Making Connections

November 2014

The utility of teacher and student surveys in principal evaluations: An empirical investigation

Keke Liu
Basis Policy Research

Jeff Springer
Basis Policy Research

David Stuit
Basis Policy Research

Jim Lindsay
American Institutes for Research

Yinmei Wan
American Institutes for Research

Key findings

Does adding teacher and student survey measures to existing measures (such as supervisor ratings and student attendance rates) increase the power of principal evaluation models to explain across-school variance in value-added achievement gains? This study found that the classroom instructional environment measure and the instructional leadership measure significantly increase the explained across-school variance in value-added achievement gains—by 28.8 percentage points in math and by 26.5 percentage points in a composite of math and reading.

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences
U.S. Department of Education

REL MIDWEST
Regional Educational Laboratory
At American Institutes for Research

REL 2015–047

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts unbiased large-scale evaluations of education programs and practices supported by federal funds; provides research-based technical assistance to educators and policymakers; and supports the synthesis and the widespread dissemination of the results of research and evaluation throughout the United States.

November 2014

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-IES-12-C-0004 by Regional Educational Laboratory Midwest, administered by American Institutes for Research. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Liu, K., Springer, J., Stuit, D., Lindsay, J., & Wan, Y. (2014). *The utility of teacher and student surveys in principal evaluations: An empirical investigation.* (REL 2015–047). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest. Retrieved from http://ies.ed.gov/ncee/edlabs.

This report is available on the Regional Educational Laboratory website at http://ies.ed.gov/ncee/edlabs.

Summary

In recent years states and school districts in the Midwest Region have developed new evaluation models that hold principals accountable for their contributions to student achievement. Many districts are turning to teacher and student feedback surveys to evaluate principals based on school conditions that promote better teaching and learning.

This study examines whether adding such survey measures to an existing principal evaluation model that includes supervisor ratings and student attendance rates improves the model's ability to explain variation in schools' average value-added achievement gains. Using data from one midsize urban school district in the Midwest Region, this study investigates the incremental utility of six candidate survey measures—four teacher survey measures (instructional leadership, professional learning community, quality of professional development, and cultural press for excellence) and two student survey measures (classroom instructional environment and school safety and climate). Incremental utility is defined as the degree to which a candidate survey measure explains the across-school variance in value-added achievement gains above and beyond the district's existing principal evaluation measures (supervisor ratings and student attendance rates).

Data from school year 2011/12 for 39 elementary and secondary schools in the district included responses from teacher and student feedback surveys, supervisor ratings of principals, and student achievement, demographic, and attendance records. A two-step multivariate regression analysis was used to answer the research questions. The first step assessed the incremental utility of the six survey measures in significantly increasing the model's explained variance. The second step examined whether the survey measures that demonstrated significant incremental utility in the first step could be reduced to an optimal subset of measures that made the most significant incremental contributions.

Findings indicate that adding teacher and student survey measures on school conditions to the principal evaluation model can strengthen the relationship between principals' evaluation results and their schools' average value-added achievement gains in math and in a composite of math and reading. Neither teacher nor student survey measures showed significant incremental utility in explaining across-school variance in reading.

The complete set of candidate survey measures could be reduced to an optimal subset of two measures: instructional leadership from the teacher survey and classroom instructional environment from the student survey. These two measures were found to have significant incremental utility in explaining across-school variance in value-added achievement gains in math and in a composite of math and reading. Together, these two measures increased the explained across-school variance in value-added achievement gains in math by 28.8 percentage points, or 73.6 percent of the incremental utility contributed by all six survey measures. The same two measures also increased the explained across-school variance in value-added achievement gains in a composite of math and reading by 26.5 percentage points, or 73.2 percent of the incremental utility contributed by all six survey measures.

The measure of classroom instructional environment represents the core school condition most closely linked to student learning. The finding of significant incremental utility for this measure echoes the results of prior studies on the important influence of classroom and teacher quality on student achievement. The finding of significant incremental utility for the instructional leadership measure aligns with principal leadership studies that have found a strong influence on student achievement for leadership practices that focus on curriculum and instruction.

Contents

Boxes

Figures

Tables

Why this study?

Improving the evaluation of principal performance is a priority among states and school districts in the Midwest Region. Since 2009, five of the seven states in the region have adopted new administrative rules, legislative codes, or state education policies on principal evaluation models.[1] These new models require districts to move beyond a single evaluation instrument and rely on multiple performance measures, such as growth in student achievement, leadership competency assessments, and school climate surveys, to create a more complete picture of principal effectiveness (Clifford, Behrstock-Sherratt, & Fetters, 2012; Illinois Principals Association & Association of School Administrators, 2012; Mattson Almanzán, Sanders, & Kearney, 2011; The New Teacher Project, 2012; Ohio Department of Education, 2011; Roeber, 2011; Wisconsin Educator Effectiveness Design Team, 2011).

Policymakers' and researchers' calls for multiple-measure evaluation models have compelled many districts to search for new measures to add to their existing set of principal evaluation tools. Districts are particularly interested in understanding the utility of incorporating teacher and student feedback surveys (Illinois State Board of Education, 2011; Mattson Almanzán et al., 2011; National Conference of State Legislatures, 2011; Wacyk, Reeves, McNeill, & Zimmer, 2011). Feedback surveys can provide rich information about a principal's role in shaping school conditions that promote better teaching and learning (Clifford, Menon, Gangi, Condon, & Hornung, 2012; Hallinger & Heck, 1996; National Center for Education Statistics, 1997).

This study contributes new information on the degree to which adding teacher and student survey measures to existing principal evaluation measures strengthens the correlation between these measures and a school's value-added achievement gains

To make an informed decision on whether to add feedback surveys to principal evaluation models, districts need to understand the associated costs and benefits. The costs include the expense of administering the surveys and the time required for teachers and students to complete them. The benefits can be judged by incremental utility (Haynes & O'Brien, 2000), or the degree to which the new survey measures improve the power of principal evaluation models to explain the across-school variance in student achievement outcomes for which principals are held accountable (Milanowski & Kimball, 2012; Sanders, Kearney, & Vince, 2012). Evaluation model designers on tight budgets need to know that investing in a new measure will yield relevant information that is not contained in existing measures.

Yet few resources are available to help districts assess the incremental utility of measures considered for inclusion in their performance evaluation models. The research literature offers criteria for judging the technical quality of standalone measures[2] but little on how to determine whether introducing a new measure will improve the evaluation model's overall quality of information (see appendix A for a literature review).

This study contributes new information on the degree to which adding teacher and student survey measures to existing principal evaluation measures strengthens the correlation between these measures and a school's value-added achievement gains. This information will help district superintendents, principals, and other district leaders in the Midwest Region and elsewhere understand the quality and utility of these surveys and make informed decisions on whether and how to include them in principal evaluations (Clifford, Menon, et al., 2012). More generally, this study contributes to the emerging body of research on principal evaluation by demonstrating a process for evaluating the incremental utility of measures that are candidates for inclusion in evaluation models.

What the study examined

Using data from one midsize urban school district in the Midwest Region, this study investigated the incremental utility of four teacher survey measures and two student survey measures in strengthening the correlation between principal evaluation results and school value-added achievement gains.

Two research questions guided the study:

- Does adding the teacher and student feedback survey measures to an existing set of principal performance measures improve the power of the principal evaluation model to explain variance in across-school value-added achievement gains?
- Can the full set of six survey measures be reduced to an optimal subset of measures that make significant incremental contributions to the link between principal evaluation models and school value-added achievement gains?

This study investigated the incremental utility of four teacher survey measures and two student survey measures in strengthening the correlation between principal evaluation results and school value-added achievement gains

The study team analyzed survey, evaluation, achievement, and other administrative data from school year 2011/12 for 39 elementary and secondary schools in the Midwest Region district to explain the across-school variance in value-added achievement gains in math, reading, and a composite of both subjects. School value-added achievement gains in each subject and grade (grades 3–11) were estimated using students' math and reading test scores on the Northwest Evaluation Association's (2011) Measures of Academic Progress (MAP) assessment. These grade-level value-added gains were aggregated by subject to the school level to produce the average math and reading gains and then combined across subjects to obtain the composite math and reading gain. To aid in understanding the analyses and interpreting the findings, box 1 defines key terms used in the report, and box 2 briefly describes the data, measures, and methodology (more detailed information is in appendix B).

Box 1. Key terms

Candidate measures. The six performance measures under consideration for inclusion in the principal evaluation model. Four measures are from the teacher feedback survey: instructional leadership, professional learning community, quality of professional development, and cultural press for excellence. Two measures are from the student feedback survey: classroom instructional environment and school safety and climate.

Core Competency Assessment. The participating district's principal evaluation instrument. It is an evaluation instrument consisting of two rubrics that assess principals on 11 core competencies.

Correlation. A statistic that indicates the degree to which two measures are related. Correlation coefficients range from –1 to 1. A correlation of 0 indicates no relationship between the two measures. A correlation of 1 indicates a perfect positive relationship, and a correlation of –1 indicates a perfect negative (inverse) relationship.

Cronbach's alpha. A commonly used statistic to estimate internal consistency, or the reliability with which a set of test, evaluation, or survey items delivers consistent results. In this study Cronbach's alpha is estimated to confirm the internal consistency of the Core Competency Assessment.

(continued)

Box 1. Key terms *(continued)*

Existing measures. The two state-required performance measures already in the principal evaluation model: principal supervisor ratings and school attendance.

Explained variance. The proportion of the variance in the outcome measure (or variable) that a regression model accounts for. In linear regression the explained variance equals the coefficient of determination, or R^2.

Incremental utility. The degree to which the introduction of a new measure increases the power of a regression model to explain the variance in the outcomes of interest relative to the model's existing measures.

Measures of Academic Progress. The Northwest Evaluation Association's (2011) benchmark assessment, which is administered in math and reading three times a year to all students in grades 3–11 in the district studied.

Student survey. The Tripod Student Perception Survey (developed by Harvard researcher Ronald Ferguson and distributed by Cambridge Education), which consists of 36 items measuring students' perceptions of their classroom instructional environment in seven domains and six items measuring students' perceptions of school safety and climate. The survey was administered to all students in grades 3–12 in the district studied. The two student survey measures (classroom instructional environment and school safety and climate) were candidate measures for inclusion in the principal evaluation model. See appendix E for the survey items used to measure classroom instructional environment, which were released publicly in a 2010 report by the Bill & Melinda Gates Foundation's Measures of Effective Teaching Project (Kane & Cantrell, 2010). The survey items used to measure school safety and climate are not presented because Cambridge Education maintains exclusive intellectual property rights to them.

Supervisor ratings. Principals' overall evaluation results on the district's Core Competency Assessment. District principals were observed and evaluated by supervisors during school year 2011/12.

Teacher survey. Cambridge Education's Tripod Teacher Survey, which consists of 82 items that measure teachers' perceptions of school organizational conditions and self-reflections on their instructional practice. The survey was administered to approximately 800 district teachers in spring 2012. This study derives four measures from the teacher survey as candidate measures for inclusion in the principal evaluation model: instructional leadership, professional learning community, quality of professional development, and cultural press for excellence. The survey items from the teacher survey are not presented because Cambridge Education maintains exclusive intellectual property rights to them.

Value-added achievement gains. Grade-level results (for grades 3–11) from the value-added model that are aggregated at the school level for math, reading, and a composite of both subjects. Other terms that appear in the research literature are "value-added scores," "value-added estimates," "value-added measures," and "value-added effects."

Value-added model. A statistical technique to estimate the contributions of schools to their students' achievement growth by examining changes in test scores over time. The value-added model used in this study is a two-stage covariate adjustment model.

Box 2. Data, measures, and methods

Data. The participating district provided districtwide data files from school year 2011/12. The final analytical sample included 20 elementary schools (grades K–5), 13 secondary schools (grades 6–12), and 6 schools with grades spanning both ranges. These 39 schools were selected because they had the most complete data on all three sets of variables: student outcome measures, the existing principal evaluation measures, and the teacher and student survey measures.

School value-added achievement gains were estimated from the math and reading test scores of 7,709 grade 3–11 students in the fall and spring administrations of the Northwest Evaluation Association's (2011) Measures of Academic Progress assessments. The survey data included 541 teacher responses to Cambridge Education's Tripod teacher feedback survey and 8,345 responses from students in grades 3–12 to the Tripod student feedback survey. The school average attendance rate was based on the attendance records of 16,537 students in grades K–12.

Measures. The analysis used three sets of measures: student outcome measures, existing principal evaluation measures, and teacher and student survey measures (candidate measures).

Student outcome measures. The analysis focused on three outcome measures: subject-specific school value-added achievement gains in math and reading and value-added achievement gains in a composite of both subjects. The value-added estimates were calculated with the widely used covariate-adjustment model (Kane & Staiger, 2008; McCaffrey, Lockwood, Koretz, & Hamilton, 2003). Scores from the fall administration of the Measures of Academic Progress assessment were included as covariates in the value-added model, and scores from the spring administration served as outcome variables. Student background information—gender, English language learner status, special education status, and mobility—was used for control variables.

Existing principal evaluation measures. The existing principal evaluation measures include two measures mandated by state law in the principal evaluation model: principal supervisor ratings and school average attendance rates. Results from the district's supervisor rating instrument were used to construct two measures of principal competency: one in job function competency and one in leadership skill competency. Scores on these two measures were then averaged to obtain a composite supervisor rating for each principal. Student attendance records were aggregated to the school level to calculate school average attendance rates.

Teacher and student survey measures (candidate measures). Six multiscale variables, four derived from the Tripod teacher survey and two from the Tripod student survey, were considered candidate measures for the principal evaluation model. The four teacher survey measures reflect teachers' perceptions of school instructional leadership, professional learning community, quality of professional development, and cultural press for excellence. The two student survey measures reflect students' perceptions of classroom instructional environment and school safety and climate. (The literature review in appendix A explains the school working conditions as measured by the teacher and student survey measures and the relationship between the survey measures and school performance; see appendix E for the student survey items used to measure perceptions of classroom instructional environment.)

