Title

Evaluating Berkeley's Principal Leadership Institute: Experiments With Student Outcomes

Authors

Grubb, W. Norton
Liao, Patrick
Cheung, Rebecca

Publication Date

2014-09-01

Evaluating Berkeley's Principal Leadership Institute:

Experiments With Student Outcomes

Professor W. Norton Grubb, Faculty Coordinator*

Patrick Liao, Research Associate

Rebecca Cheung, Ed.D., Program Coordinator

University of California, Berkeley

Principal Leadership Institute

Graduate School of Education

September 2014

*The analysis in this paper has been supported by funds from the David Gardner Chair in Higher Education, held by W. Norton Grubb.

Education is engulfed in an age of accountability and evaluation. *Teachers* have been subject to several waves of evaluation, first in the 1980s with teacher competency exams, and more recently in efforts to measure effectiveness with value-added measures (or test score gains) plus other mechanisms like teacher observations. States imposed their own test-based accountability systems on *schools* during the 1990s; then the federal government created its own accountability system in No Child Left Behind, based on state tests but with different consequences. In theory these accountability systems could be used to improve the quality of schooling, but much more often they have served to punish (and only sometimes reward) teachers and schools.

Now accountability and evaluation have spread to *principals* and other school-level leaders, usually with the understanding that they may be second in importance only to teachers.[1] A small empirical literature has tried to ascertain what specific dimensions of leadership affect student outcomes.[2] Formal methods of evaluating principals have been developed, including the widely-used VALED system from Vanderbilt University and several others. The two national organizations of principals have produced a framework for principal evaluation based on six domains of leadership, and other general frameworks have been developed.[3] Following the lead of teacher evaluation, a few states now require that some part of every administrator's evaluation be based on student test scores, and some districts are experimenting with the use of test scores to determine salaries or bonuses.[4] Some of these policy efforts are misguided and premature since the difficulties in evaluating principals through test scores are even more difficult than they are for teachers,[5] as we will show in this paper. But there's no question that evaluating principals is firmly on the agenda of districts and policy-makers.

The next step from evaluating principals themselves is to examine *preparation programs* for school-level leaders,[6] to see whether they select for and teach the capacities necessary for effective leaders. A few frameworks for evaluating principal preparation

programs exist,[7] using multiple ways of ascertaining effectiveness including examination of the formal curriculum, satisfaction surveys of graduates, stakeholder evaluations, and outcome evaluations, presumably the most important kind. Two large and high-profile programs (the Aspiring Principals Program in New York City and New Leaders for New Schools) have tried to link their efforts to student test scores only, though with distinctly mixed (and often negative) results and many unavoidable flaws in their methods (see Section V).[8] Most recently, the federal Department of Education is funding an effort to see whether a principal preparation program can be "effective" when replicated in a large number of districts and schools. While "effectiveness" has not yet been defined, it would not be surprising, given the proclivities of the Department under Arne Duncan, if effectiveness comes to be measured by simple math and reading scores.

Closer to home, our program — the Principal Leadership Institute (PLI) at the University of California, Berkeley — has been asked by several foundations how we prove our impact, including our effects on student test scores; many foundations have adopted a "results-oriented" stance, and do not intend to fund programs without evidence of "results". Federal grants from the Department of Education require evaluations asking grantees to quantify their effects, something that might be done in several ways but that leads ultimately to test scores and perhaps other student outcomes. At the state level, the California Commission on Teacher Credentialing now requires an assessment" for the preliminary administrative credential, and there's an on-going discussion about the need to "improve" or "monitor" leadership programs, though specific metrics to measure improvement or quality have not yet emerged. There's no question that evaluation and accountability have arrived at our niche within the educational system.

In theory, any leadership program could use student outcomes to evaluate its effectiveness, and we suspect that others have tried. This paper presents our efforts to do so for the PLI. In practice, however, evaluating the effects of principals on outcomes, and then

of leadership programs, proves to be horrendously complex. The existing efforts have not even realized the many problems with doing so, never mind resolved them. Because of these issues, we consider our efforts to evaluate the PLI with student outcomes to be *experimental*, useful for clarifying how difficult it is to use student outcomes for these purposes but not to conclude much about the effectiveness of our particular program. In the process we have discovered many reason that student outcome evaluations for leadership programs are well-nigh impossible, as we summarize in Section V.

