 7, indicate that the statistically significant relationships in the full sample appeared to be driven mostly by evaluations of principals with at least three years of school tenure. In particular, the same patterns emerged for the longer-serving principals as for the full sample—statistically significant regression coefficients for the full FFL and in both domains 2 (systems leadership) and 4 (professional and community leadership), and for math outcomes. For this principal group, their value-added in science was also related to several FFL measures (full FFL score, and domain 1 and 4 scores). In contrast, none of the estimated relationships between FFL and value-added measures were statistically significant for principals with one or two years of school tenure.

An alternative explanation for these findings is that supervisors of longer-serving principals have more information than other supervisors about the academic performance of students at the school under the principal's leadership. Instead of rating longer-serving principals on FFL practices, supervisors may implicitly rate them on principal value-added in prior years. We

[9] Math value-added had a marginally significant relationship ($p = 0.056$) with FFL scores in domain 1 (strategic/cultural leadership). In addition, several components in domains 1 and 3 may be associated with value-added scores (McCullough, et al., 2016). That study found that higher scores on components 1b (uses data for informed decisionmaking), 1c (builds a collaborative and empowering work environment), 3c (implements high-quality instruction), and 3e (maximizes instructional time) were related at a marginal level of significance to value-added in all subjects combined and in math.

[10] Recall that principals in the study sample could have at most six years of school tenure, due to the research design and limitations on data availability.

examine this possibility in Table 8, which regresses principal value-added in 2012/13 on FFL scores in 2013/14. None of the findings for either group of principals is statistically significant, suggesting that supervisors' prior knowledge about principal value-added does not appear to relate to their ratings of principals' leadership practices. As a result, we conclude that the FFL most effectively reflects contributions to student achievement for principals with at least three years of school tenure, because they have had an opportunity to make an impact on their school.

Table 7. Associations between the Framework for Leadership scores in 2013/14 and principal value-added estimates in 2013/14, by tenure length as school's leader

		Predicted difference in FFL scores between principals at the 84th and 50th percentiles of value-added			
		Tenure: 1 or 2 years		Tenure: 3 to 6 years	
Outcome	Value-added measure	Estimate	p value	Estimate	p value
Score on the full FFL	All subjects	0.03	.252	0.07	.113
	Math	0.03	.213	0.07	.031
	English language arts	0.02	.379	0.02	.570
	Science	0.01	.701	0.08	.040
Score on domain 1 (Strategic/cultural leadership)	All subjects	0.02	.499	0.05	.286
	Math	0.04	.245	0.06	.102
	English language arts	0.00	.952	-0.01	.831
	Science	-0.00	.912	0.08	.036
Score on domain 2 (Systems leadership)	All subjects	0.02	.359	0.09	.042
	Math	0.03	.223	0.08	.010
	English language arts	0.00	.874	0.05	.208
	Science	0.02	.470	0.08	.051
Score on domain 3 (Leadership for learning)	All subjects	0.03	.228	0.05	.346
	Math	0.02	.407	0.05	.263
	English language arts	0.03	.184	0.02	.758
	Science	0.02	.511	0.07	.134
Score on domain 4 (Professional and community leadership)	All subjects	0.03	.310	0.08	.079
	Math	0.03	.372	0.10	.013
	English language arts	0.03	.229	0.04	.430
	Science	0.00	.894	0.09	.043

Bold indicates statistical significance at p = .05.

Notes: FFL=Framework for Leadership. The sample size of principals with one or two years of tenure is 154. The sample size of principals with three to six years of tenure is 151.

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

Table 8. Associations between the Framework for Leadership scores in 2013/14 and principal value-added estimates in 2012/13

		Predicted difference in FFL scores between principals at the 84th and 50th percentiles of value-added		
Outcome	Value-added measure	Estimate	p value	Number of principals
Score on the full FFL	All subjects	0.04	.272	107
	Math	0.04	.325	107
	English language arts	0.04	.305	107
	Science	-0.01	.640	107
Score on domain 1 (Strategic/cultural leadership)	All subjects	0.09	.053	107
	Math	0.07	.125	107
	English language arts	-0.07	.139	107
	Science	0.03	.387	107
Score on domain 2 (Systems leadership)	All subjects	0.02	.493	107
	Math	0.03	.519	107
	English language arts	0.04	.358	107
	Science	-0.04	.165	107
Score on domain 3 (Leadership for learning)	All subjects	0.04	.366	107
	Math	0.04	.404	107
	English language arts	0.05	.265	107
	Science	-0.02	.512	107
Score on domain 4 (Professional and community leadership)	All subjects	0.01	.789	107
	Math	0.02	.609	107
	English language arts	0.01	.823	107
	Science	-0.03	.548	107

Bold indicates statistical significance at p = .05.

Notes: FFL=Framework for Leadership.