(continued)

Box 2. Data, measures, and methodology *(continued)*

Methods. A two-step regression analysis was used to address the research questions. The first step assessed the incremental utility of the candidate measures in explaining the across-school variance in value-added achievement gains beyond the existing principal evaluation measures. The second step examined whether the candidate measures could be reduced to an optimal subset of measures that made significant incremental contributions to strengthening the correlation between the principal evaluation models and the three school value-added outcomes.

Step 1: Estimating the incremental utility of the six survey measures. The basic analytic strategy was to test whether a regression model that included the candidate survey measures and the existing evaluation measures as independent variables explained more of the across-school variance in value-added achievement gains than the baseline regression model that included only the existing measures.

First, the study team tested whether adding the full set of six candidate measures to the existing measures led to a statistically significant increase in explained variance (R^2). An F-test was conducted to compare the R^2 of the model that added the six candidate survey measures with the R^2 of the baseline model that used only the two existing measures. A p-value of 0.10 was required to reject the null hypothesis of no significant difference in the explained variance between the two models.[1] A significant difference in the R^2 between the two models would be evidence that adding the full set of six candidate measures strengthens the relationship between the principal evaluation model and school value-added achievement gains.

Next, the study team tested the joint significance of two subsets of the candidate measures: the subset of the two student survey measures and the subset of the four teacher survey measures. Finally, the six candidate measures were entered into the regression models separately to examine the individual incremental utility of each survey measure. All regressions were based on the same sample of 39 schools.

Step 2: Determining an optimal subset of survey measures. In this step, candidate measures were entered sequentially into the regression model according to their estimated incremental utility in the first-step analysis. Only candidate measures found to significantly increase the explained across-school variance in value-added achievement gains were used in this step. These measures were entered into the regression model in descending order of incremental utility from the first-step analysis. For example, if the classroom instructional environment measure was found to have the largest incremental utility in the math model from the first-step analysis, it would be the first measure added into the baseline model for math. After each candidate measure was entered, an F-test was conducted to determine the significance of its incremental utility. The optimal subset of candidate measures was attained when the entry of the next candidate measure failed to make a significant incremental contribution to explaining the variance in the outcome measure.

Note

[1] A p-value of 0.10 is a common cutoff value in variable selection processes and incremental validity research (Bendel & Afifi, 1977; Mickey & Greenland, 1989). Using lower significance levels such as 0.05 increases the risk of eliminating candidate measures that research and theory suggest are important to school performance but cannot achieve statistical significance due to limited sample size.

What the study found

Two survey measures—classroom instructional environment and instructional leadership—significantly contributed to the incremental utility of the existing principal evaluation models in explaining the across-school variance in math and composite value-added achievement gains (see box 3 for a brief description of these measures and appendix A for more detail). Together, these two measures increased the explained across-school variance in value-added gains by 28.8 percentage points in math and by 26.5 percentage points in a composite of math and reading.

The first measure, classroom instructional environment, represents the core school condition most closely linked to student learning; its significant incremental utility echoes prior studies on the important influence of classroom and teacher quality on student achievement (see Aaronson, Barrow, & Sander, 2007; Gordon, Kane, & Staiger, 2006; Kane & Cantrell, 2010; Kane, Rockoff, & Staiger, 2006; Rivkin, Hanushek, & Kain, 2005). The significant incremental utility of the second measure, instructional leadership, agrees with principal leadership studies that suggest leadership practices focusing on curriculum and instruction can improve student achievement (Hattie, 2009; Robinson, Lloyd & Rowe, 2008; Witziers, Bosker, & Krüger, 2003).

Two survey measures—classroom instructional environment and instructional leadership—significantly contributed to the incremental utility of the existing principal evaluation models in explaining the across-school variance in math and composite value-added achievement gains

The grouped sets of survey measures (teacher survey, student survey, or both surveys together) were also significant in explaining the across-school variance in math and composite value-added achievement gains. However, these increases can largely be attributed to the two individual survey measures identified previously.

Box 3. What are classroom instructional environment and instructional leadership?

Classroom instructional environment

Effective principals directly influence the quality of the classroom instructional environment through the strategic hiring, development, and retention of good teachers. Although some districts limit principals' authority to hire their own staff, research suggests that good principals take a proactive stand in teacher recruitment. Principals also directly influence the classroom instructional environment when they connect with teachers in their classrooms during formal and informal observations. These interactions often lead to immediate changes in instructional practice as teachers respond to feedback.

Instructional leadership

The instructional leadership measure represents the leadership activities in the school that address instruction and curriculum. It extends beyond the role of the principal to include all leadership activities of the school's staff. Strong instructional leaders are able to create and sustain a clear vision for learning, communicate school instructional goals, and garner schoolwide commitment to those goals. They also promote coherence in the instructional program by frequently visiting classrooms to monitor instruction and dialog with teachers. Numerous studies have documented the relationship between instructional leadership and student outcomes. A meta-analysis of 22 leadership studies found that the average effects of instructional leadership practices on student achievement and other outcomes (such as absenteeism and engagement) were three to four times as large as the average effects of other leadership practices that do not explicitly focus on curriculum and instruction (Robinson et al., 2008).

Neither the teacher nor the student survey measures showed significant incremental utility in explaining across-school variance in reading. The lack of effect on reading achievement is consistent with the finding from the Measures of Effective Teaching Project (Kane & Cantrell, 2010) that value-added estimates in reading are less correlated with prior value-added estimates and student feedback from the Tripod student survey.

Existing principal evaluation measures explained 3–8 percent of the across-school variance in value-added gains

The baseline models with only the two existing principal evaluation measures (principal supervisor ratings and school attendance rate) explained about 8.0 percent of the across-school variance in value-added achievement gains in math, 3.4 percent of the variance in gains in reading, and 5.1 percent of the variance in gains in a composite of math and reading (table 1).

Among the four teacher survey measures, instructional leadership significantly increased explained across-school variance in value-added achievement gains in math and in a composite of math and reading

Two teacher survey measures—instructional leadership and cultural press for excellence—showed significant incremental utility in explaining the across-school variance in value-added gains in math and a composite of math and reading

Among the four teacher survey measures, instructional leadership had the most incremental utility, with a significant increase in explained across-school variance (R^2) in value-added achievement gains of 12.1 percentage points in math and 10.7 percentage points in a composite of math and reading (see table 1). A second teacher survey measure, cultural press for excellence, also significantly increased the R^2, albeit to a lesser degree, by 8.3 percentage points in math and 8.1 percentage points in a composite of math and reading. The other two teacher survey measures, professional learning community and quality of professional development,

Table 1. Incremental utility of candidate survey measures: Explained across-school variance in school value-added achievement gains

	Math		Reading		Composite	
Measure	Increase in R^2 over baseline models	p value	Increase in R^2 over baseline models	p value	Increase in R^2 over baseline models	p value
Baseline R^2 of existing measures (supervisor ratings and school attendance)	0.080	na	0.034	na	0.051	na
Teacher survey (four measures)						
A. Instructional leadership	0.121	0.040*	0.067	0.126	0.107	0.048*
B. Professional learning community	0.048	0.186	0.036	0.265	0.051	0.180
C. Quality of professional development	0.001	0.825	0.013	0.509	0.007	0.636
D. Cultural press for excellence	0.083	0.081*	0.053	0.176	0.081	0.090*
Joint significance (A+B+C+D)	0.222	0.073*	0.090	0.553	0.179	0.168
Student survey (two measures)						
E. School safety and climate	0.055	0.287	0.007	0.761	0.031	0.350
F. Classroom instructional environment	0.145	0.041*	0.086	0.192	0.135	0.062*
Joint significance (E+F)	0.189	0.074*	0.098	0.362	0.159	0.137
Joint significance of all measures (A+B+C+D+E+F)	0.391	0.034*	0.227	0.389	0.362	0.074*

* Explained variance is significant ($p < 0.10$).

na is not applicable.

Source: Authors' analysis based on data provided by the district.

did not show significant incremental utility in explaining the across-school variance in any outcome measure. As a subset, the four teacher survey measures jointly increased the explained across-school variance in value-added achievement gains by 22.2 percentage points in math, 9 percentage points in reading, and 17.9 percentage points in a composite of math and reading, although only the increase in the math model was statistically significant.

One student survey measure—classroom instructional environment—showed significant incremental utility in explaining across-school variance in value-added gains in math and in a composite of math and reading

One student survey measure, classroom instructional environment, significantly increased the explained across-school variance in value-added achievement gains by 14.5 percentage points in math and by 13.5 percentage points in a composite of math and reading (see table 1). In reading, the instructional environment measure was associated with an 8.6 percentage point increase in explained variance, but the increase was not statistically significant. The other student survey measure, school safety and climate, did not significantly increase the explained variance in any subject. As a subset, the two student survey measures significantly improved the R^2 in the math model by 18.9 percentage points. The increase in the explained variance attributed to the subset of two student survey measures was 9.8 percentage points in the reading model and 15.9 percentage points in the composite model, but neither increase was statistically significant.

The full set of six candidate survey measures significantly increased the explained across-school variance in value-added achievement gains in math and in a composite of math and reading

The full set of six survey measures showed significant incremental utility in explaining across-school variance in value-added gains in math and a composite of math and reading

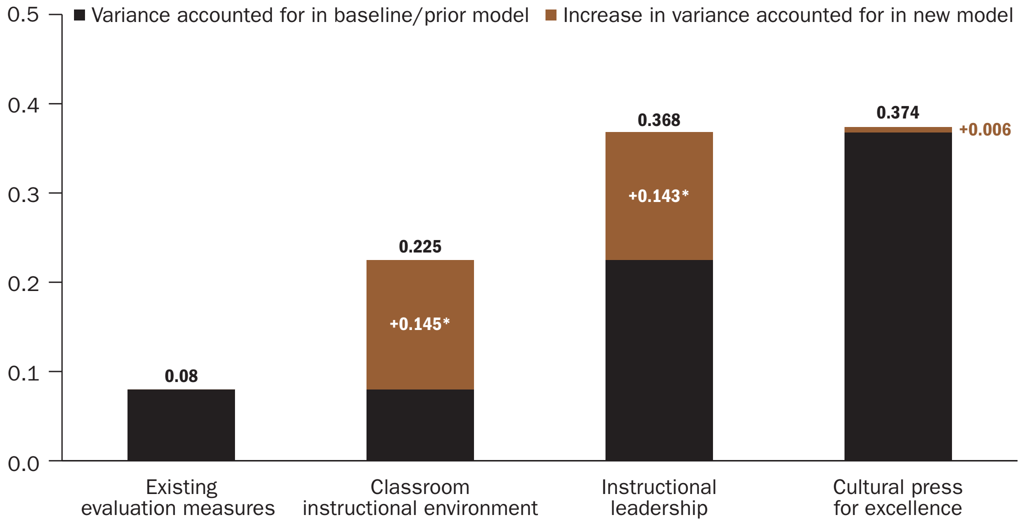
The full set of six candidate survey measures significantly increased the explained across-school variance in value-added achievement gains by 39.1 percentage points in math and by 36.2 percentage points in a composite of math and reading. The full set of measures increased the explained across-school variance in value-added achievement gains by 22.7 percentage points in reading, but the increase was not significant. Only three of the six survey measures (instructional leadership, cultural press for excellence, and classroom instructional environment) showed significant individual incremental utility.

The optimal subset of survey measures with significant incremental utility includes the classroom instructional environment and the instructional leadership measures

The three candidate survey measures that individually showed significant incremental utility in explaining the across-school variance in value-added gains—instructional leadership, cultural press for excellence, and classroom instructional environment—were used in the second step of the analysis to determine an optimal set of survey measures for the math and composite models.

Because the classroom instructional environment measure was associated with the largest incremental utility, it was entered first into the regression models, followed by the instructional leadership measure. The addition of both measures significantly increased the power of existing principal evaluation measures to explain the across-school variance in value-added achievement gains in math (figure 1) and a composite of math and reading (figure 2). Adding the cultural press for excellence measure did not improve the explained variance of either the math or the composite model.

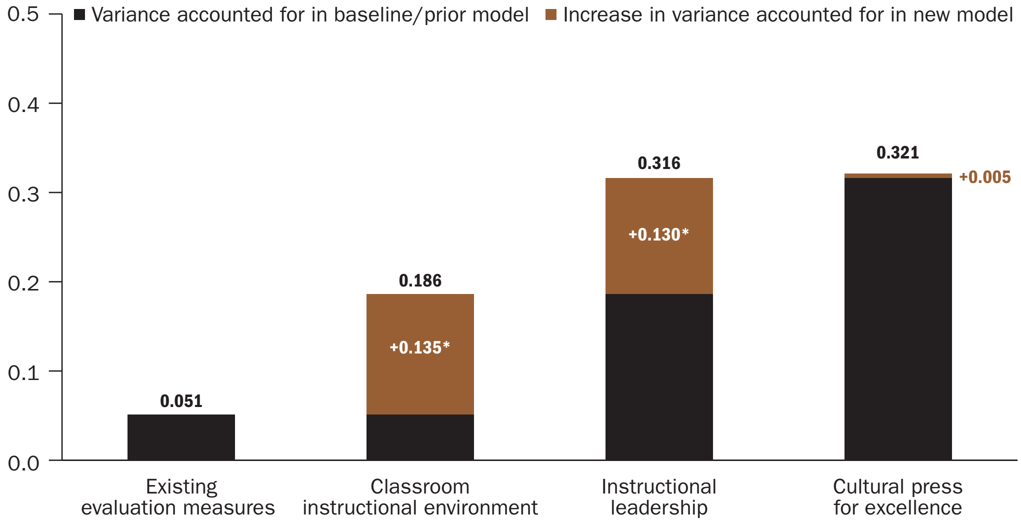
Figure 1. Improvements in the proportion of explained across-school variance in value-added achievement gains in math when adding three survey measures to the principal evaluation model



* Increase in explained variance is significant (p < 0.10).

Source: Authors' analysis based on data provided by the district.