The Principal Leadership Institute opened in 1990 and enrolls between 25 and 35 individuals per year, all of whom have a minimum of 3 – 5 years of experience teaching. The program prepares aspiring leaders specifically for urban schools, and consequently stresses the issues of equity and the treatment of low income and racial minority students. PLI students enroll for 14 months (two summers and an intervening school year), have a variety of instructors and coaches who work with them in their schools, and take many problem- and project-based classes as well as participating in four practicums and a variety of other leadership experiences. The program now has almost 500 graduates, most of them in the Bay Area and virtually all of them still in the educational system.[9] It has developed a strong reputation, particularly with four partner districts with which it has particularly close relationships. But reputation is not necessarily a measure of effectiveness, and so we have begun the process of evaluating the program in more formal ways. Our purpose in carrying out these evaluations is strictly to identify strengths and weaknesses in the program, to maintain the strengths and remedy the weaknesses — rather than to respond to any accountability system.

The PLI collects annual data on completion rates in the program, the gender and racial/ethnic composition of graduates, placement rates in administrative positions, the types of positions graduates have, the districts in which graduates work, whether they stay in the Bay Area or not, and whether they remain in urban education, the focus of the PLI.

Such information is purely descriptive, of course, and so we have supplemented these statistics with four different evaluations or impact reports:

• A very early report by Mary Alice Callahan, "A Study of the Principal Leadership Institute at UC Berkeley" (2008), used questionnaires to administrators, both graduates of the PLI and graduates of other programs, about their satisfaction with preparation in four areas — a form of evaluation we call a satisfaction survey. The results revealed high levels of satisfaction among PLI graduates, and consistently lower levels of satisfaction among graduates of other programs. However, while the response rate among PLI graduates was quite high (with 52 responses), the response rate for graduates of other programs 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 432 graduates to date, based on a questionnaire about where they worked and what kinds of school and district positions they occupied. The report also described the various elements of the PLI model.

• Rebecca Cheung's "Alumni and District Partner Feedback" (2013), an example of "stakeholder evaluation", reports the results of alumni surveys and district feedback from the four partner districts: San Francisco, Oakland, Berkeley, and West Contra Costa.

• Norton Grubb and Rebecca Cheung's "Collective and Team Leadership: Preparation for Urban Schools" (2014) is based on interviews with graduates working in schools with two or more PLI alumni. This report revealed the many benefits of having multiple PLI graduates at one school, a practice we have labeled "collective leadership", and also confirmed that many dimensions of the PLI program are valuable on the job.

In this paper we report our efforts to use student outcome data to examine the effectiveness of PLI graduates, usually compared to graduates of other programs. We regard these efforts as experimental because going into this project we weren't sure what kinds of data we could find in our partnership districts. We also failed to appreciate how many

evaluation challenges would come up. This effort was therefore a learning experience, clarifying how difficult it is — indeed, how nearly impossible it is — to carry out a credible outcome evaluation of a leadership preparation program. In addition to data limitations, we were unable to spend large sums on our evaluation since we (like almost all leadership programs) did not have any special grants to carry this out, and in particular we did not have the millions of dollars that some evaluations have cost.[10]

In the first section we review the data in this study. We also describe the sample of administrators we used, and why we were initially restricted the sample as we did. We also note the limitations of these data given the weakness of data collection in California. Then Section II examines test scores, both the California State Test (CST) scores and the state's Academic Performance Index (API); we also examine results according to Similar Schools Rankings, themselves based on API scores. Section III analyzes the patterns of suspension and truancy rates, and Section IV analyzes the California Healthy Kids Survey (CHKS), providing information about student perceptions of teachers, their plans for continuing in school, and their perceptions of belonging. Finally, the concluding summarizes what we learned from this experiment, both about the PLI itself and about the conditions necessary for a more complete evaluation.