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

In addition to examining the FFL's concurrent validity for different school tenure groups, we estimated separate relationships for principals in elementary (n=155), middle (n=81), and high (n=69) schools. Higher FFL scores were associated with larger value-added among middle school principals, but we did not detect any relationships for elementary school principals or high school principals (Table 9). Among middle school principals, scores in domain 1 (strategic and cultural leadership) were significantly positively associated with value-added in all subjects combined. Middle school principals' value-added in all subjects combined also had a marginally significant and positive relationship with their full FFL scores and domain 4 scores. The magnitude of the three estimated relationships exceeded the size of all those detected across the full sample. We did not detect any associations between full FFL or domain scores and principal value-added in all subjects combined for either elementary school principals or high school principals. This finding may reflect that value-added estimates typically cover a larger proportion of grades for middle schools than for elementary and high schools and thus are more accurate measures of schoolwide performance. The smaller sample sizes for this grade span analysis also made it more difficult to detect statistically significant relationships.

Table 9. Associations between the Framework for Leadership scores in 2013/14 and principal value-added estimates (all subjects) in 2013/14, by grade span of the school

		Predicted difference in FFL scores between principals at the 84th and 50th percentiles of value-added		
Grade span	Outcome	Estimate	p value	Number of principals
Elementary	Score on the full FFL	0.02	.487	155
	Score on domain 1 (Strategic/cultural leadership)	0.00	.928	155
	Score on domain 2 (Systems leadership)	0.03	.144	155
	Score on domain 3 (Leadership for learning)	0.00	.887	155
	Score on domain 4 (Professional and community leadership)	0.04	.349	155
Middle	Score on the full FFL	0.12	.072	81
	Score on domain 1 (Strategic/cultural leadership)	0.14	.040	81
	Score on domain 2 (Systems leadership)	0.12	.114	81
	Score on domain 3 (Leadership for learning)	0.12	.148	81
	Score on domain 4 (Professional and community leadership)	0.11	.096	81
High	Score on the full FFL	-0.02	.627	69
	Score on domain 1 (Strategic/cultural leadership)	-0.04	.255	69
	Score on domain 2 (Systems leadership)	-0.01	.782	69
	Score on domain 3 (Leadership for learning)	-0.02	.518	69
	Score on domain 4 (Professional and community leadership)	0.01	.750	69

Bold indicates statistical significance at p = .05.

Notes: FFL=Framework for Leadership. Analyses are based on a value-added measure that combines all subjects, and the analysis sample consists of all principals participating in the 2013/14 pilot year who have a value-added measure. Elementary schools are defined as those with no grade above 6; middle schools are defined as those with at least one grade above 6 but no grades above 8; high schools are defined as those with at least one grade above 8.

Source: Authors' calculations based on Framework for Leadership pilot evaluation data from 2013/14, student achievement and background data, and principals' job assignment data provided by the Pennsylvania Department of Education.

VI. CONCLUSION

The findings from this study provide the first evidence of the FFL's concurrent validity, and the first evidence that any measure of principals' professional practice is correlated with their contributions to student achievement. FFL scores differentiate (to a limited extent) principals who make larger or smaller contributions to student achievement. Higher full scores and scores in two of the four domains are significantly or marginally significantly associated with value-added in all subjects combined and with value-added in math specifically, despite limited variation in FFL scores. The coefficient estimates on the marginally significant and nonsignificant relationships suggest that future studies with larger principal samples may be able to detect additional statistically significant relationships. The findings also suggest that including a measure of professional practice similar in content and structure to the FFL is a viable option for states and districts that seek to employ a multiple measures approach to evaluating principals.

Evidence of concurrent validity sets the FFL apart from other principal evaluation tools examined in the literature. Only two other studies have examined validity of principal evaluation tools for student achievement outcomes, focusing on a small number of district-specific evaluation instruments. Neither study found any robust evidence of a relationship between the instruments and principals' value-added (Grissom et al., 2015a; Milanowski & Kimball, 2012). In contrast, findings from this study are based on a sample of principals from nearly 200 school districts in Pennsylvania.

The study also describes a method for estimating principal contributions to student outcomes that less severely restricts the ability to compare effectiveness across a broad set of school principals. Our approach adjusts the value-added of a principal's school for its value-added in the year before the principal began leading the school, thereby controlling for school-specific factors beyond the principal's control. In contrast, previous studies have either mistakenly attributed the effectiveness of entire schools to the effectiveness of the principal alone or used methods that permitted comparisons only among small sets of principals connected through leadership of the same schools. We view this method as a contribution to a field that currently lacks consensus on the most theoretically satisfying and practically realistic way of making large-scale comparisons of principal value-added.

This method also allows us to obtain separate findings for several subgroups of principals, including those based on school tenure and grade span. For example, we find evidence suggesting that the FFL has concurrent validity primarily among principals with three to six years of tenure as their school's leader. We interpret this evidence as consistent with other research suggesting that principals require multiple years to have an impact on a school (Coelli & Green, 2012). A limitation of the study is that we cannot estimate value-added for principals with more than six years of tenure, although this limitation will become less important for policy and research as data systems accumulate additional years of information.