Figure 2. Improvements in the proportion of explained across-school variance in value-added achievement gains in a composite of math and reading when adding three survey measures to the principal evaluation model



* Increase in explained variance is significant (p < 0.10).

Source: Authors' analysis based on data provided by the district.

For the math model, including the first two survey measures (classroom instructional environment and instructional leadership) increased explained variance by 28.8 percentage points, accounting for 73.6 percent of total variance explained by the full set of six survey measures (39.1 percentage points; see table 1).

For the composite of math and reading model, including the classroom instructional environment measure and the instructional leadership measure increased explained variance by 26.5 percentage points, accounting for 73.2 percent of total variance explained by the full set of six survey measures (36.2 percentage points; see table 1).

Limitations of the study

This study has four notable limitations, which are important for education policymakers to keep in mind as they consider the implications of the study findings for their districts or states.

First, the analysis was based on a sample of only 39 schools. This sample is comparable in size to those used in validity testing other principal evaluation measures (Goldring, Cravens, Murphy, Porter, & Elliott, 2012; Milanowski & Kimball, 2012).[3] However, a power analysis showed that the statistical test used to evaluate the candidate measures' incremental utility may not consistently detect measures that explain less than 11.7 percentage points of the across-school variance in value-added achievement gains (see table C3 in appendix C). Thus some candidate measures that were excluded from the optimal subset may in fact explain additional variation in school achievement gains, but the sample size did not yield enough statistical power to pick up their incremental effects.

The whole-school student and teacher feedback survey measures provide valuable information on school conditions through which principals can influence teaching and student learning

Second, the data used in the analysis are from one school district, a midsize urban district in the Midwest Region serving more than 18,000 students, more than 85 percent of whom are eligible for free or reduced-price lunch. Approximately 36 percent of the district's students are Black, 31 percent are Hispanic, 25 percent are White, 5 percent are multiracial, 2 percent are Asian, and 1 percent are American Indian. The findings do not necessarily generalize to other districts in the Midwest Region with different demographics, organizational structures, or student and professional cultures.

Third, this study examined a restricted set of principal performance measures from the district. Because the teacher and student feedback surveys may not represent those used in other districts, the findings may not apply directly to other districts. Findings also depend on what principal evaluation measures a district uses. In this study, the district's existing measures (principal supervisor ratings and school average attendance rates) explained low baseline amounts of the across-school variance in value-added achievement gains, which created more opportunity for the survey measures to demonstrate incremental value. The survey measures may not have the same incremental utility in a model that includes a stronger set of baseline nonsurvey measures.

Fourth, this study was not able to examine the incremental utility of subject-specific teacher and student survey measures. The teacher survey data identify the subjects taught by the surveyed teachers, but limiting student survey measures to math and reading classes would have entailed a large loss of student survey responses, especially in secondary schools. An examination of the relationship between subject-specific survey measures and school value-added achievement gains could find significant correlations. Despite this limitation, the whole-school student and teacher feedback survey measures provide valuable information on school conditions through which principals can influence teaching and student learning.

Appendix A. Literature review

An extensive body of research indicates that principals have strong effects on student achievement (Waters, Marzano, & McNulty, 2003). These effects are largely indirect and result from the ways that principals shape school conditions that promote effective teaching and learning (Hallinger & Heck, 1996; Heck & Hallinger, 2009; Leithwood, Louis, Anderson, & Wahlstrom, 2004; Sebastian & Allensworth, 2012). As Murphy, Elliot, Goldring, and Porter (2007, p. 181) note, "Leaders influence the factors that, in turn, influence the outcomes." To be comprehensive, principal evaluation models require technically sound measures of the school conditions that fall within principals' sphere of influence and that associate with improvement in student outcomes (Murphy et al., 2007).

This study examines the incremental utility of teacher and student feedback surveys in increasing the power of principal evaluation results to predict school value-added achievement gains beyond existing evaluation measures. The surveys consider six school conditions through which principals influence student achievement (Bryk et al., 2010; Wahlstrom, Louis, Leithwood, & Anderson, 2010).

The teacher perception survey measures four conditions: instructional leadership, professional learning community, professional development, and cultural press for excellence. These four conditions all promote specific behaviors and attitudes among teachers that shape the quality of their classroom instruction. Teachers are well positioned to assess the quality of these conditions because the conditions influence their daily work (Clifford, Menon, et al., 2012; National Center for Education Statistics, 1997).

The student perception survey measures the other two school conditions: school safety and climate and classroom instructional environment. These conditions promote behaviors and attitudes among students that lead to more productive learning (Carroll, 2006; Kane & Cantrell, 2010). As daily observers of their school and classroom environments, students offer an important perspective on the conditions that foster better student outcomes (Aleamoni, 1999; Clifford, Menon, et al., 2012; Worrell & Kuterbach, 2001). How principals influence these six organizational conditions to promote student achievement is described in the following literature.

Instructional leadership

Instructional leadership represents the leadership activities in the school that address instruction and curriculum (Hallinger, 2003). Instructional leadership extends beyond the role of the principal to include all leadership activities of the school's staff (Elmore, 2000; Spillane, Halverson, & Diamond, 2004). Strong instructional leaders are able to create and sustain a clear vision for learning, communicate school instructional goals, and garner schoolwide commitment to those goals (Leithwood & Riehl, 2003; Stronge, Ricard, & Catano, 2008). They also promote coherence in the instructional program by frequently visiting classrooms to monitor instruction and dialog with teachers (Cooper, Ehrensal, & Bromme, 2005; Leithwood & Riehl, 2003, Portin, Schneider, DeArmond, & Gundlach, 2003).

Numerous studies have documented the relationship between instructional leadership and student outcomes (Hattie, 2009; Robinson et al., 2008; Witziers et al., 2003). A meta-analysis of 22 leadership studies found that the average effects of instructional leadership practices on student achievement and other outcomes (such as absenteeism and engagement) were three to four times as large as the average effects of other leadership practices that do not explicitly focus on curriculum and instruction (Robinson et al., 2008).

Professional learning community

Principals also influence student achievement by promoting an effective professional learning community (Marzano, Waters, & McNulty, 2005; Valentine, Clark, Hackmann, & Petzko, 2004). Strong professional learning communities, characterized by teacher collaboration on instruction, can predict student achievement gains according to some studies (Goddard, Goddard, & Tschannen-Moran, 2007; Lomos, Hofman, & Bosker, 2011; Louis, Marks, & Kruse, 1996; Vescio, Ross, & Adams, 2008). For example, a longitudinal analysis found significant increases in average student achievement (effect sizes of 0.63, 0.64, and 0.88 in the final three years of implementation) across reading, math, language, and spelling in elementary schools where teacher teams collaborated on their instructional practices (Saunders, Goldenberg, & Gallimore, 2009).

Professional development

Principals also influence teachers' instructional quality through professional development (Newmann, King, & Youngs, 2000; Sebastian & Allensworth, 2012). Principals' involvement in professional development "provides them with a deep understanding of the conditions required to enable staff to make and sustain the changes required for improved outcomes" (Robinson et al., 2008, p. 667). A recent meta-analysis of six studies with 17 effect sizes identified promotion of and participation in teacher learning and professional development as the principal leadership dimension most strongly associated with positive student outcomes (average effect size of 0.84; Robinson et al., 2008).

Research demonstrates that effective professional development for teachers focuses on subject matter content and student learning, encourages the active involvement of teachers, and aligns with teacher knowledge and beliefs, as well as school, district, and state policies and reforms (Desimone, 2009; Garet, Porter, Desimone, Birman, & Yoon, 2001). It also promotes teacher collaboration. This type of professional development requires nimble principals able to gain access to resources and match development activities with school strategic goals (Portin et al., 2003). Even if principals have less influence over the quality of professional development as it relates to teacher knowledge and skills, they influence teacher access to professional development opportunities (Louis, Leithwood, Wahlstrom, & Anderson, 2010; Portin et al., 2009).

Within a district, the central office usually specifies general policies on the type, frequency, and duration of teacher professional development in accordance with the collective bargaining agreement with the local teachers union. District principals are responsible for helping plan and schedule, as well as participating in, formal sessions in which teachers review student achievement data, plan curriculum and lesson changes, discuss student needs, review student projects, and plan appropriate instruction to promote student learning. This professional development may be individual or collaborative and applies to both elementary and secondary schools.

Cultural press for excellence

Cultural press for excellence refers to the extent to which principals clearly and publicly articulate high standards of academic performance and rigorous learning goals for students, teachers, leadership, and staff at the individual, team, and school levels (Porter et al., 2008). Principals play an important role in shaping a culture of excellence within the school (Hallinger & Heck, 2002; Murphy et al., 2007). High expectations for all students and staff have been shown to associate with improvement in student achievement (Betts & Grogger, 2003; Newmann, 1998). For example, a study based on data from the sophomore cohort of the High School and Beyond survey found that a one standard deviation increase in the rigor of grading standards was associated with a 40 percent increase in the average rate of student progress in math between grades 10 and 12 (Betts & Grogger, 2003).

School safety and climate

A key responsibility of the principal is to ensure a safe and orderly school environment (Sebastian & Allensworth, 2012). This requires maintaining safe, clean, and visually attractive physical facilities (Murphy et al., 2007). It also requires ensuring that school and classroom rules for student behavior and disciplinary procedures are clearly defined and communicated to students, teachers, and parents (Marzano et al., 2005).

Studies measuring school safety using student perception surveys and student disciplinary records have found that unsafe and disorderly school environments are associated with lower student achievement results (American Institutes for Research, 2007; Barton, Coley, & Wenglinsky, 1998; Carroll, 2006; Ripski & Gregory, 2009). Exposure to violence and disorder in schools can negatively affect student performance in the classroom (Carrell & Hoekstra, 2011; Henrich, Schwab-Stone, Fanti, Jones, & Ruchkin, 2004). One study estimated that adding 1 additional disruptive student to a classroom of 20 was associated with a decrease in composite student achievement in math and reading of 1.5 percentage points among peers who are less inclined to behavioral problems (Carrel & Hoekstra, 2011).

Classroom instructional environment

Principals influence school classroom instructional environments through a number of channels. Effective principals directly influence the quality of the classroom instructional environment through the strategic hiring, development, and retention of good teachers (Béteille, Kalogrides, & Loeb, 2009). Although some districts limit principals' authority to hire their own staff (Bottoms & Schmidt-Davis, 2010), research suggests that good principals take a proactive stand in teacher recruitment (Brewer, 1993; Grissom & Loeb, 2009; Levine & Lezotte, 1990). Although, procedures and collective bargaining rules may limit principals' authority to replace tenured teachers, principals do have discretion in hiring new teachers. In one state more than 80 percent of principals indicated that they have major influence over the hiring of new full-time teachers in their schools (National Center for Education Statistics, 2008).

Principals also directly influence the classroom instructional environment when they connect with teachers in their classrooms during formal and informal observations (Leithwood et al., 2004; Portin et al., 2009). These interactions often lead to immediate changes in instructional practice as teachers respond to feedback (Hallinger & Heck, 1996; Kimball, Milanowski, & McKinney, 2007; Leithwood et al., 2004; Murphy et al., 2007).

A study using the Tripod student survey developed by Cambridge Education found that student perceptions of their classroom instructional environments were predictive of achievement gains in math and reading (Kane & Cantrell, 2010). Across the seven dimensions of classroom instructional quality measured, correlations ranged from 0.31 to 0.49 in math and from 0.01 to 0.32 in English language arts. Students' perceptions that their teachers "clarified" difficult academic content (0.49) and "challenged" them to give their best effort (0.44) were most strongly associated with achievement gains in math, and their perceptions that their classrooms "challenged" them (0.32) and "controlled" student behavior (0.29) were most strongly associated with achievement gains in English language arts.

Although research shows that the six school conditions described here are associated with effective school leadership and improvement in student achievement, it offers little guidance on how to measure these conditions or on how principals influence these conditions for the purpose of principal evaluations (Clifford, Menon, et al., 2012). This study was designed to help fill this gap by investigating the degree to which adding teacher and student survey measures to the existing set of principal evaluation measures can increase the power of principal combined evaluation results to predict school average achievement gains. This information will help states and districts understand the utility of these surveys and make informed decisions about whether and how to include student and teacher surveys in their principal evaluation models (Clifford, Menon, et al., 2012).

Appendix B. Data and methodology

This appendix describes the data and methodology used in the study.

Data

A midsize urban district in the Midwest Region provided the study team with districtwide data files for this study. All the districtwide data files are for school year 2011/12, and the data are disaggregated at the student, teacher, and principal levels. To answer the two research questions, all the data sources were used to create three sets of variables: student outcome measures; existing principal evaluation measures, and teacher and student survey measures (candidate measures).

Student outcome measures. The outcome measures were school value-added achievement gains in math, reading, and a composite of math and reading. The value-added outcome measures were calculated using student-level reading and math test scores from the fall 2011 and spring 2012 Northwest Evaluation Association's (2011) Measures of Academic Progress (MAP) assessments. Additional variables were individual student characteristics (race/ethnicity, gender, English language learner status, special education status, and the like). These variables were drawn from the district's administrative data file on student background characteristics.

The district's MAP test data from school year 2011/12 included 8,246 students in grades 3–11 with valid pretest and posttest scores from 51 schools, which was 74.1 percent of the district's total enrollment in grades 3–11. These students were included to estimate the three outcome measures of school value-added achievement gains. For precision, 6 schools with fewer than 10 students in each tested grade were dropped from the sample. The final school-level value-added data file included 45 schools with three outcome measures.

Existing principal evaluation measures. The existing principal evaluation measures, used to establish the baseline validity of the principal evaluation model to predict school value-added achievement gains in math and reading, include two sets of measures that are mandated by state law: principal supervisor ratings and school average attendance rates. Results from the district's evaluation instrument were used to construct two measures of principal job function competency and leadership skill competency. Scores on these two measures were then averaged to obtain a composite supervisor rating for each principal. Student attendance records were aggregated at the school level to calculate school average attendance rates.