I. Data Sources and the Study Sample

The first and most important issue in any outcome evaluation is to determine which educational outcomes can be measured and analyzed, from among the dozens and dozens of consequences we can easily imagine. In our test-besotted age, math and English (really reading) multiple-choice test scores have been the most prominent outcomes, and most evaluations don't include any other results. But obviously there are many more outcomes we would like to know: test scores in other subjects like science, history, and civics or social

studies; measures of progress through school, including rates of grade progression (or of being retained in grade) for elementary and middle schools and of credits completed, college-appropriate courses taken, and graduation rates in high school; measures of attachment to schooling like homework completion and grades, as well as measures of alienation like poor attendance, truancy, and disciplinary actions; educational and occupational aspirations, including plans to attend college; perhaps measures of socio-emotional learning (SEL), which many commentators perceive as crucial both to academic success and to various other life endeavors. But states and districts don't collect this range of information, and such a wide variety of outcomes is usually available only in the special longitudinal surveys conducted by the U.S. Department of Education.[11] In practice, then, we were able to collect data on only three kinds of outcomes: test scores; suspension and truancy rates; and student perceptions measured by the California Healthy Kids Survey.

The data for this report come from a variety of sources. Demographic data about PLI graduates comes from the annual PLI Alumni Survey which contains demographic information on all current and past PLI students as well as detailed work history including school, year, and position. In order to analyze the results of the four educational measures of our study, data was collected from each individual measure's data source. For the CST, the California Department of Education's (CDE) DataQuest website was utilized to collect individual years' CST results dating back to 2002-03. Suspension and truancy rates were also collected utilizing CDE's DataQuest tool under different report filters. Annual API scores and data files were collected directly from the CDE's website under the testing and accountability section. (API files for earlier years differed in formatting from more recent years, and required substantial format standardization.) The final data analyzed were derived from the California Healthy Kids Survey. The raw data for CHKS is not readily accessible to the public. As a result, PLI contacted WestEd and OUSD representatives directly to collect raw data files dating back to 2005-06, the earliest year CHKS data was

available. There is still a great deal of variation in the CHKS data, with the response rates for certain schools varying from year to year (particularly when it was first administered) and also a great deal of inconsistency in the phrasing of questions, questions asked, and scaling of responses. Some subjective standardization was necessary to insure these results were interpretable and comparable from year to year.

Because various factors outside the control of school leadership affect student outcomes, our analysis required the control of several other variables, particularly in statistical models presented below. Fortunately, the yearly API files provide several such variables for every school. They include the number and percentage of socioeconomically-disadvantaged students (as measured by the proportion eligible for free and reduced-price meals, a crude measure of income); the percentage of teachers with full and emergency credentials; the percent of racial minority students in eight different categories; the percentage of students in the Gifted and Talented Education program (GATE); the percentage of migrant students; the percent of English learners, percent of students re-designated as fluent English proficient; and the percent of students with disabilities.

In analyzing the effects of principals, it is crucial to confine the analysis to those in specific districts. Districts vary enormously in their practices: some assign new principals to the most difficult schools; others try to assign the strongest (or most experienced) principals to the weakest schools. They vary substantially in their support for principals, with some providing continuing professional development and induction programs, while others provide no support whatsoever; some even undermine their principals' effectiveness by imposing too many bureaucratic requirements and distracting them from instructional improvement. Therefore the effectiveness of principals may vary according to *district* policy, and it's important to disentangle the effects of individual competency from those of district policies. Different districts may also vary in the kinds of data they have available, so that evaluations using data from different districts — and certainly those using data from

different states, with non-comparable test score measures[12] — are likely to use inconsistent data.

In addition, principals cannot possibly be expected to influence a school in their first or second year; it may take three or more years for a leader to influence a school's teachers, instruction, and climate. We therefore defined the initial study population to be graduates of the Principal Leadership Institute with at least 3 years of principal experience at the same school in one of our large partnership districts, the Oakland Unified School District.[13]

Although there are about 500 PLI graduates, a sample restricted the Oakland District and to principals with three or more years of experience generated a sample of only 24 PLI alumni from the years 2001 through 2013, in seven different cohorts. Thus the restriction of a sample by school district and principal experience reduces the sample size dramatically — another unavoidable data problem, which we will get around in various ways in subsequent sections of this paper.

