BUILDING AN INCLUSIVE SCHOOL ENVIRONMENT

CREATING SAFE CONDITIONS FOR LGBT STUDENTS
MAKING STUDENT MENTAL HEALTH A PRIORITY
COMBATING CYBERBULLYING THROUGH STUDENT VOICE
how do you evaluate leadership?

The impossibility of evaluating principals and leadership programs using student outcomes

W. Norton Grubb, Patrick Liao, and Rebecca Cheung

In our age of accountability and evaluation, teachers were first subject to competency exams in the 1980s; now they’re subject to value-added measures. States imposed accountability systems on schools during the 1990s, and the federal government created accountability for districts in No Child Left Behind. Now accountability and more formal evaluation have spread to school leaders, because they may be second in importance only to teachers. The two national organizations of principals have produced frameworks for evaluation (NAESP and NASSP, n.d.), and formal methods of evaluation like VAL-ED are widely used. Inconclusive statistical literature has tried to ascertain what dimensions of leadership affect outcomes (Grissom et al., 2014). A few states require that every administrator’s evaluation be partly based on test scores, and some districts are using test scores to determine salaries or bonuses (Murphy et al., 2014). The U.S. Department of Education is now turning its attention to principals after proposing plans to evaluate teacher preparation programs with various measures, including test score gains.

Now the focus is turning to preparation programs for leaders. Frameworks for evaluating leadership programs exist, using multiple ways of ascertaining effectiveness (Tredway et al., 2012). Two high-profile programs (the Aspiring Principals Program in New York City and New Leaders for New Schools) have linked their efforts to test scores, with distinctly mixed and often negative outcomes (Corcoran et al., 2012; Gates et al., 2014), along with methodological flaws that we will review.

Closer to home, foundations have asked our program—the Principal Leadership Institute (PLI) at the University of California, Berkeley—how we can prove our impact, including effects on test scores, since many foundations do not fund programs without outcome evidence. Federal grants ask grantees to quantify their effects, including those on student outcomes. In California the need to improve and monitor leadership programs is
problems, but not for conclusions about effectiveness.

Existing Evaluations of the PLI
The PLI opened in 1999 and enrolls between 25 and 35 individuals annually. It prepares leaders for urban districts and consequently stresses equity and the treatment of historically underserved students. PLI students enroll for 14.5 months, learning from various instructors and coaches who work with them in their schools; they take numerous problem- and project-based classes, as well as participate in four practicums and several leadership experiences. The program has almost 500 graduates, virtually all in urban education. (More information on the PLI, the impact reports noted below, and the longer study underlying this article can be found at http://gse.berkeley.edu/policy-organization-measurement-evaluation/pli.)

While the PLI collects annual data on its graduates, such information is purely descriptive. We have, therefore, supplemented these with four impact reports:
- An early report used questionnaires to graduates of PLI and other programs to learn about their satisfaction with preparation. The results revealed high levels of satisfaction among PLI graduates and consistently lower satisfaction among other graduates. However, the response rate for other graduates was low (with only 20 responses), making comparisons uncertain.
- A 2012 report, “Breadth of Regional Impact and Strength of Model,” analyzed data on the program’s graduates. Based on a questionnaire about what positions they occupied, 98 percent worked in education, 66 percent in administrative positions, and 24 percent as teacher-leaders. The report also described the PLI model.
- Cheung’s “Alumni and District Partner Feedback” (2013), an example of stakeholder evaluation, described alumni surveys and feedback from four partner districts: San Francisco, Oakland, Berkeley, and West Contra Costa. Of all respondents, 94 percent agreed that PLI provided strong preparation for their positions, and 100 percent of district partners provided positive feedback over 10 years.
- Grubb and Cheung’s “Collective and Team Leadership: Preparation for Urban Schools” (2014) was based on interviews with graduates working in schools with two or more PLI alumni. It revealed many benefits of having multiple PLI graduates at one school, a practice we labeled “collective leadership,” and it confirmed that many dimensions of PLI are valuable on the job.

However, none of these impact reports focused on student outcomes, so we decided to carry out an outcome evaluation despite the challenges. The outcomes that can be used for any program are largely limited to those available in state and district data, especially math and English test scores, rather than the many other consequences important for students, and most evaluations don’t include any other results. We also collected data on suspension and truancy rates, and student perceptions measured by the California Healthy Kids Survey. Because factors other than leadership influence outcomes, our analysis required variables reflecting such factors, particularly the percentage of low-income students (measured crudely by the proportion eligible for free and reduced-price meals);
Many school-level influences on student outcomes need to be considered before concluding that leaders are responsible for outcomes.

teachers with full and emergency credentials; racial minority students in eight different categories; students in the Gifted and Talented Education program; migrant students; English learners and those redesignated as English proficient; and students in special education. The data available to us, and to virtually all preparation programs, are quite limited, and include little information about other dimensions of schools.