The district's evaluation instrument consists of two rubrics in 2011/12. The first rubric assesses principals on seven core competencies related to job functions, and the second rubric measures four core competencies related to leadership skills. Principals received a rating of 1–4 on each competency (1 = ineffective, 2 = minimally effective, 3 = effective, and 4 = highly effective). A composite supervisor rating was constructed for each principal by first averaging the ratings on core competencies within each rubric and then averaging the two rubric scores. To calculate each school's attendance rate, the actual number of attendance days and the number of possible attendance days for all students in the school were first separately aggregated. The school's attendance rate was then calculated as the aggregate actual attendance days divided by the aggregate possible attendance days.

Attendance records from 17,623 students in grades pre-K–12 were used to calculate school average attendance rate for the same 51 schools that were used for the value-added analysis. And the district provided principal evaluation data for the same 51 schools.

Teacher and student survey measures. The candidate measures are six survey measures from teacher and student feedback surveys. The district used the teacher and student Tripod surveys developed by Ron Ferguson of Harvard University's Kennedy School of Government and Cambridge Education. School year 2011/12 was the first time the district administered the student Tripod survey and the second time it administered the teacher Tripod survey. Four candidate measures were extracted from the teacher survey: instructional leadership, professional learning community, quality of professional development, and cultural press for excellence. Two candidate measures were taken from the student survey: classroom instructional environment and school safety and climate. These candidate measures were entered into the principal evaluation model in addition to the two existing measures to predict school value-added achievement gains.

The student Tripod survey data include 8,601 students in grades 3–12 from 47 schools. The teacher Tripod survey data include 581 teachers with valid survey responses from 53 schools. All individual survey responses were aggregated at the school level to create the six school-level candidate measures that reflect school organizational conditions: instructional leadership, quality of professional learning community, quality of professional development, cultural press for excellence, school safety and climate, and classroom instructional environment.

Sampling

The research questions were answered using a sample of 39 schools from the participating district, including 20 elementary schools (grades K–5), 13 secondary schools (grades 6–12), and 6 schools with grades spanning both ranges. The schools were selected because they had nonmissing data on all three sets of variables: the student outcome measures, the existing principal evaluation measures, and the teacher and student survey measures. Limiting the sample to the 39 schools with nonmissing data ensured that the comparison of predictive power of different models was based on the same sample.

The value-added analysis includes 7,709 students in grades 3–11 (table B1). The mean school size for the value-added analysis is 198 students, with a minimum of 31 and a maximum of 568. The school average attendance rate is based on the attendance records of 16,537 students in grades K–12. The student Tripod survey data include responses from 8,345 students in grades 3–12, with a mean sample size of 214 students per school, a minimum of 11 students, and a maximum of 733 students.[4] The teacher Tripod survey data include responses from 541 teachers, with a mean sample of 14 teachers per school. Finally, each school had only one principal, and thus only one set of supervisor ratings.

On the whole, the student racial/ethnic composition and English language learner student population in the final sample of 39 schools were very similar to those in the full sample of 57 schools (table B2). The 39 schools in the final sample had lower percentages of students eligible for special education programs and larger enrollments than the schools excluded from the analysis did. This result was expected because schools with small enrollments (fewer than 10 students in each tested grade) were dropped from the value-added analysis,

Table B1. Sample sizes by measure for 39 schools in the study sample

Measure	Total number of observations	Observations per school		
		Average	Minimum	Maximum
School value-added gains	7,709 students	198 students	31 students	568 students
School attendance rate	16,537 students	424 students	62 students	998 students
Student feedback survey	8,345 students	214 students	11 students	733 students
Teacher feedback survey	541 teachers	14 teachers	2 teachers	35 teachers
Supervisor ratings of principals	39 principals	1 principal	1 principal	1 principal
School value-added gains	7,709 students	198 students	31 students	568 students

Source: Authors' analysis based on data provided by the district.

Table B2. Averages for characteristics of sampled schools

Sample of schools (number of schools)	Black students (percent)	Hispanic students (percent)	White students (percent)	English language learner students (percent)	Special education students (percent)	School enrollment (number of students)
Schools with valid value-added data (45)	38.5	27.0	25.9	23.0	20.8	442
Schools with valid student survey data (47)	38.4	29.6	23.9	25.4	19.4	416
Schools with valid teacher survey data (52)	39.1	28.4	24.5	24.1	20.1	411
Schools with data from at least one source (57)	38.6	26.8	26.6	22.9	22.1	380
Final sample (39)	38.5	29.8	23.2	25.3	17.9	485

Source: Authors' analysis based on data provided by the district.

and schools with a large special-education population tend to have fewer valid responses to the student surveys. A set of *t*-tests suggests that the final sample of 39 schools differs from the other 18 schools significantly in only two school characteristics: percentage of special education students and enrollment size.

Candidate principal evaluation measures

Six multiscale variables derived from the teacher and student Tripod surveys were considered candidate measures for the principal evaluation model. The six survey measures were also entered into the regression models together to examine the joint incremental utility of all measures as a set. Then the joint significance of incremental utility from the two subsets of the six candidate measures—the subset of the four teacher survey measures and the subset of the two student survey measures—were examined. Finally, the six survey measures were separately entered into the regression models to examine the incremental utility of each individual candidate measure.

Teacher survey measures

All survey items were standardized to have a mean of 0 and standard deviation of 1. Individual survey items were first averaged to form the four measures of school organizational conditions at the teacher level and then each measure was aggregated at the school level.

Instructional leadership. Measured by 18 items on teachers' perception of the expertise of school instructional leaders in promoting a climate of learning, managing instruction, and defining the school mission.

Professional learning community. Measured by 12 items related to the amount of time spent in professional learning community activities and in collaboration with teachers on curriculum design and assessment.

Quality of professional development. Measured by 14 items on teachers' perception of the effectiveness of professional development activities and the support they receive from school leadership in their professional development.

Cultural press for excellence. Measured by 3 items on the school culture of holding adults accountable for excellence and setting and achieving important goals.

Student survey measures

Classroom instructional environment. A single measure that reflects students' perceptions of their classroom instructional environment in seven domains (Ferguson, 2011; Kane & Cantrell, 2010): caring about students, controlling behavior, clarifying lessons, challenging students, captivating students, conferring with students, and consolidating knowledge. The indices of the seven domains include 36 survey items; the number of survey items composing each domain ranges from three (for caring about students) to eight (for captivating students). The indices of the seven domains were first created for each classroom and then summed into a single measure of classroom instructional environment. Although students provided feedback on a specific course, the classroom instructional quality measures were not aggregated to the school level by subject.[5] Instead, a single school-level measure of classroom instructional environment was constructed and used as a candidate survey measure in later analysis.

Six steps were applied in calculating the single school-level measure of classroom instructional environment:

1. Aggregating all student survey items to the classroom level to create a classroom raw score.
2. Standardizing the classroom item mean score to create the *z*-value for each item at the classroom level.
3. Averaging the standardized item scores for each classroom within each of the seven domains.
4. Standardizing each of the seven domain scores at the classroom level so that each domain would contribute the same information to the composite measure of the classroom instructional environment.
5. Creating a composite classroom instructional quality measure by averaging the standardized domain scores for each class.
6. Aggregating the classroom composite measure to the school level to form the school-level measure of classroom instructional environment.

The sample mean and standard deviation for the fourth step (classroom-level domain score standardization) were based on a national sample provided by Cambridge Education and currently used by the study's district.

School safety and climate. A single measure that reflects students' perceptions of their school safety and climate constructed from six items. This measure was aggregated at the school level using the same procedure as the classroom instructional environment measure.

The value-added model

A two-step covariate adjustment value-added model (McCaffrey et al., 2003) was used to measure school performance in the district. This model was selected in part because of its prevalence among large school districts in the Midwest Region and elsewhere. Covariate adjustment models are used in the Chicago Public Schools, Madison Metropolitan School District, and Milwaukee Public Schools.[6] They also are used in the District of Columbia Public Schools' IMPACT evaluation system, the New York City educator evaluation system, and Florida's state evaluation model. Another reason for using the covariate adjustment model is that it is easy to specify and does not require proprietary software, so districts can replicate this analysis using standard statistical computing packages. Finally, the covariate adjustment model was appropriate for this analysis because it does not require test scores to be linked across grades with a vertical scale, making it well suited to districts that use both norm-referenced and state tests.

The value-added estimates of school performances are based on students' MAP assessment scores in math and reading in grades 3–11. The analysis calculated within–school year achievement growth based on the test score difference between the fall and spring tests.

Estimation equations. The first step predicts student spring test scores as a function of prior performance in the fall test in math and reading, student characteristics, and school characteristics. The first-stage model is:

$$Y_{i,g,spring} = \beta_0 + \beta_1 Y^{same}_{i,g,fall} + \beta Y^{oth}_{i,g,fall} + \beta'_3 X_{i,g} + \beta'_4 \overline{X}_{i,g} + \varepsilon_{i,g}, \qquad \text{(B1)}$$

where $Y_{i,g,spring}$ is the posttest score for student i in the spring test in grade g. $Y^{same}_{i,g,fall}$ is the same subject pretest score for student i in the fall test, and $Y^{oth}_{i,g,fall}$ is the pretest score in the other subject in the fall test. For example, if the dependent variable refers to student posttest achievement in math ($Y^{math}_{i,g,spring}$), $Y^{same}_{i,g,fall}$ denotes the math pretest score in the fall test and $Y^{oth}_{i,g,fall}$ denotes the reading pretest score in the fall test. All test scores are converted into a common metric (or *z*-score) with a sample mean of 0 and standard deviation of 1 to address possible across-year or grade differences in score scaling. The vector $X_{i,g}$ represents the covariates for individual characteristics, and $\overline{X}_{i,g}$ denotes a vector of school averages of individual pretest scores in both subjects as well as other demographic characteristics. The last term, $\varepsilon_{i,g}$, is an individual residual. The covariates in the first-stage model are described in table B3.

The first-stage model (equation B1) was run separately by subject and grade and forecasts student current performance ($\hat{Y}_{i,g,spring}$) from all available information about the students and schools. The estimated residual $\hat{\varepsilon}_{i,g}$ represents the deviation of the actual performance $Y_{i,g,spring}$ from the predicted performance $\hat{Y}_{i,g,spring}$. For example, a positive $\hat{\varepsilon}_{i,g}$ means that student *i* outperforms his or her forecast, or student *i* performs better than other same-grade students who have similar fall pretest scores and similar individual and school characteristics. A negative $\hat{\varepsilon}_{i,g}$ means that student *i* performs worse than the forecast.

Table B3. Covariates included in the first-stage value-added model

Covariates	Description
$Y^{same}_{i,g,fall}$	Same subject pretest score in the fall
$Y^{oth}_{i,g,fall}$	Other subject pretest score in the fall
$X_{i,g}$	Racial/ethnic indicators: a set of dichotomous variables for racial/ethnic groups, including Asian, Black, Hispanic, Native American, White, and other races.[a] Gender indicator: a dichotomous variable for female or male that equals 1 if the student is female and 0 if the student is male. English language learner: a dichotomous variable that equals 1 if the student is identified as an English language learner and 0 otherwise. Special education status: a dichotomous variable that equals 1 if the student is in a special education program and 0 otherwise. School mobility: a dichotomous variable that equals 1 if the student is new to the school in 2011/12 for reasons other than normal grade promotion and 0 otherwise.
$\overline{X}_{i,g}$	Aggregated school-level student prior test scores in both subjects as well as background characteristics; for example, percentage of students who are female and percentage of students who receive special education services

a. The racial/ethnic terms from the original source are used.

Source: Authors' compilation.

The second step regresses $\hat{\varepsilon}_{i,g}$ on a vector of dummy indicators of school enrollments to estimate a school's contribution to student performance:

$$\hat{\varepsilon}_{i,g} = \theta' S_{i,g} + u_{i,g}, \tag{B2}$$

where the school enrollment indicators are included in the vector $S_{i,g}$: if student *i* enrolled in school *s* in the current year, $S_{i,g}$ equals 1; otherwise, $S_{i,g}$ equals 0. If student *i* was not enrolled in school *s* for the full school year, the school indicator $S_{i,g}$ equals 1 but is assigned a weight that equals the proportion of school days that student *i* stayed in school *s*:

$$\frac{enrollmentdays_{i,s,g}}{totalschooldays}.$$

In this study, the district did not provide enrollment data, and each student was linked to one school where he or she took the posttest. Therefore, each student received an exposure weight of 1 in the regression. If enrollment data are available, the aggregated school variables ($\overline{X}_{i,g}$) for students who enrolled in multiple schools is an average of school characteristics of the enrolled schools weighted by the enrollment days. Like the first-stage model, the second-stage model was run by subject and grade. The school-level coefficients, captured in the vector $\theta_{s,g}$ represent the average deviations from the forecasts of equation B1 for all students in a given school and grade. For example, if students in one grade in school *j* systematically outperform their forecasts, school *j* receives a larger estimate of $\hat{\theta}_{j,g}$ than other schools whose students in the same grade perform close to or below expectations.

Measurement error. Standard tests are not perfect measures of students' true ability. Measurement error accounts for a sizeable portion of test score variability, which tends to cause ordinary least squares regressions to produce biased value-added estimates of teacher or school effectiveness (Hanushek & Rivkin, 2010). Thus, this study used a statistical approach called errors-in-variables regression to control for measurement error in pretest scores (see, for example, Isenberg & Hock, 2011, 2012; Value-Added Research Center, 2010). Specifically, the errors-in-variables regression approach divides the ordinary least squares regression estimator by the reliability ratio of the regressor of interest (that is, the pretest scores). The reliability ratio of the regressor is provided by the test vendor and represents the ratio of variance of the true explanatory variable to the total ratio of the measured variable (the pretest scores). Through this procedure, the errors-in-variables regression approach adjusts the coefficient on pretest scores upward by the size of average measurement error variance of the test population in each grade and thereby produces a consistent estimator of the true coefficient (Greene, 2003).