The average age of the study population was 44 years old; the age distribution is shown in Figure 1. PLI graduates were over 60% female and 54% people of color, with five ethnicities represented. Just under 71% of graduates were principals at elementary schools, 12.5% were in middle schools, and 26.7% were in high schools (see Figure 2). This distribution reflects the fact that the majority of principals are in smaller elementary schools than in larger middle and high schools. Note that the demographic data in Figures 1 and 2 was gathered from the yearly PLI Alumni Survey, which asks PLI graduates to detail their work history including school, year, and position. Such data cannot easily be collected for graduates of other programs without substantial sums of money, so comparisons with non-PLI graduates cannot be carried out — yet another data limitation.

We can use our initial sample of 24 graduates to generate some descriptive statistics, therefore, but 24 observations are clearly far too few for any other kinds of analyses. In

subsequent sections, therefore, we use other samples to examine possible outcomes of the PLI.

Figure 1: Age distribution of our sample

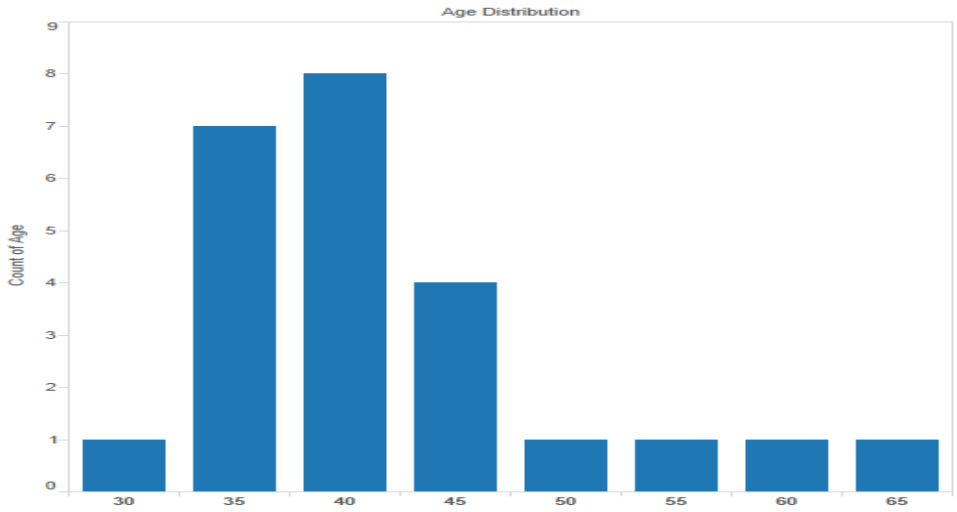
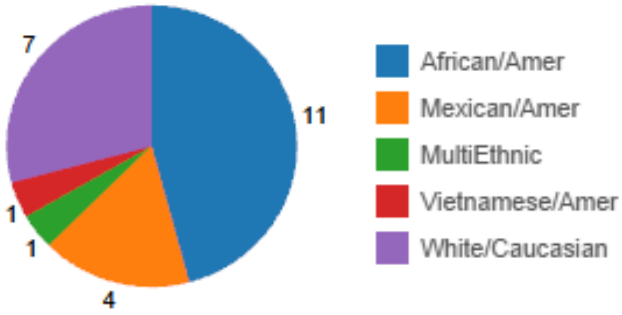


Figure 2: Age, school type and racial/ethnic composition

Group Characteristics	
Average Age	44.04
Male	9
Female	15
School Type	
Elementary	17
Middle	3
High School	4



II. Examining Student Outcomes: Test Scores

California has developed several measures by which it holds schools accountable. Despite their limitations, we use these because they are the only outcome measures available — as is true in virtually all state accountability systems.