The research base on principals' effects on student outcomes is small but growing. We suggest two promising avenues for future research. First, we recommend that future studies attempt to validate principal evaluation tools in a high-stakes setting. The findings from this study came from a low-stakes pilot, and it remains to be seen whether principals' practices or their supervisors' ratings of their practices would be different under circumstances in which the ratings have consequences. Second, we recommend that future studies attempt to validate principal value-added measures as multiple studies have done for teacher value-added measures (Chetty, Friedman, & Rockoff, 2014; Kane & Staiger, 2008). Although we argue that our principal value-added measure is valid on conceptual grounds, no principal value-added measure has been analytically validated using experimental or quasi-experimental methods. Obtaining this information would contribute directly to efforts by policymakers and researchers to assess the accuracy of principal evaluation tools.

REFERENCES

Branch, G., Hanushek, E., & Rivkin, S. (2012). Estimating the effect of leaders on public sector productivity: The case of school principals (NBER No. 17803). National Bureau of Economic Research Working Paper. Cambridge, MA.

Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). Professional practice, student surveys, and value.added: Multiple measures of teacher effectiveness in the Pittsburgh Public Schools. (REL 2014–024). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). “Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates.” American Economic Review, 104(9): 2593-2632.

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

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

Condon, C., & Clifford, M. (2012). Measuring principal performance: How rigorous are commonly used principal performance assessment instruments? Washington, DC: American Institutes for Research.

Deutsch, J. (2014). Proposing a test of the value-added model using school lotteries. Working paper. Chicago, IL: Mathematica Policy Research.

Dhuey, E., & Smith, J. (2014). How effective are school principals in the production of school achievement? Canadian Journal of Economics, 47(2), 634-663.

Dhuey, E., & Smith, J. (2013). How school principals influence student learning. Working paper. Toronto, ON: University of Toronto.

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., Cravens, X. C., Murphy, J., Porter, A. C., & Elliot, S. N. (2012). The convergent and divergent validity of the Vanderbilt Assessment of Leadership in Education (VAL-ED): Instructional leadership and emotional intelligence. Working paper. Nashville, TN: Vanderbilt University.

Grissom, J. A., Blissett, R. S. L., & Mitani, H. (2015b). Supervisor ratings as measures of principal performance: Evidence from the TEAM evaluation system in Tennessee. Working paper.

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

Kane, T. J., & Staiger, D. O. (2008). “Estimating teacher impacts on student achievement: An experimental evaluation.” NBER working paper no. 14607.

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Seattle, WA: Bill & Melinda Gates Foundation.

Lipscomb, S., Chiang, H., & Gill, B. (2012). Value-added estimates for phase 1 of the Pennsylvania teacher and principal evaluation pilot. Cambridge, MA: Mathematica Policy Research.

Lipscomb, S., Terziev, J., & Chaplin, D. (2015). Measuring teachers’ effectiveness: Report from phase 3 of Pennsylvania’s pilot of the Framework for Teaching. Cambridge, MA: Mathematica Policy Research.

McCullough, M., Lipscomb, S., Chiang, H., Gill, B., and Cheban, I. (2016). Measuring school leaders’ effectiveness: Final report from a multiyear pilot of Pennsylvania’s Framework for Leadership (REL 2016–106). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic.

Milanowski, A., & Kimball, S. (2012). The relationship between standards-based principal performance evaluation ratings and school value-added: Evidence from two districts. Rockville, MD: Westat.

Morris, C. (1983). Parametric Empirical Bayes Inference: Theory and Applications (with discussion). Journal of the American Statistical Association, 78(381), 47-65.

Pennsylvania Department of Education. (2014). Framework for Leadership. Retrieved from the Pennsylvania Department of Education website: http://www.education.pa.gov/Documents/Teachers-Administrators/Educator%20Effectiveness/Principals%20and%20CTC%20Directors/Principal%20Effectiveness%20Framework%20for%20Leadership.pdf.

Porter, A. C., Murphy, J., Goldring, E., Elliott, S. N., Polikoff, M. S., & May, H. (2008). Vanderbilt Assessment of Leadership in Education: Technical manual. New York, NY: Wallace Foundation.

Porter, A. C., Polikoff, M. S., Goldring, E., 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.

Rotz, D., Johnson, M., & Gill, B. Value-added models for the Pittsburgh Public Schools, 2012–13 school year. Cambridge, MA: Mathematica Policy Research.

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). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic.

ABOUT THE SERIES

Policymakers and researchers require timely, accurate, evidence-based research as soon as it’s available. Further, statistical agencies need information about statistical techniques and survey practices that yield valid and reliable data. To meet these needs, Mathematica’s working paper series offers access to our most current work.

For more information about this paper, contact Moira McCullough, Researcher, at mmccullough@mathematica-mpr.com.

Suggested citation: Moira McCullough, Stephen Lipscomb, Hanley Chiang, and Brian Gill. “Do Principals’ Professional Practice Ratings Reflect Their Contributions to Student Achievement? Evidence from Pennsylvania’s Framework for Leadership.” Working Paper 46. Cambridge, MA: Mathematica Policy Research, June 2016.

www.mathematica-mpr.com

Improving public well-being by conducting high quality, objective research and data collection

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■ TUCSON, AZ ■ WASHINGTON, DC

MATHEMATICA
Policy Research