Shrinkage estimates. The empirical Bayes shrinkage procedure was employed to reduce the instability of value-added estimates that is often caused by small sample sizes. The shrinkage technique views the estimate of an individual school effect as an optimal combination of two sources: the estimated effect of the school ($\hat{\theta}_s$)[7] and the average estimate of all schools evaluated ($\overline{\theta}$). The weight placed on each effect depends on the amount of information used to estimate $\hat{\theta}_s$. If $\hat{\theta}_s$ is precisely estimated from a large number of students taught in school *s*, the weight on $\hat{\theta}_s$ will be large, and the shrinkage estimate of effectiveness for school *s* will not be very different from the actual $\hat{\theta}_s$. Conversely, if school *s* has only a few students, the weight on $\hat{\theta}_s$ will be small, and the shrinkage estimate will drop toward the sample mean $\overline{\theta}$. Specifically, the shrinkage estimates are expressed as

$$\hat{\theta}_{s,shrinkage} = \lambda_s \hat{\theta}_s + (1 - \lambda_s)\overline{\theta}. \qquad \text{(B3)}$$

After standardization, the overall mean of school effects $\overline{\theta}$ is centered on zero, so the second term in equation B3 would disappear, and equation B3 can be written as

$$\hat{\theta}_{s,shrinkage} = \lambda_s \hat{\theta}_s, \qquad \text{(B4)}$$

where λ_s is the weight or the reliability of $\hat{\theta}_s$:

$$\lambda_s = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{\sigma}_s^2}, \qquad \text{(B5)}$$

where $\hat{\sigma}^2$ is the true variance of all school value-added estimates and is constant for all schools in the sample and $\hat{\sigma}_s^2$ is the squared standard error of $\hat{\theta}_s$. When the estimate of $\hat{\theta}_s$ is precise (based on a large sample of students), the standard error $\hat{\sigma}_s$ is small, so $\lambda_s = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{\sigma}_s^2}$ is close to 1 and the postshrinkage estimate is close to the original estimate. On the contrary, if the estimate of $\hat{\theta}_s$ is based on a small number of students, the standard error $\hat{\sigma}_s$ will be large, and $\lambda_s = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{\sigma}_s^2}$ will be close to 0; consequently, there is substantial shrinkage, and $\hat{\theta}_{s,shrinkage}$ drops toward the sample mean.

The standard errors for the shrinkage estimates can be computed by taking the square root of the product of λ_s and $\hat{\sigma}_s^2$:

$$\hat{\theta}_{s,shrinkage} = \sqrt{\lambda_s * \hat{\sigma}_s^2}. \qquad \text{(B6)}$$

Composite value-added school scores across subjects and grades. The two-stage value-added model produces a set of grade-specific estimates by subject for each school. To generate a school-level, subject-specific value-added estimate, it is necessary to combine the grade-specific value-added estimates by subject into one composite measure.[8]

Thus, the next step is to combine these grade-specific estimates across grades into a single effect for each school. Because the average and variability of errors-in-variables regression estimates differ across grades, it is necessary to standardize the grade-specific estimates within each grade. For grade *g* in school *s*, the standardized errors-in-variables regression estimate is

$$\hat{\sigma}_{s,g}^{standardized} = \frac{\hat{\theta}_{s,g} - \overline{\theta}_{g}}{\hat{\sigma}_{g}}. \tag{B7}$$

To simplify the illustration, the analysis omits the notations of subject-composite and shrinkage from equation B7, but $\hat{\theta}_{s,g}$ refers to the postshrinkage subject-specific value-added estimate for grade *g* in school *s*, and $\hat{\sigma}_{s,g}^{standardized}$ is the standardized value of $\hat{\theta}_{s,g}$. $\overline{\theta}_{s,g}$ is the average errors-in-variables regression estimate in grade *g* across all schools, and $\hat{\sigma}_{s}$ is the standard deviation of errors-in-variables regression estimates in grade *g*. The analysis then calculates a weighted average of errors-in-variables regression estimates across the grades in each given school, with the weight equal to the proportion of students in the school who enrolled in grade *g* (denoted by $p_{s,g}$)[9]:

$$\hat{\theta}_{s} = \sum \hat{\theta}_{s,g}^{standardized} * p_{s,g}, \tag{B8}$$

where $\hat{\theta}_{s}$ is the postshrinkage single estimate of school performance combing the estimates across subjects and grades in school *s*. Assume that the covariance of errors-in-variables regression estimates across grades is 0, and then obtain the variance of this combined school estimate as the following:

$$Var(\hat{\theta}_{s}) = \sum Var(\hat{\theta}_{s,g}) * p_{s,g}^{2}. \tag{B9}$$

Regression analysis of principal evaluation models

Before conducting the regression analysis to examine the incremental utility of the six candidate measures, the correlations between the school value-added outcome measures and the existing and candidate principal evaluation measures were computed. This analysis was conducted to examine the function of each principal evaluation measure in the regression model as well as the contribution of each measure to the overall variance in the outcome variables. The results of the correlation analysis are reported in table C2 in appendix C.

The two research questions were addressed with a two-step regression analysis. The first step assessed the incremental utility of the candidate measures above the existing measures in explaining across-school variance in value-added achievement gains in math and reading and a composite of math and reading. The second step examined whether the six candidate measures could be reduced to an optimal subset of measures that make significant incremental contributions to strengthen the relationship between the principal evaluation models and the three school value-added outcomes.

Step 1: Estimating the incremental utility of the six survey measures

The first step in the analysis addresses research question 1: Does adding the teacher and student feedback survey measures to an existing set of principal performance measures improve the power of the principal evaluation model to explain variance in across-school value-added achievement gains?

The basic analytic strategy used to answer research question 1 is to test whether a regression model that includes the candidate survey measures as independent variables in addition to the existing evaluation measures explained more across-school variance in value-added achievement gains than the baseline regression model that includes only the existing measures.

First, the study team tested whether adding the full set of six candidate measures to the existing measures led to a statistically significant increase in explained variance (R^2). An F-test was conducted to compare the R^2 of the model that added the six candidate measures to that of the baseline model. A p-value of less than 0.10 was required to reject the null hypothesis that there is no significant difference in the explained variance between the two models.[10] A significant difference in the R^2 between the two models would be evidence that adding the full set of six candidate measures strengthens the relationship between the principal evaluation model and school value-added achievement gains. Because the full set of survey measures is centered on six school conditions through which principals can influence student learning, this evidence would send an important message to the district that the surveys jointly contribute new information on the link between principal practice and student achievement.

Next, the joint significance of two subsets of the candidate measures—the subset of the four teacher survey measures and the subset of the two student survey measures—was tested. Finally, the six candidate measures were entered into the regression models separately to examine the individual incremental utility of each survey measure.

For each school value-added achievement gain outcome, the study team examined the incremental utility of nine sets of candidate measures: the full set of all six survey measures, a subset of teacher survey measures, a subset of student survey measures, and six sets of individual survey measures. Twenty-seven F-tests (nine sets of candidate measures and three outcomes) were conducted to compare the R^2 of the subsequent regression model that included the additional survey measure or measures with that of the baseline regression model (see table 1 in the main text). For each test, rejecting the null hypothesis provided evidence that adding the candidate measure or measures significantly improved the power of the principal evaluation model to explain the across-school variance in average value-added achievement gains. All regressions were based on the same sample of 39 schools to ensure that reported differences in the R^2 between the models cannot be attributed to differences in samples.

Step 2: Determining an optimal subset of survey measures

The second step in the analysis addresses research question 2: Can the full set of six survey measures be reduced to an optimal subset of measures that make significant incremental contributions to the link between principal evaluation measures and school value-added achievement gains?

To answer this question, the study team entered candidate measures sequentially into the regression model as suggested by their estimated incremental utility in the first-step analysis. Only candidate measures found to significantly increase the explained across-school variance in value-added achievement gains at the statistical level of p-value < 0.10 were included in the second-step analysis. The selected candidate measures were entered sequentially into the regression model in descending order of incremental utility from the first-step analysis. For example, if the classroom instructional environment measure was found to have the largest incremental utility in the math model from the first-step analysis, it was the first measure added into the baseline model for math. After each candidate measure was entered, an F-test was conducted to determine the significance of its incremental utility. The optimal subset of candidate measures was considered attained when the entry of the next candidate measure failed to make a significant incremental contribution to explain the variance in the outcome measure.

Detailed results of regression analyses are provided in appendix D.

Appendix C. Supplemental analysis

This appendix reports the supplemental analyses conducted and the results.

Technical quality of the district's principal evaluation measures

Like many districts in the United States, the district in this study uses a homemade principal evaluation instrument with unknown psychometric properties (Porter et al., 2008). The study team examined the technical quality of the district's principal evaluation instrument in terms of its reliability and validity.

With regard to reliability, three measures from the district's principal evaluation instrument —the job function rating, the leadership skills rating, and the composite supervisor rating of all competencies—were examined. Internal consistency values (Cronbach's alpha) were calculated on the ratings of each rubric's competencies and the instrument's 11 competencies as whole. Pairwise deletion was used for missing data. The validity of the principal evaluation instrument was based on its relationship to school value-added achievement gains (Milanowski & Kimball, 2012). The correlations between the principal evaluation measures and the three school value-added achievement growth measures were also examined.

On the principal evaluation measures, the reliability coefficients are 0.61 for the job function rating, 0.54 for the leadership skill rating, and 0.73 for the composite supervisor rating (table C1). The reliability of the composite supervisor rating exceeds the conventional 0.70 minimum threshold for internal consistency based on Cronbach's alpha (Nunnally, 1978).

The correlations between the three principal evaluation measures and school value-added achievement growth range from 0.05 to 0.27, which are very similar to the correlations of 0.1–0.3 found in studies by Waters et al. (2003) and Milanowski and Kimball (2012). Two of the three correlations between the principal evaluation measures and school value-added achievement gains in math and in the composite of math and reading meet the practical validity threshold of 0.15 suggested by Milanowski and Kimball.

There are three possible reasons for the low correlations between principal supervisor evaluation measures and school value-added achievement gains. First, principals tend to influence student achievement indirectly through other school factors such as teachers, communities, and school climates (see, for example, Hallinger & Heck, 1996; Milanowski & Kimball, 2012), which is likely to result in a small association between the principal evaluation measures and school value-added gains, even if the instruments used to evaluate them have strong psychometric properties. Second, the variation in principal ratings is

Table C1. Technical quality of the principal evaluation instrument

	Reliability: Cronbach s alpha internal consistency	Validity: correlations with school value added achievement gains		
Measure		Math	Reading	Composite
Principal job function rating	0.61	0.266	0.049	0.178
Principal leadership skill rating	0.54	0.121	0.068	0.105
Composite supervisor rating	0.73	0.215	0.067	0.158

Source: Authors' analysis based on data provided by the district.

small: 77 percent of the study's principals were rated in a range between 2.5 and 3.5.[11] This makes it less likely to find large correlations between principal evaluation measures and school value-added achievement gains. Third, the study sample comprises only 39 schools and therefore may not yield sufficient statistical power to detect large and significant correlations between principal evaluation measures and school value-added achievement gains.

Correlations between principal evaluation measures and school outcome measures

Before conducting the regression analysis to examine the incremental utility of the six candidate measures, the study team first computed the correlations between the three value-added outcome measures and the existing and candidate principal evaluation measures. This correlation analysis was conducted to examine the function of each principal evaluation measure in the regression model as well as the contribution of each measure to the overall variance in the outcome variables.

The results show that 25 of 27 measures were positively correlated with school value-added performance (table C2). One teacher survey measure, quality of professional development, was negatively correlated with school value-added achievement gains in math and a composite of math and reading, although the magnitude of the correlation was almost zero. The positive correlations between the principal evaluation measures and school value-added outcomes suggest that both the existing measures and candidate measures of the principal evaluation model possess certain validity to explain the variance in school performances.

None of the correlations between the principal evaluation measures and school value-added outcomes exceeds 0.3; the magnitudes of correlations all fall within the weak correlation category of Cohen's (1988) guideline. Because of the small sample size, the correlations are not statistically significant.[12]

Table C2. Correlations of principal evaluation measures with school value-added achievement gains

	School value added achievement gains		
Measure	**Math**	**Reading**	**Composite**
Existing measures			
Principal job function rating	0.266	0.049	0.178
Principal leadership skill rating	0.121	0.068	0.105
School attendance rate	0.103	0.145	0.135
Candidate measures			
Teacher survey			
A. Instructional leadership	0.255	0.228	0.266
B. Professional learning community	0.157	0.166	0.177
C. Quality of professional development	–0.047	0.038	–0.007
D. Cultural press for excellence	0.255	0.221	0.262
Student survey			
E. School safety and climate	0.221	0.081	0.168
F. Classroom instructional environment	0.289	0.232	0.277

Source: Authors' analysis based on data provided by the district.

One reason for the low correlations may be the fact that principals' influence on student achievement is indirect. Waters et al. (2003) found that the correlations between various principal behaviors and student achievement ranged from 0.16 to 0.33. Similarly, Milanowski and Kimball (2012) more recently found correlations below 0.3 between standards-based principal performance evaluation ratings (which are similar to the supervisor ratings in this study) and school value-added achievement growth in math and reading from various samples. According to Milanowski and Kimball (2012), the largely indirect impact of principal behaviors on student learning poses a major challenge in detecting substitute correlations between principal behaviors and student performance.

Another possible reason for the low correlations is attenuation bias due to measurement error in the principal evaluation measures. The potential for attenuation bias in the value-added measures was addressed by using the errors-in-variables regression method in the estimation of school value-added achievement gains. To disattenuate the correlations of measurement error, the study team also divided the correlation coefficients by the square root of the reliability coefficients for the principal evaluation measures (Spearman, 1904). After the disattenuation adjustment, the correlation between the principal job function measure and school value-added achievement gains in math is 0.33, exceeding Cohen's moderate correlation cutoff. The disattenuated correlation between the classroom instructional environment measure and school value-added gains in math was 0.304; the disattenuated correlation coefficients on other principal evaluation measures were below 0.30. All Cronbach's alpha reliability coefficients were obtained from the internal consistency analysis based on the district's evaluation or survey data.