California Standardized Tests (CST)

The CST is one component of California's annual Standardized Testing and Reporting (STAR) program. The CST assesses students in a variety of subjects, and then categorizes results under six basic reporting measures: percentage of students far below basic level, below basic, at basic, at or above proficient, proficient, and advanced. These categorizations have the advantage of being readily interpretable, even though the methods for determining proficiency are both opaque and arbitrary.

For our analysis, we focused on CST tests in both language arts and mathematics. We decided to further simplify the results into three categories: percentage of students far below or below basic, at basic, and at or above proficient. We started our analysis by comparing those schools with PLI leadership and those without, but because of the need to create a larger sample we were forced to ignore the experience levels of principals. We ended up with about 9,200 observations for PLI and non-PLI schools, for the years 2002-03 until 2012-13 though sample sizes vary somewhat because of missing data. The CST results for this expanded sample are as shown in Table 1, with a simple average of annual results in the last column.[14]

In these results, schools with PLI leadership have fewer students than schools with non-PLI leadership in the "below or far below basic" category for both math and language arts (8% and 5% differences respectively) while simultaneously having more students in the "at or above proficient" category (9% and 7% differences respectively). From these results, then, PLI graduates are in "better", or at least higher-performing, schools. Of course causality cannot be determined from these simple results. It may be that PLI graduates improve CST scores more than other leaders do. Alternatively, the results in Table 1 may simply reflect the assignment practices of the Oakland school district: If PLI graduates are assigned to schools with *lower* proportions of low-income and racial-minority students, then

their student CST scores would be higher, as in Table 1, simply because of these assignment practices. In the section below on Similar Schools Rankings, in Figure 5 and Table 2, we examine these possibilities.

Table 1: CST scores for math and language arts

CST - Math PLI Present	2002-03	2003-04	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-12	2012-13	Average
Below or Far Below Basic	60.75%	32.00%	42.38%	31.24%	34.66%	30.43%	24.21%	16.97%	20.94%	20.64%	18.94%	26.66%
Basic	30.13%	26.20%	22.31%	23.10%	22.45%	23.48%	21.62%	21.25%	19.41%	20.29%	20.36%	21.85%
At or Above Proficient	9.00%	41.60%	35.41%	45.76%	43.03%	46.16%	54.20%	61.84%	59.77%	59.18%	60.75%	51.56%
PLI Not Present												
Below or Far Below Basic	48.10%	46.74%	42.68%	42.45%	40.80%	33.63%	27.72%	20.61%	18.27%	21.12%	24.70%	34.14%
Basic	24.74%	26.28%	24.58%	23.91%	24.24%	24.20%	23.39%	22.34%	21.46%	21.93%	23.76%	23.80%
At or Above Proficient	27.23%	27.01%	32.79%	33.67%	35.04%	42.26%	49.05%	57.13%	60.30%	56.98%	51.55%	42.11%

CST - Language Arts PLI Present	2002-03	2003-04	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-12	2012-13	Average
Below or Far Below Basic	53.17%	31.32%	35.52%	33.11%	37.10%	34.57%	29.75%	25.47%	24.49%	24.71%	18.51%	29.51%
Basic	35.00%	27.26%	29.29%	26.78%	28.59%	29.35%	28.31%	27.78%	28.24%	28.40%	26.35%	28.24%
At or Above Proficient	11.67%	41.42%	35.35%	40.20%	34.46%	36.18%	42.10%	46.74%	47.39%	46.77%	55.19%	42.32%
PLI Not Present												
Below or Far Below Basic	44.42%	45.54%	40.77%	40.50%	39.37%	35.54%	31.56%	27.99%	25.60%	23.58%	24.32%	34.32%
Basic	31.97%	32.04%	31.95%	31.23%	30.99%	31.48%	29.68%	28.87%	28.36%	28.78%	30.54%	30.51%
At or Above Proficient	23.69%	22.50%	27.37%	28.36%	29.74%	33.07%	38.84%	43.14%	46.08%	47.68%	45.21%	35.24%