In considering the effects of principals, it is crucial to confine the analysis to a specific district. Districts vary enormously in their practices: Some assign new principals to the most difficult schools, while others assign the strongest (or most experienced) to the weakest schools. Districts vary substantially in their support for leaders, and some undermine principals by constraining their autonomy and imposing too many bureaucratic requirements. In addition, it takes time for principals to improve schools. Initially, therefore, we defined the study population to include graduates of the PLI in the Oakland district with at least three years of experience at the same school. Unfortunately, while PLI has about 500 graduates, this restricted sample included only 24 alumni from 2001 to 2013, too few for any statistical analyses. In subsequent work, therefore, we used samples unrestricted by experience.

The most widely used test score measure in California is the Academic Performance Index (API), a rescaling of California State Tests to a range of 200–1,000. PLI graduates are typically placed at schools with slightly higher API scores than other schools. However, higher scores may reflect students’ family backgrounds, their preparation levels, or school characteristics only partially controlled by principals. To adjust for these influences, we used a statistical technique—regression analysis—to determine which other variables might influence API scores. When we allowed for effects on test score growth rather than levels—a value-added approach—the coefficient associated with a PLI leader was small and statistically insignificant, suggesting that PLI graduates did not have greater effects than graduates of other programs. Instead, scores were higher in schools with higher-income students; a higher proportion in gifted programs; lower proportions of African-Americans, Latinos, Pacific Islanders, English language learners, special education students, and students redesignated as English proficient. Finally, scores were higher in schools with more fully and emergency-credentialed teachers, compared to those lacking credentials. These results are unsurprising, since they appear in most analyses of test scores.

Suspension and truancy rates are important because they lead to students being out of school and are never good for students’ academic progress. When we analyzed these using the same value-added formulation as for test scores, roughly the same results emerged. There were no statistically significant differences between PLI-led schools and those with principals from other programs. Instead, suspension rates were higher in schools with more low-income students and more African-American students, reflecting the common finding that African-American boys are likely to be disciplined even for offenses for which other students are not. Truancy rates were higher in schools with more African-Americans, low-income students, and students redesignated fluent English proficient. As with test scores, then, any differences in suspension and truancy rates that might be due to principals were overwhelmed by the characteristics of students.

The California Healthy Kids Survey collects information about students’ perceptions of belonging at school, leadership, and home lives. Schools with negative responses are poor places to learn. Many responses to these questionnaires were quite similar for PLI and non-PLI schools. However, the answers to whether “I plan to graduate from high school” and “I know where to go for help” were much more positive for PLI-led schools. Students in PLI-led schools also had significantly more positive responses to questions about their sense of belonging and support. Within schools designated by the state as being in the lowest 10 percent measured by high levels of poverty and racial-minority students, these differences were particularly large, especially for responses about teachers and others who care about students, and about teachers who listen when students have something to say.

PLI emphasizes preparing graduates for urban schools, to improve the common conditions of inexperienced and demeaning teachers, harsh discipline policies, class, and racial discrimination. Furthermore, graduates agree that these elements of PLI are important in their positions. We like to think, therefore, that the differences in student responses are due to the preparation principals receive in PLI, though more detailed analysis would be necessary to demonstrate this.
Unavoidable Predicaments

Overall, we found only a little evidence that PLI does a better job of preparing school leaders than other programs serving the Oakland district. There were no differences in test scores, suspension or truancy rates, though students in PLI-led schools had more positive conceptions of their schools as safe and supportive places.

It’s possible, of course, that student composition and the resources of a school overwhelm the effects of principals who may not be powerful enough to prevail over other strong influences. We doubt this, however; other analyses with more detailed data reveal that dozens of school resources affect students, many of which can be influenced by principals and leadership teams (Grubb, 2009).

Instead, we think that all evaluations, including ours, suffer from at least 11 problems that cannot be resolved with existing data and methods:

- The need to confine analysis to one district, and to principals with several years of experience, reduces the sample size dramatically. Programs like PLI that prepare graduates for several districts are, therefore, difficult to evaluate.
- Just as principals need some experience in a school to have any effect, students must have some experience with a principal for any effect to develop. Therefore, the most appropriate samples include principals with experience in a school and students with time under that principal. Such samples are almost impossible to get, and sample size would be small.
- Outcome measures are limited, usually to test scores. In most states, test score measures are complex and opaque. They are rarely vertically equated so that a point in one year means the same as a point in another year; strictly speaking, they cannot be used to measure added value. The lack of variables measuring progress through schooling is particularly serious because test scores and progress respond to different school resources (Grubb, 2009, 70–72).
- The effectiveness of principals depends on what schools they are assigned to and the composition of students. Every study of student outcomes has found family background among the strongest predictors of success, but only a crude measure of family income is available; other dimensions like parental education, occupation, and aspirations for children are inaccessible except in large, longitudinal data sets. Without these measures, principals in schools with high socioeconomic statuses always look more effective than those in schools with low socioeconomic statuses, and accountability systems will drive principals from low-performing to high-performing schools and districts.
- Similarly, many school-level influences on student outcomes need to be considered (or controlled) before concluding that leaders are responsible for outcomes. Some—the quality of teachers and instruction, a school’s climate, the availability of student supports—may be partially under a principal’s control, but others (especially funding and other resources, union agreements, and civic support) are not.
- Schools have varying levels of district support. In Oakland, low-performing schools are coded according to test scores, test score growth, and progress in closing achievement gaps. Schools with the lowest codes are targeted for special district support and monitoring. As a result, school-level evaluations must include information on district practices.
- Just as value-added measures for teachers vary from year to year (Darling-Hammond et al., 2012), principal effects on outcomes may vary as students come and go, district and state policies shift, funding increases or declines, and different statistical models and test data are used. With unstable measures, judging the effectiveness of principals and leadership preparation becomes unreliable.
- Many leadership programs, including the three mentioned in this article, select their students carefully, while others are nonselective. For example, PLI selects students on the basis of their experience in and commitment to urban schools and equity. But no evaluation has managed to account for selection and self-selection, so they cannot distinguish between a program’s effectiveness and its selectiveness.
- Evaluations use the wrong model of leadership. Principals don’t affect students directly; teachers do that through their instruction. Instead, leaders affect students indirectly, through effects on teachers, climate, the curriculum, and reforms adopted (though these are restricted in many districts); through linkages to parents, external organizations, and community services; and through their success in capturing additional resources. These resources, in turn, impact student outcomes. The right way to describe the effects of principals is, therefore, to estimate two kinds of equations, one describing the influences of principals on school processes and resources, and
the second analyzing the effects of these resources plus student characteristics on outcomes. But if the two equations are combined (as in all existing analyses), then all coefficients including those describing the effects of principals are certain to be biased. (For this demonstration, see the longer paper on which this article is based, viewable on the PLI website.)

In some schools, efforts to measure the influence of principals mistakes the nature of leadership. The conventional approach assumes that leadership is embodied in the principal only, with a top-down hierarchy. But many schools have different practices: high schools with leadership teams; those adopting distributed leadership, with responsibilities shared among principals, assistant principals, teacher-leaders, and coaches (Spillane, 2006); schools with collective leadership based on the similar backgrounds and values of several leaders. Evaluation must then focus on a school’s leadership team or its leadership environment, as the CALL system does (Halvorsen et al., 2014). But existing evaluations do not reflect such distributed or collective leadership.

The analysis of student outcomes can tell only whether leaders, or principal preparation programs, affect outcomes, not how they do so. Such evaluations are useful only for threatening principals with dismissal or leadership programs with closure, not for improving them. An emphasis on punishment rather than improvement cannot make schools better places for students.

Finally, we should consider the cost of evaluation. The analysis of New Leaders by RAND, producing several studies, cost $3.5 million; its replication in Oakland by Mathematica must have cost $100,000 or more; and even so those evaluations are inconclusive and flawed. (In contrast, the PLI evaluation cost about $30,000.) Few leadership programs have such resources, and most have to hire statistical expertise from outside. State and federal governments and foundations should understand the effort and funding required to find proof of effectiveness.

Unfortunately, we conclude that it is impossible at this stage to use student outcomes to evaluate principals and leadership programs; educators ought to move away from the narrative that leadership and preparation can be judged by test scores. We must fall back on conventional but informative kinds of evaluation, collecting information about practices that presumptively influence outcomes. These might include content evaluations examining a program’s courses, syllabi, and practicums; satisfaction surveys asking graduates how well they have been prepared; stakeholder surveys asking teachers, district officials, parents, and students about the effectiveness of leaders (as VAL-ED and CALL).

Some of these methods have distinct advantages over outcome evaluations. They can readily identify what dimensions of leadership need to be improved and whether disagreements exist among stakeholders. Their results are more comprehensible and transparent than complex statistical analyses. In the absence of better data and methods, these traditional forms of evaluation will have to do, despite the pressures of this age of accountability. PL

REFERENCES


Gates, S. et al. (2014). Preparing principals to raise student achievement. Santa Monica: RAND.


Murphy, J., Goodring, E., & Porter, A. (2014). Principal evaluation takes center stage. Principal, 93(3).


W. Norton Grubb was the David Pierpont Gardner Professor in Higher Education Emeritus at the UC Berkeley Graduate School of Education. He was the founding faculty coordinator of the Principal Leadership Institute, a program to prepare principals for urban schools in the Bay Area.

Patrick Liao is a graduate student researcher studying in the field of Public Policy at the University of California, Berkeley.

Rebecca Cheung is program director of the Principal Leadership Institute. Prior to that, she worked in urban schools for 15 years as a teacher, middle school principal, and district office administrator.