The reliability coefficients are 0.61 for the job function rating, 0.54 for the leadership rating, 0.90 for the instructional leadership measure, 0.87 for the quality of professional community measure, 0.84 for the quality of professional development measure, 0.89 for the cultural press for excellence measure, 0.80 for the school safety measure, and 0.90 for the classroom instructional environment measure.

Power analysis

A power analysis was conducted to understand the implications of the study's sample size of 39 schools on the results of the statistical tests used to evaluate the candidate measures' incremental utility. Statistical power is important because it determines the degree to which a sample size of 39 will consistently identify candidate measures as statistically significant if their "true" incremental effect on school value-added outcomes is different from 0.

As described in the main report, the statistical test used to evaluate incremental utility was an incremental *F*-test (Greene, 2011).[13] The *F*-test compares the variance in school value-added achievement gains that is explained by the two existing principal performance measures in the baseline model (R^2) to the R^2 of a series of unrestricted models that include between one and six candidate measures along with the two existing measures. The null hypothesis of the *F*-test is that the increase in R^2 associated with the additional measure or measures in the unrestricted model is equal to 0. The question of interest to the power analysis is: With a sample size of 39, what is the smallest increase in explained variance that must be observed to correctly reject the null hypothesis with a high probability?

To answer this question, the study team estimated the minimum increases in R^2 that will correctly reject the null hypothesis 80 percent of the time (power = 0.80, Cronbach's alpha = 0.10). These values were calculated for the four specifications of the number of candidate measures included in the unrestricted models run in step 1: one candidate measure (individual student or teacher survey measures), two candidate measures (the set of student survey measures), four candidate measures (the set of teacher survey measures), and six candidate measures (the full set of student and teacher measures).

The findings of the power analysis show that, according to the standards proposed by Cohen (1988), the sample is sufficient for detecting moderate increases in explained variance, but more observations are required to detect smaller changes in variance with 80 percent probability (table C3). The minimum increase in R^2 ranged from 0.117 for a test of the significance of a single candidate measure to 0.210 for a test of the joint significance of all six candidate measures. Statistically significant effects for changes in R^2 that are smaller than the values for the minimum detectable effects in table C3 were also found. The values in table C3 are estimates of the "true" incremental effects in the population that can be observed when statistical power of 0.80 is maintained, whereas the statistical significance of the observed effects is dictated by the actual power derived from this study sample.

Table C3. Minimum detectable increases in explained variance for sample size of 39 schools (power of 0.80, Cronbach's *alpha* = 0.10)

Number of existing measures included in baseline model	Number of candidate measures included in unrestricted model	Minimum increase in R^2
2	1	0.117
2	2	0.144
2	4	0.180
2	6	0.210

Source: Authors' analysis based on data provided by the district.

Appendix D. Descriptive statistics and regression coefficients for the principal evaluation measures

All three sets of measures (outcome measures, existing measures, and candidate measures) were used to answer the two research questions.

Descriptive statistics for the principal evaluation measures

The descriptive statistics of all variables used in the analysis from the district's 39 schools in the study are shown in table D1. The mean values of the three outcome measures approximate zero because all individual test scores were standardized before being used for value-added estimation. For the 39 principals the average supervisor rating was 3 (on a scale of 1–4) in both rubrics. The lowest rating in the leadership skills rubric was 2, suggesting that no principal is rated at the bottom of the scale (similar to the findings of Milanowski and Kimball, 2012, in two relatively large districts). School attendance rates ranged from 0.538[14] to 0.997, with a sample mean of 0.906. All six candidate measures are constructed from multiple survey items using a scale of 1–5. Although the descriptive statistics in table D1 are based on the natural scales of these survey measures, all items from both surveys were standardized into a common metric before being aggregated at the school level for the analysis. The means of the six survey measures are all above 3. The standard deviations of the two student survey measures are smaller than the standard deviations of the four teacher survey measures, suggesting smaller between-school variation in the two student survey measures. Among the six measures, principals and their schools generally scored lower in the measures of professional learning community (3.11) and quality of professional development (3.14).

Table D1. Descriptive statistics for the principal evaluation measures

Measure	Mean	Standard deviation	Minimum	Maximum
Outcome measure (criteria)				
School value-added math	–0.116	0.708	–1.922	1.153
School value-added reading	–0.024	0.790	–1.856	1.249
School value-added composite	–0.070	0.679	–1.759	1.162
Existing measure				
Principal job function rating	3.040	0.427	2.286	3.714
Principal leadership skill rating	3.026	0.489	2.000	3.750
School attendance rate	0.906	0.075	0.538	0.997
Candidate measure				
Teacher survey				
Instructional leadership	3.814	0.506	2.426	4.717
Professional learning community	3.114	0.382	2.292	3.980
Quality of professional development	3.138	0.365	1.725	3.758
Cultural press for excellence	3.950	0.485	2.556	4.833
Student survey				
School safety and climate	3.682	0.280	3.078	4.503
Classroom instructional environment	3.722	0.284	3.081	4.232

Note: All variables reported are school-level aggregates given that the unit of analysis in this study is the school/principal.

Source: Authors' analysis based on data provided by the district.

Regression coefficients on the principal evaluation measures

The regression coefficients on the principal evaluation measures (two existing measures and six candidate measures) from the nine regression models that examined the incremental utility of survey measure (or measures) in explaining variance in school value-added achievement gains in math, reading, and composite of math and reading are reported in tables D2–D4. In all three tables, columns 1–4 report the coefficients from the models that added individual teacher survey measures separately into the baseline model containing the two existing measures; column 5 reports the coefficients from the model where the subset of four teacher survey measures were controlled together as a group; columns 6 and 7 report the coefficients from the models where the two student survey measures were examined separately; column 8 reports the coefficients on the subset of two student survey measures as a group; and column 9 reports the coefficients from the full model where all six survey measures were included along with the two existing measures.

As shown in the three models, all coefficients that are statistically significant are positive except the one on the quality of professional development measure in column 5 of table D2. When the four teacher survey measures were added to the baseline model (with the two existing principal evaluation measures), the coefficient on the quality of professional

Table D2. Regression coefficients on principal evaluation models: Math models

	School value added achievement gains								
Measure	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Teacher survey									
Instructional leadership	0.56** (0.26)	na	na	na	0.78* (0.40)	na	na	na	0.86** (0.43)
Professional learning community	na	0.47 (0.36)	na	na	0.03 (0.48)	na	na	na	–0.58 (0.51)
Quality of professional development	na	na	0.11 (0.36)	na	–0.98** (0.49)	na	na	na	–0.79 (0.46)
Cultural press for excellence	na	na	na	0.44* (0.24)	0.39 (0.33)	na	na	na	0.40 (0.36)
Student survey									
School safety and climate	na	na	na	na	na	0.21 (0.22)	na	0.16 (0.23)	0.16 (0.23)
Classroom instructional environment	na	na	na	na	na	na	0.61** (0.26)	0.56* (0.26))	0.73* (0.41)
Existing measures									
Supervisor rating	0.37 (0.23)	0.35 (0.22)	0.31 (0.23)	0.34 (0.21)	0.33 (0.21)	0.30 (0.25)	0.35 (0.23)	0.34 (0.22)	0.36 (0.23)
Attendance rate	0.07 (0.13)	0.06 (0.12)	0.08 (0.14)	0.03 (0.13)	-0.09 (0.14)	0.01 (0.13)	–0.03 (0.13)	–0.02 (0.13)	–0.16 (0.14)

** Significant at $p < 0.05$; * significant at $p < 0.1$.

na is not applicable.

Note: Columns 1–4 report the coefficients from the models that added individual teacher survey measures separately into the baseline model containing the two existing measures, column 5 reports the coefficients from the model where the subset of four teacher survey measures were controlled together as a group, columns 6 and 7 report the coefficients from the models where the two student survey measures were examined separately, column 8 reports the coefficients on the subset of two student survey measures as a group, and column 9 reports the coefficients from the full model where all six survey measures were included along with the two existing measures. Values in parentheses are standard errors.

Source: Authors' analysis based on data provided by the district.

Table D3. Regression coefficients on principal evaluation models: Reading models

	School value added achievement gains								
Measure	1	2	3	4	5	6	7	8	9
Teacher survey									
Instructional leadership	0.49 (0.32)	na	na	na	0.53 (0.42)	na	na	na	0.73 (0.55)
Professional learning community	na	0.47 (0.44)	na	na	0.04 (0.60)	na	na	na	–0.57 (0.68)
Quality of professional development	na	na	0.28 (0.41)	na	–0.45 (0.61)	na	na	na	–0.27 (0.63)
Cultural press for excellence	na	na	na	0.38 (0.29)	0.26 (0.43)	na	na	na	0.45 (0.50)
Student survey									
School safety and climate	na	na	na	na	na	0.14 (0.27)	na	0.06 (0.23)	–0.12 (0.35)
Classroom instructional environment	na	na	na	na	na	na	0.63 (0.54)	0.52 (0.59)	0.69 (0.69)
Existing measures									
Supervisor rating	0.13 (0.25)	0.13 (0.25)	0.11 (0.26)	0.11 (0.25)	0.12 (0.22)	0.07 (0.25)	0.11 (0.23)	0.09 (0.26)	0.17 (0.29)
Attendance rate	0.14 (0.15)	0.14 (0.15)	0.17 (0.16)	0.10 (0.14)	0.07 (0.18)	0.14 (0.15)	0.06 (0.16)	0.03 (0.16)	–0.06 (0.20)

na is not applicable.

Note: Columns 1–4 report the coefficients from the models that added individual teacher survey measures separately into the baseline model containing the two existing measures, column 5 reports the coefficients from the model where the subset of four teacher survey measures were controlled together as a group, columns 6 and 7 report the coefficients from the models where the two student survey measures were examined separately, column 8 reports the coefficients on the subset of two student survey measures as a group, and column 9 reports the coefficients from the full model where all six survey measures were included along with the two existing measures. Values in parentheses are standard errors.

Source: Authors' analysis based on data provided by the district.

development measure is negative and statistically significant at p-value < 0.05. The negative coefficient may suggest that in the district included in this study, more professional development resources may have gone to schools with low math performance.

Table D4. Regression coefficients on principal evaluation models: Composite models

	School value added achievement gains								
Measure	1	2	3	4	5	6	7	8	9
Teacher survey									
Instructional leadership	0.53** (0.26)	na	na	na	0.69 (0.42)	na	na	na	0.78* (0.43)
Professional learning community	na	0.47 (0.37)	na	na	0.01 (0.48)	na	na	na	–0.45 (0.52)
Quality of professional development	na	na	0.19 (0.36)	na	–0.73 (0.50)	na	na	na	–0.54 (0.48)
Cultural press for excellence	na	na	na	0.41* (0.24)	0.34 (0.33)	na	na	na	0.42 (0.39)
Student survey									
School safety and climate	na	na	na	na	na	0.18 (0.22)	na	0.13 (0.23)	–0.09 (0.27)
Classroom instructional environment	na	na	na	na	na	na	0.58** (0.25)	0.51* (0.27)	0.71* (0.39)
Existing measures									
Supervisor rating	0.26 (0.23)	0.25 (0.22)	0.22 (0.22)	0.23 (0.21)	0.23 (0.21)	0.20 (0.22)	0.25 (0.23)	0.22 (0.22)	0.27 (0.21)
Attendance rate	0.11 (0.12)	0.10 (0.12)	0.12 (0.13)	0.07 (0.12)	–0.02 (0.14)	0.10 (0.12)	0.03 (0.13)	0.00 (0.13)	–0.07 (0.16)

** Significant at $p < 0.05$; * significant at $p < 0.1$.

na is not applicable.

Note: Columns 1–4 report the coefficients from the models that added individual teacher survey measures separately into the baseline model containing the two existing measures, column 5 reports the coefficients from the model where the subset of four teacher survey measures were controlled together as a group, columns 6 and 7 report the coefficients from the models where the two student survey measures were examined separately, column 8 reports the coefficients on the subset of two student survey measures as a group, and column 9 reports the coefficients from the full model where all six survey measures were included along with the two existing measures. Values in parentheses are standard errors.

Source: Authors' analysis based on data provided by the district.

Appendix E. Tripod Student Perception Survey

The survey items shown here are limited to those that were publicly released with the survey developer's permission in a 2010 report by the Bill & Melinda Gates Foundation's Measures of Effective Teaching Project (Kane & Cantrell, 2010). The student survey items used to measure school safety and climate are not presented because Cambridge Education maintains exclusive intellectual property rights to those items.