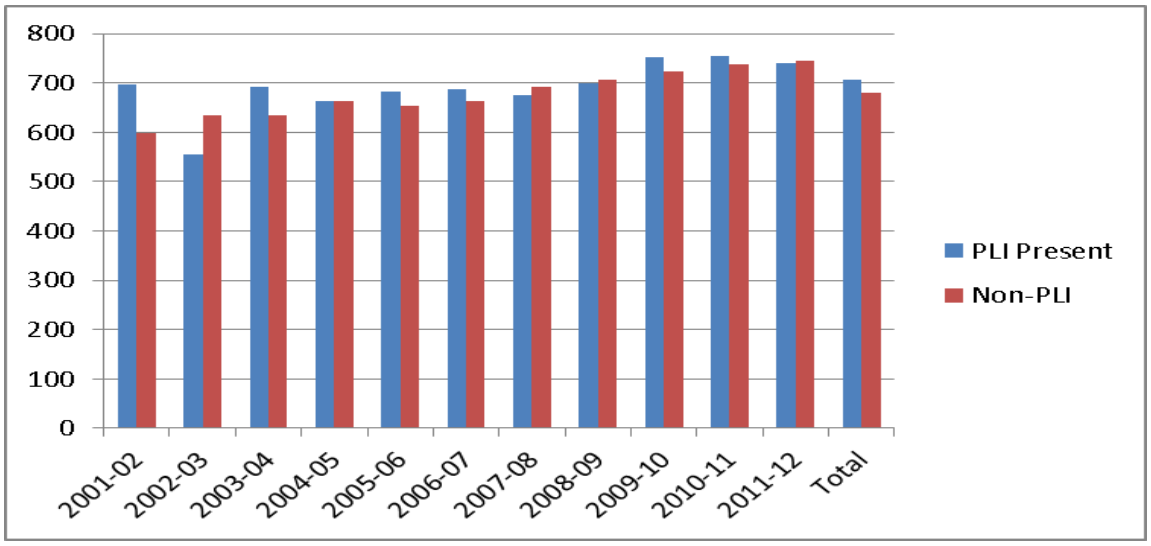
The California Academic Performance Index (API)

The California Academic Performance Index is a more complicated measure than the CST. The state legislature intended the API to be based on multiple measures of outcomes, including graduation rates, teacher attendance and performance, and other reliable measures. In practice, however, the API is based entirely on STAR tests, including CST outcomes and results from the California High School Exit Exam (the CAHSEE). These are combined and transformed into a scale between 200 and 1000.

The API measures a school's *level* of performance for each year, called the base API score. The California Department of Education also calculates a mis-described API growth score, examined below. Base scores and growth scores are contained in separate data files, with each containing roughly 1,400 relevant records since the 2000-01 academic year. The data for this report were collected in fall 2013, with the most recent base API data available detailing results for school year 2011-12.[15] Unlike CST scores, the API base scores have no intuitive meaning, and again their construction is both arbitrary and opaque.[16]

As Figure 3 shows, PLI graduates are typically at schools with slightly higher API scores than graduates of other programs, confirming the results in Table 1. These differences prove to be statistically significant. However, they may have nothing to do with principals and other leaders, but may instead reflect differences in the family backgrounds of students, and their preparation levels prior to entering a school; they may reflect the backgrounds,

Figure 3: API Base Scores, by Year and PLI/Non-PLI Schools



experience, and instructional approaches of teachers; they may be due to other dimensions of the school like its climate and culture, the support it provides students, levels of trust within the schools, and other factors that are only partially under the control of principals. In order

to adjust for these potential influences, we have used a relatively simple statistical technique — regression analysis — to ascertain which of these other variables might influence CST scores. Of course, we (like all other evaluators) are limited by the available data, and so can use only those variables mentioned in Section I, again based on state data for schools. These have very complete data on the racial/ethnic composition of students, reflecting the American preoccupation with race; the data on other dimensions of student backgrounds are much more limited, as is information about teachers; and other dimensions of schools are completely missing.[17]

The simplest regression we can run is presented in Appendix Table 1, where the dependent variable is the school's API score. All of the results are consistent with what virtually all researchers have found: Scores are much higher in schools with higher-income students; they are also higher where a school has a higher proportion of students in gifted (GATE) programs. Scores are lower in schools with high proportions of African Americans, Latinos, and Pacific Islanders, high proportions of English Learners, high proportions of special education students, and high proportions of students designated as fluent English proficient (RFEP), a measure of English Learners who have attained fluency in English. Finally, scores are highest in schools with more fully-credentialed and emergency-credentialed teachers, compared to those with teachers lacking credentials (like full-time substitutes). But the most important result for our purposes is that once all these variables have been considered, the difference associated with having a PLI leader is statistically insignificant — that is, it is too small to be reliably considered positive or negative.