Please indicate how true each statement is by checking the appropriate box.	Totally Untrue	Mostly Untrue	Somewhat True	Mostly True	Totally True
1. My teacher in this class makes me feel that s/he really cares about me.	□	□	□	□	□
2. My teacher seems to know if something is bothering me.	□	□	□	□	□
3. My teacher really tries to understand how students feel about things.	□	□	□	□	□
4. Student behavior in this class is under control.	□	□	□	□	□
5. I hate the way that students behave in this class.	□	□	□	□	□
6. Student behavior in this class makes the teacher angry.	□	□	□	□	□
7. Student behavior in this class is a problem.	□	□	□	□	□
8. My classmates behave the way my teacher wants them to.	□	□	□	□	□
9. Students in this class treat the teacher with respect.	□	□	□	□	□
10. Our class stays busy and doesn't waste time.	□	□	□	□	□
11. If you don't understand something, my teacher explains it another way.	□	□	□	□	□
12. My teacher knows when the class understands, and when we do not.	□	□	□	□	□
13. When s/he is teaching us, my teacher thinks we understand even when we don't.	□	□	□	□	□
14. My teacher has several good ways to explain each topic that we cover in this class.	□	□	□	□	□
15. My teacher explains difficult things clearly.	□	□	□	□	□
16. My teacher asks questions to be sure we are following along when s/he is teaching.	□	□	□	□	□
17. My teacher asks students to explain more about answers they give.	□	□	□	□	□
18. In this class, my teacher accepts nothing less than our full effort.	□	□	□	□	□
19. My teacher doesn't let people give up when the work gets hard.	□	□	□	□	□
20. My teacher wants us to use our thinking skills, not just memorize things.	□	□	□	□	□
21. My teacher wants me to explain my answers—why I think what I think.	□	□	□	□	□
22. In this class, we learn a lot almost every day.	□	□	□	□	□
23. In this class, we learn to correct our mistakes.	□	□	□	□	□
24. This class does not keep my attention—I get bored.	□	□	□	□	□
25. My teacher makes learning enjoyable.	□	□	□	□	□
26. My teacher makes lessons interesting.	□	□	□	□	□
27. I like the ways we learn in this class.	□	□	□	□	□
28. My teacher wants us to share our thoughts.	□	□	□	□	□
29. Students get to decide how activities are done in this class.	□	□	□	□	□
30. My teacher gives us time to explain our ideas.	□	□	□	□	□
31. Students speak up and share their ideas about class work.	□	□	□	□	□
32. My teacher respects my ideas and suggestions.	□	□	□	□	□
33. My teacher takes the time to summarize what we learn each day.	□	□	□	□	□
34. My teacher checks to make sure we understand what s/he is teaching us.	□	□	□	□	□
35. We get helpful comments to let us know what we did wrong on assignments.	□	□	□	□	□
36. The comments that I get on my work in this class help me understand how to improve.	□	□	□	□	□

Scoring of Tripod student perception measures

CARE	Items 1–3	CHALLENGE	Items 24–27
CONTROL	Items 4–10	CONFER	Items 28–32
CLARIFY	Items 11–15	CONSOLIDATE	Items 33–36
CAPTIVATE	Items 16–23		

Notes

1. Examples of state laws are Illinois's Performance Evaluation Reform Act, 96–0861 (2011); Indiana's Senate Enrolled Act No. 1 (2011); Michigan's Public Act 205 of 2009 (2010); Minnesota's revised statute 122A.40 subd. 8 (2011); and Ohio's Amended Substitute House Bill Number 153 (2011). In exchange for flexibility on provisions of the federal Elementary and Secondary Education Act (commonly known as the No Child Left Behind Act of 2001), Wisconsin will (in part) incorporate schoolwide value-added data from statewide standardized assessments as a component of educator evaluation (Wisconsin Department of Public Instruction, 2012).
2. For example, the Standards for Educational and Psychological Testing by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999) provides the criteria most commonly used for evaluating tests, testing practices, assessments, and scales.
3. Goldring et al. (2012) used samples of 36–45 schools to examine the correlation between the principal evaluation tool, Vanderbilt Assessment of Leadership in Education, with other widely used leadership evaluation measures. Milanowski and Kimball (2012) examined the relationship between principal performance ratings and school value-added achievement growth in math and reading in two districts (72 schools in district A and 92 schools in district B).
4. Almost all students in elementary schools provided feedback on self-contained general education courses, and students in secondary schools provided feedback on subject-specific courses. Among the 8,345 students taking the 2011/12 student Tripod survey, 3,391 took the survey on general education (42.2 percent), 795 students on math (9.5 percent), and 1,131 on English language arts (13.6 percent). If the self-contained general education classrooms at the elementary level are counted as math and reading, 65 percent of students took the survey on math or reading courses.
5. Aggregating classroom instructional quality measures by subject would have caused a large loss of student survey responses, especially in secondary schools. For the 13 secondary schools included in the final sample, an average of 19 percent of survey classes are in math, and 22 percent are in English language arts.
6. On Chicago Public Schools, see http://www.cps.edu/Pages/valueadded.aspx (retrieved February 7, 2013); on Madison Metropolitan School District, see http://www.wcer.wisc.edu/projects/projects.php?project_num=554 (retrieved February 7, 2013); and on Milwaukee Public Schools, see http://www.wcer.wisc.edu/projects/projects.php?project_num=476 (retrieved February 7, 2013).
7. For this basic illustration of the empirical Bayes shrinkage method, a simple school indicator *s* is used in place of *s,g* for each school and grade combination. All estimates produced in this analytic step are grade- and subject-specific.
8. Separate school-level value-added estimates are produced for math and reading. In addition, a composite school-level value-added estimate across math and reading is produced by averaging the two subject-specific estimates on *q* weighted basis by test taker count.
9. A similar approach to combining value-added estimates across grades is used by Milanowski and Kimball (2012) for errors-in-variables regression estimates of school effects and by Isenberg and Hock (2011, 2012) for errors-in-variables regression estimates of teacher effects.
10. A *p*-value of 0.10 is a common cutoff value in variable selection processes and incremental validity research (Bendel & Afifi, 1977; Mickey & Greenland, 1989). Using

lower significance levels such as $p < 0.05$ increases the risk of eliminating candidate measures that research and theory suggest are important in school performance but cannot achieve statistical significance as a result of lower sample sizes.

11. The distribution of principal evaluation ratings in this study is less skewed than that in Milanowski and Kimball's (2012) study, which finds 97 percent of principals rated in the top scale (a range of 1) in one district and 70 percent of principals rated in a range of 0.25 in the second district.
12. The *p*-values range from 0.102 to 0.118 for the correlations between school value-added gains in math and the following evaluation measures: principal job function rating, instructional leadership, classroom instructional environment, and cultural press for excellence.
13. The incremental *F*-test can be specified as follows:

$$\mathrm{F}_{J,n-K-1} = \frac{R^2_{\text{unrestricted}} - R^2_{\text{baseline}}\; n-K-1}{1 - R^2_{\text{unrestricted}}\; J},$$

where R^2_{baseline} is the variance explained by the baseline model, which includes the two existing measures, $R^2_{\text{unrestricted}}$ is the explained variance of an unrestricted model that includes one or more candidate measures along with the two existing measures, K is the number of variables in the unrestricted model (three when testing the incremental utility of the individual measures, four when testing the joint significance of the two student survey measures, six when testing the joint significance of the four teacher survey measures, and eight when the joint significance of the full set of candidate measures), and J is the number of linear restrictions being tested, which equals the total number of variables in the unrestricted model minus the number of variables in the baseline model. To reject the null hypothesis, $\mathrm{F}_{J,n-K-1}$ must exceed the critical F value associated with a Cronbach's alpha value of 0.10.

14. The attendance rate of 0.54 is significantly lower than the average school attendance (0.91) in the sample and lower than that in other schools (the second lowest rate is 0.79). This school is a high school serving grades 9–12. Regardless of the low attendance rate, this school was kept in the final sample for several reasons. It contributed 127 student survey responses and 8 teacher survey responses. Dropping this school would further reduce the sample size, especially the small high school sample of only six schools. Although its attendance rate is considerably lower than that of other schools, the school's other variables are within the 20–75 percentile range among the sample schools. And as shown in later analysis, the attendance rate (included as an existing measure) contributed little to the power of the evaluation model in explaining the variance in school value-added outcomes.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95–135. Retrieved August 1, 2012, from http://www.jstor.org/stable/pdfplus/10.1086/508733.pdf

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*(2), 153–166. http://www.eric.ed.gov/?id=EJ600516

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association. http://www.eric.ed.gov/?id=ED436591

American Institutes for Research. (2007). *School climate and connectedness and student achievement*. Washington, DC: Author. Retrieved August 21, 2012, from http://alaskaice.org/wordpress/wp-content/uploads/2010/11/070918_SCCSandAchievement_AIRTechPaper.pdf

Barton, P. E., Coley, R. J., & Wenglinsky, H. (1998). *Order in the classroom: Violence, discipline, and student achievement*. Princeton, NJ: Educational Testing Service. Retrieved August 1, 2012, from http://www.ets.org/Media/Research/pdf/PICORDER.pdf

Bendel, R. B. & Afifi, A. A. (1977). Comparison of stopping rules in forward "stepwise" regression. *Journal of the American Statistical Association, 72*(357), 46–53. Retrieved March 1, 2013, from http://www.jstor.org/stable/2286904

Béteille, T., Kalogrides, D., & Loeb, S. (2009). *Effective schools: Managing the recruitment, development, and retention of high-quality teachers*. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research, the Urban Institute. http://eric.ed.gov/?id=ED509688

Betts, J. R., & Grogger, J. (2003). The impact of grading standards on student achievement, educational attainment, and entry-level earnings. *Economics of Education Review, 22*(4), 343–352. http://eric.ed.gov/?id=EJ669559

Bottoms, G., & Schmidt-Davis, J. (2010). *The three essentials: Improving schools requires district vision, district and state support, and principal leadership.* Atlanta, GA: Southern Regional Education Board. Retrieved September 1, 2012, from http://publications.sreb.org/2010/10V16_Three_Essentials.pdf

Brewer, D. J. (1993). Principals and student outcomes: Evidence from U.S. high schools. *Economics of Education Review, 12*(4): 281–292. http://eric.ed.gov/?id=EJ477488

Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organizing schools for improvement: Lessons from Chicago.* University of Chicago Press.

Carrell, S. E., & Hoekstra, M. L. (2011). Externalities in the classroom: How children exposed to domestic violence affect everyone's kids. *American Economic Journal: Applied Economics, 2*(1), 211–228.

Carroll, B. R. (2006). *The effects of school violence and crime on academic achievement*. Davidson, NC: Davidson College. Retrieved August 1, 2012, from http://econ.duke.edu/uploads/assets/dje/2006_Symp/Carroll.pdf

Clifford, M., Behrstock-Sherratt, E., & Fetters, J. (2012). *The ripple effect: A synthesis of research on principal influence to inform performance evaluation design*. Quality School Leadership Issue Brief. Washington, DC: American Institutes for Research. http://eric.ed.gov/?id=ED530748

Clifford, M., Menon, R., Gangi, T., Condon, C., & Hornung, K. (2012). *Measuring school climate for gauging principal performance: A review of the validity and reliability of publicly accessible measures*. Quality School Leadership Issue Brief. Washington, DC: American Institutes for Research. http://eric.ed.gov/?id=ED531401

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Erlbaum.

Cooper, B. S., Ehrensal, P. A., & Bromme, M. (2005). School-level politics and professional development: Traps in evaluating the quality of practicing teachers. *Educational Policy, 19*(1), 112–125. http://eric.ed.gov/?id=EJ690113

Desimone, L. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38*(3), 181–199. http://eric.ed.gov/?id=EJ883327

Elmore, R. F. (2000). *Building a new structure for school leadership*. Washington, DC: Albert Shanker Institute. Retrieved July 2, 2012, from http://www.shankerinstitute.org/Downloads/Building.pdf

Ferguson, R. F. (2011, May 3). *Tripod classroom-level student perceptions as measures of teaching effectiveness* [Slide presentation]. Cambridge, MA: Harvard University, National Center for Teacher Effectiveness. Retrieved June 20, 2012, from http://www.gse.harvard.edu/ncte/news/NCTE_Conference_Tripod.pdf

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective: Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915–945. http://eric.ed.gov/?id=EJ648260

Goddard, Y. L., Goddard, R. D., & Tschannen-Moran, M. (2007). A theoretical and empirical investigation of teacher collaboration for school improvement and student achievement in public elementary schools. *Teachers College Record, 109*(4), 877–896. http://eric.ed.gov/?id=EJ820449

Goldring, E., Cravens, X., Murphy, J., Porter, A., & Elliott, S. (2012, March 17). *The convergent and divergent validity of the Vanderbilt Assessment of Leadership in Education™*

(VAL-ED): Instructional leadership and emotional intelligence. Paper presented at the 2012 Annual Meeting of the Association for Education Finance and Policy, Boston, MA. Retrieved June 25, 2013, from http://aefpweb.org/sites/default/files/webform/AEFP%202012%20Final%20Draft_March%206%202012.pdf

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job* (Hamilton Project Discussion Paper). Washington, DC: Brookings Institution. Retrieved February 12, 2013, from http://www.brookings.edu/~/media/Research/Files/Papers/2006/4/education%20gordon/200604hamilton_1.PDF

Greene, W. H. (2003). *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Greene, W. H. (2011). *Econometric analysis* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Grissom, J., & Loeb, S. (2009). *Triangulating principal effectiveness: How perspectives of parents, teachers, and assistant principals identify the central importance of managerial skills* (Working Paper 35). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research. Retrieved May 17, 2012 from http://www.urban.org/uploadedpdf/1001443-Triangulating-Principal-Effectiveness.pdf

Hallinger, P. (2003). Leading educational change: Reflections on the practice of instructional and transformational leadership. *Cambridge Journal of Education, 33*(3), 329–352. Retrieved September 17, 2012 from http://philiphallinger.com/old-site/papers/CCJE%20Instr%20and%20Trans%20Lship%202003.pdf

Hallinger, P., & Heck, R. (1996). Reassessing the principal's role in school effectiveness: A review of empirical research, 1980–1995. *Educational Administration Quarterly, 32*(1), 5–44.

Hallinger, P., & Heck, R. (2002). What do you call people with visions? The role of vision, mission, and goals in school leadership and improvement. In K. A. Leithwood & P. Hallinger (Eds.), *The second international handbook of educational leadership and administration* (pp. 9–40). Dordrecht, The Netherlands: Kluwer.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2), 267–271.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Haynes, S. N., & O'Brien, W. B. (2000). *Principals of behavioral assessment: A functional approach to psychological assessment*. New York: Kluwer.