However, these results describe *levels* of test scores, affected by many aspects of schools aside from their leadership. This is why many analysts (and policy-makers following them) have recommended using test score *growth*, a measure of value added, and an outcome that some think is more likely to be related to what schools do rather than outside influences. The wrong way to do this is to examine differences in test scores from one year to the next, or TS_t

– TS_{t-1}, as a function of other independent variables; the right way is to add test scores from prior years to the independent variables in Appendix Table 1.[18] When we do this by including test scores from three years before, we get the results in Appendix Table 2. Now the coefficient associated with having a PLI leader is even smaller and much less statistically significant than in the previous regression. The coefficient on the lagged test scores is .68, between zero and one as expected, and it has by far the greatest explanatory power of any of the independent variables. But it is much smaller than one, indicating that prior tests scores influence subsequent test scores, as everyone should expect, but that a one point difference on prior test scores leads to less than a one point increase on current tests. The other results in this table are much the same as before, except that the effect of the percent Latino is now insignificant.

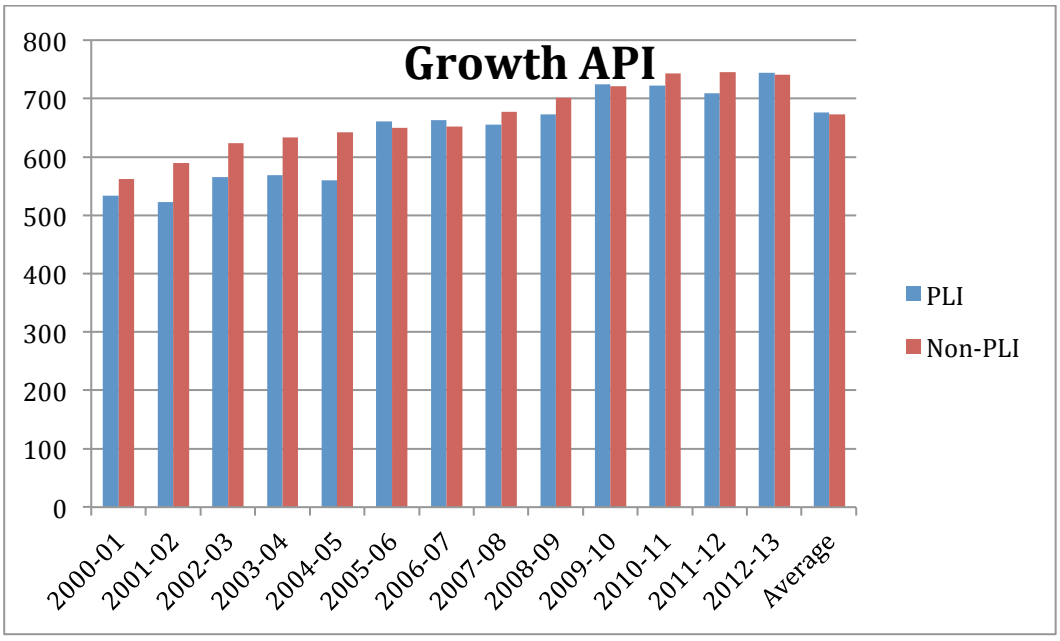
What we have learned from these analyses is that, while there may be some differences in API scores between schools with and without PLI leaders, these differences vanish when we consider other factors. PLI graduates might have a positive influence on students' academic learning, but we cannot detect it with these data.