Heck, R. H., & Hallinger, P. (2009). Assessing the contribution of distributed leadership to school improvement and growth in math achievement. *American Educational Research Journal, 46*(3), 659–689. http://eric.ed.gov/?id=EJ883287

Henrich, C. C., Schwab-Stone, M., Fanti, K., Jones, S. M., & Ruchkin, V. (2004). The association of community violence exposure with middle-school achievement: A prospective study. *Journal of Applied Developmental Psychology, 25*(3), 327–348. http://eric.ed.gov/?id=EJ731751

Illinois Principals Association & Illinois Association of School Administrators. (2012). *A guide to implementing principal performance evaluation in Illinois*. Springfield, IL: Illinois Principals Association. Retrieved May 5, 2012, from http://www.ilprincipals.org/resources/resource-documents/principal-evaluation/ipep/Guide%20to%20Implementing%20Principal%20Performance%20Evaluation%20in%20Illinois.pdf

Illinois State Board of Education. (2011). *2011 teacher/principal evaluation systems—Final collection*. Springfield, IL: Author. Retrieved May 6, 2012, from http://www.isbe.net/peac/pdf/survey/teacher_prin_eval_survey11_final.pdf

Isenberg, E., & Hock, H. (2011). *Design of value-added models for IMPACT and TEAM in DC public schools, 2010–2011 school year.* (Final Report.) Princeton, NJ: Mathematica Policy Research. Retrieved March 14, 2013, from http://dcps.dc.gov/DCPS/Files/downloads/In-the-Classroom/Design%20of%20Value-Added%20Models%20for%20DCPS%202010–2011.pdf

Isenberg, E., & Hock, H. (2012). *Measuring school and teacher value-added in DC, 2011–2012 school year.* (Final Report.) Princeton, NJ: Mathematica Policy Research. Retrieved March 14, 2013, from http://www.mathematica-mpr.com/publications/pdfs/education/value-added_DC.pdf

Kane, T. J., & Cantrell, S. (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching Project*. Seattle, WA: Bill & Melinda Gates Foundation

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2006). *What does certification tell us about teacher effectiveness? Evidence from New York City* (NBER Working Paper No. 12155). Cambridge, MA: National Bureau of Economic Research. Retrieved May 8, 2012, from http://www.nber.org/papers/w12155.pdf

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (NBER Working Paper No. 14607). Cambridge, MA: National Bureau of Economic Research. Retrieved March 1, 2012, from http://www.nber.org/papers/w14607.pdf

Kimball, S. M., Milanowski, A. T., & McKinney, S. A. (2007, April 10). *Implementation of standards-based principal evaluation in one school district: First year results from randomized trial.* Paper presented at the 2007 Annual Meeting of the American Educational Research Association, Chicago. Retrieved August 21, 2012, from http://cpre.wceruw.org/publications/KimballMilanowskiMcKinney.pdf

Leithwood, K., Louis, K. S., Anderson, S., & Wahlstrom, K. (2004). *How leadership influences student learning.* University of Minnesota, Center for Applied Research and Educational Improvement & University of Toronto, Ontario Institute for Studies in Education. Retrieved August 21, 2012, from http://www.wallacefoundation.org/knowledge-center/school-leadership/key-research/Documents/How-Leadership-Influences-Student-Learning.pdf

Leithwood, K., & Riehl, C. (2003). *What we know about successful school leadership.* Washington, DC: American Educational Research Association. Retrieved March 1, 2012, from http://dcbsimpson.com/randd-leithwood-successful-leadership.pdf

Levine, D. U., & Lezotte, L. W. (1990). *Unusually effective schools: A review and analysis of research and practice*. Madison, WI: National Center for Effective Schools Research and Development. http://eric.ed.gov/?id=ED330032

Lomos, C., Hofman, R. H., & Bosker, R. J. (2011). Professional communities and student achievement: A meta-analysis. *School Effectiveness and School Improvement, 22*(2), 121–148. http://eric.ed.gov/?id=EJ925055

Louis, K. S., Leithwood, K., Wahlstrom, K. L., & Anderson, S. E. (2010). *Investigating the links to improved student learning.* University of Minnesota, Center for Applied Research and Educational Improvement & University of Toronto, Ontario Institute for Studies in Education. Retrieved March 1, 2012, from http://www.wallacefoundation.org/knowledge-center/school-leadership/key-research/Documents/Investigating-the-Links-to-Improved-Student-Learning.pdf

Louis, K. S., Marks, H. M., & Kruse, S. D. (1996). Teachers' professional community in restructuring schools. *American Educational Research Journal, 33*(4), 757–798.

Marzano, R. J., Waters, T., & McNulty, B. A. (2005). *School leadership that works: From research to results.* Alexandria, VA: ASCD. http://eric.ed.gov/?id=ED509055

Mattson Almanzán, H., Sanders, N., & Kearney, K. (2011). *How six states are implementing principal evaluation systems*. San Francisco: WestEd. Retrieved March 1, 2012, from http://www.wested.org/online_pubs/resource1105.pdf

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/content/dam/rand/pubs/monographs/2004/RAND_MG158.pdf

Mickey, R. M., & Greenland, S. (1989). The impact of confounder selection criteria on effect estimation. *American Journal of Epidemiology, 129*(1), 125–137.

Milanowski, A., & Kimball, S. M. (2012, March). *The relationship between standards-based principal performance evaluation ratings and school value-added: Evidence from two districts.* Paper presented at the Association for Education Finance and Policy 37th Conference, Boston. Retrieved April 22, 2012, from http://www.aefpweb.org/sites/default/files/webform/MilanowskiKimballPrinEvalAEFP2010.doc

Murphy, J., Elliot, S. N., Goldring, E., & Porter, A. C. (2007). Leadership for learning: A research-based model and taxonomy of behaviors. *School Leadership and Management, 27*(2), 179–201. http://eric.ed.gov/?id=EJ786139

National Center for Education Statistics. (1997). *Teacher professionalization and teacher commitment: A multilevel analysis* (NCES 97–069). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics. Retrieved August 1, 2012, from http://nces.ed.gov/pubs/97069.pdf

National Center for Education Statistics. (2008). *Table 8: Percentage of public school principals who thought they had a major influence on evaluating teachers at their school and hiring new full-time teachers at their school, by state: 2007–08* [Webpage]. In National Center for Education Statistics, School and Staffing Survey data file. Retrieved March 1, 2013, from http://nces.ed.gov/surveys/sass/tables/sass0708_2009323_p1s_08.asp

National Conference of State Legislatures. (2011, November 15). *State approaches to evaluating school principal effectiveness webinar* [Slide presentation]. Washington, DC: Author. Retrieved May 6, 2012, from http://www.ncsl.org/documents/educ/Evaluating PrinpcipalEffecitvenessPowerPointSlides.pdf

The New Teacher Project. (2012). *The Indiana evaluation pilot: Mid-year report and recommendations*. Indianapolis: Indiana Department of Education. Retrieved August 1, 2012, from http://www.riseindiana.org/sites/default/files/files/IN_MidYear_FINAL.pdf

Newmann, F. M. (1998). How secondary schools contribute to academic success. In K. Borman & B. Schneider (Eds.), *Youth experiences and development: Social influences and educational challenges*. University of Chicago Press, National Society for the Study of Education Yearbook.

Newmann, F. M., King, M. B., & Youngs, P. (2000, April 28). *Professional development that addresses school capacity*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. Retrieved March 26, 2013, from http://www.wcer.wisc.edu/archive/pdbo/grand-aje411.doc

Northwest Evaluation Association. (2011). *Technical manual: For Measures of Academic Progress (MAP) and Measures of Academic Progress for Primary Grades (MPG)*. Portland, OR: Author.

Nunnally, J. (1978). *Psychometric theory* (2d ed.). New York: McGraw-Hill.

Ohio Department of Education. (2011). *Ohio Principal Evaluation System: Model packet*. Columbus, OH: Author. Retrieved May 6, 2012, from http://www.onetohio.org/library/Documents/OPES%20Model%20Final%20(3)09272011%20(2)_km-2.pdf.

Porter, A. C., Murphy, J., Goldring, E., Elliott, S. N., Polikoff, M. S., & May, H. (2008). *Vanderbilt Assessment of Leadership in EducationTM: Technical manual, version 1.0.* Nashville, TN: Discovery Education Assessment. Retrieved April 23, 2012, from http://www.wallacefoundation.org/knowledge-center/school-leadership/principal-evaluation/Documents/Vanderbilt-Assessment-of-Leadership-in-Education-Technical-Manual-1.pdf

Portin, B. S., Knapp, M. S., Dareff, S., Feldman, S., Russell, F. A., Samuelson, C., & Yeh, T. L. (2009). *Leadership for learning improvement in urban schools*. University of Washington, Center for the Study of Teaching and Policy. Retrieved August 21, 2012, from http://www.wallacefoundation.org/knowledge-center/school-leadership/

district-policy-and-practice/Documents/Leadership-for-Learning-Improvement-in-Urban-Schools.pdf

Portin, B. S., Schneider, P., DeArmond, M., & Gundlach, L. (2003). *Making sense of leading schools: A national study of the principalship.* Seattle, WA: Center on Reinventing Public Education. Retrieved August 21, 2012, from http://www.wallacefoundation.org/knowledge-center/school-leadership/principal-training/Documents/Making-Sense-of-Leading-Schools-Study-of-School-Principalship.pdf

Ripski, M. B., & Gregory, A. (2009). Unfair, unsafe, and unwelcome: Do high school students' perceptions of unfairness, hostility, and victimization in school predict engagement and achievement? *Journal of School Violence, 8*(4), 355–375. http://eric.ed.gov/?id=EJ864851

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

Robinson, V. M. J., Lloyd, C. A., & Rowe, K. J. (2008). The impact of leadership on student outcomes: An analysis of the differential effects of leadership types. *Educational Administration Quarterly, 44*(5), 635–674. http://eric.ed.gov/?id=EJ818931

Roeber, E. (2011). *Educator evaluation—Models, parameters, issues, and implementation.* East Lansing, MI: Michigan Education Association. Retrieved May 9, 2012, from http://www.michiganassessmentconsortium.org/sites/default/files/MAC-Whitepaper-Roeber-Evaluation-Models.pdf

Sanders, N., Kearney, K., & Vince, S. (2012). *Using multiple forms of data in principal evaluations: An overview with examples.* San Francisco: WestEd. Retrieved May 2, 2012, from http://www.wested.org/wp-content/files_mf/1394052656Multiple_forms_in_PEval_20120417_final.pdf

Saunders, W. M., Goldenberg, C. N., & Gallimore, R. (2009). Increasing achievement by focusing grade-level teams on improving classroom learning: A prospective, quasi-experimental study of Title I schools. *American Educational Research Journal, 46*(4), 1006–1033. http://eric.ed.gov/?id=EJ883290

Sebastian, J., & Allensworth, E. (2012). The influence of principal leadership on classroom instruction and student learning: A study of mediated pathways to learning. *Educational Administration Quarterly, 48*(4), 626–663. http://eric.ed.gov/?id=EJ978165

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15*(2), 201–293.

Spillane, J., Halverson, R., & Diamond, J. (2004). Towards a theory of leadership practice: A distributed perspective. *Journal of Curriculum Studies, 36*(1), 3–34. http://eric.ed.gov/?id=EJ695066

Stronge, J. H., Ricard, H. B., & Catano, N. (2008). *Qualities of effective principals.* Alexandria, VA: ASCD (formerly the Association for Supervision and Curriculum Development). http://eric.ed.gov/?id=ED509122

Valentine, J., Clark, D. C., Hackmann, D. G., & Petzko, V. N. (2004). *Leadership for highly successful middle level schools, vol. 2: A national study of leadership in middle level schools*. Reston, VA: National Association of Secondary School Principals.

Value-Added Research Center. (2010). *NYC Teacher Data Initiative: Technical report on the NYC value-added model*. Retrieved December 12, 2002, from http://schools.nyc.gov/NR/rdonlyres/A62750A4-B5F5–43C7-B9A3-F2B55CDF8949/87046/TDINYCTechnicalReportFinal072010.pdf

Vescio, V., Ross, D., & Adams, A. (2008). A review of research on the impact of professional learning communities on teaching practice and student learning. *Teaching and Teacher Education, 24*(1), 80–91. http://eric.ed.gov/?id=EJ782410

Wacyk, L., Reeves, P., McNeill, P., & Zimmer, D. (2011, September 26). *School ADvance: An administrator evaluation system*. [Slide presentation] Presented at the 2011 MASA Fall Conference, Lansing, MI. Retrieved May 1, 2012, from http://gomasa.org/sites/default/files/ADvance%20MASA%20presentation%20-%209–26.pptx

Wahlstrom, K. L., Louis, K. S., Leithwood, K., & Anderson, S. E. (2010). *Investigating the links to improved student learning: Executive summary of research findings*. University of Minnesota, Center for Applied Research and Educational Improvement & University of Toronto, Ontario Institute for Studies in Education. Retrieved June 4, 2012, from http://www.wallacefoundation.org/knowledge-center/school-leadership/key-research/Documents/Investigating-the-Links-to-Improved-Student-Learning-Executive-Summary.pdf

Waters, T., Marzano, R. J., & McNulty, B. A. (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. http://eric.ed.gov/?id=ED481972

Wisconsin Department of Public Instruction. (2012, amended 2013). *Wisconsin ESEA flexibility request*. Madison, WI: Author. Retrieved June 17, 2014, from http://esea.dpi.wi.gov/files/esea/pdf/waiver-final.pdf

Wisconsin Educator Effectiveness Design Team. (2011). *Preliminary report and recommendations*. Retrieved May 1, 2012, from http://ee.dpi.wi.gov/files/ee/pdf/ee_report_prelim.pdf

Witziers, B., Bosker, R. J., & Krüger, M. L. (2003). Educational leadership and student achievement: The elusive search for an association. *Educational Administration Quarterly, 39*(3), 398–425. http://eric.ed.gov/?id=EJ672892

Worrell, F. C., & Kuterbach, L. D. (2001). The use of student ratings of teacher behaviors with academically talented high school students. *Journal of Secondary Gifted Education, 12*(4), 236–247. http://eric.ed.gov/?id=EJ632563

The Regional Educational Laboratory Program produces 7 types of reports

Making Connections
Studies of correlational relationships

Making an Impact
Studies of cause and effect

What's Happening
Descriptions of policies, programs, implementation status, or data trends

What's Known
Summaries of previous research

Stated Briefly
Summaries of research findings for specific audiences

Applied Research Methods
Research methods for educational settings

Tools
Help for planning, gathering, analyzing, or reporting data or research