In addition to API base scores, California also provides API growth scores. Although these are commonly referred to as *scores*, they are really *targets*: The state would like all schools to achieve a base API score of 800, and for all schools to make annual progress of at least 5% of the difference between 800 and the API base score. The API base score plus this desired increment is then API growth, or the API growth target. The California Department of Education recommends that school districts examine API growth scores to see their progress, even though CST tests are not constructed to compare results over time. (The API growth scores are also the basis for the federal Average Yearly Progress measures under No Child Left Behind.) Because the API growth scores are a minor transformation of API base scores, we should not expect the results of analyzing them to differ very much.

Figure 4 presents API growth scores, by year, for schools with and without PLI leaders. Compared to the results in Figure 3 above, the differences are even smaller: PLI schools have lower growth scores in the earlier years and the later years but similar scores in the middle years, and the averages over these 13 years are the same. Based on these results, it doesn't seem to make much difference whether a school is headed by a PLI graduate or not.

As we did for API base scores, we can further examine API growth scores with a multivariate regression analysis. Appendix Table 3 presents the coefficients from such a regression. As expected, the results are similar to those in Appendix Table 1 for API base year scores.[19] The coefficient for PLI schools is negative but insignificant; growth scores are higher in schools with higher proportions of full- and emergency-credential teachers, compared to teachers without any credentials; schools with more high income students have higher growth scores; schools with high proportions of African American and Pacific

Figure 4: API Growth Scores, by Years and PLI/Non-PLI Schools



Islanders have lower growth scores, while those with more Asian-American students have higher scores; schools with more GATE students have higher scores, while those with more

English Learners, RFEP, and special education students have lower scores. There aren't any surprises here, and as expected the results parallel those of API base scores. Like the earlier results, these fail to distinguish any differences between PLI and non-PLI schools.

The Similar Schools Rankings (SIM Rankings)

Another transformation of base API scores carried out by the state is to group schools into Similar School Rankings, in an attempt to provide comparisons among similar schools. Otherwise the base and growth API scores for the state as a whole include variation from suburban schools with high proportions of white and Asian students, and few low-income students — and typically with high test scores — to urban schools with concentrations of low income and racial minority students, typically with low test scores. The schools at each rank are determined by a range of socioeconomic indicators such as percentage of students eligible for free or reduced meals, percent of English learners, and percent of racial minorities among others, in an attempt to compare schools with similar demographic characteristics — though again the methodology for doing this is both opaque and arbitrary.[20] Then the Similar School Rankings place 10% of each type of school (elementary, middle, and high school) at ranks 1 - 10 (1 being lowest, 10 being highest).

When we examine schools by their SIM rankings, our earlier conclusions about PLI placements change substantially. Figure 5 indicates that PLI students are more likely to be in lower-ranked schools, and less likely to be in higher ranked schools, compared to schools without PLI leaders. Of the PLI alumni in our study, 42% were placed at SIM rank 1 schools, 17.6% at SIM rank 2 schools, and another 9.2% at SIM rank 3 schools, for a total of over 68% of PLI alumni being placed at schools with the most disadvantageous school demographics. This is a higher proportion than the typical distribution of OUSD's SIM ranks across schools without PLI graduates, of whom 47% are in ranks 1 – 3.

We can also use these SIM rankings to look at CST scores not for all schools, as we did in Table 1, but for demographically similar schools. For example, Table 2 presents CST scores only for SIM Rank 1 schools, for comparison to Table 1 above for all schools. Remember that we speculated that the results in Table 1 might be due simply to the district's assignment practices — to the possibility that if PLI graduates are assigned to schools with *lower* proportions of low-income and racial-minority students, then their CST scores will be higher (as in Table 1) simply because of these assignments. However, the same overall patterns emerge in Table 2: students in SIM 1 schools with PLI graduates are still less likely to score below or far below basic, and more likely to be at or above proficient, compared to students in other SIM 1 schools (though the differences are smaller in Table 2). Even when we take account of the demographic conditions that affect school outcomes so powerfully (as Appendix Tables 1 and 2 show), the differences in CST scores among schools in Table 1 persist.

Figure 5: Principal assignment by SIM Rank, PLI vs. Non-PLI

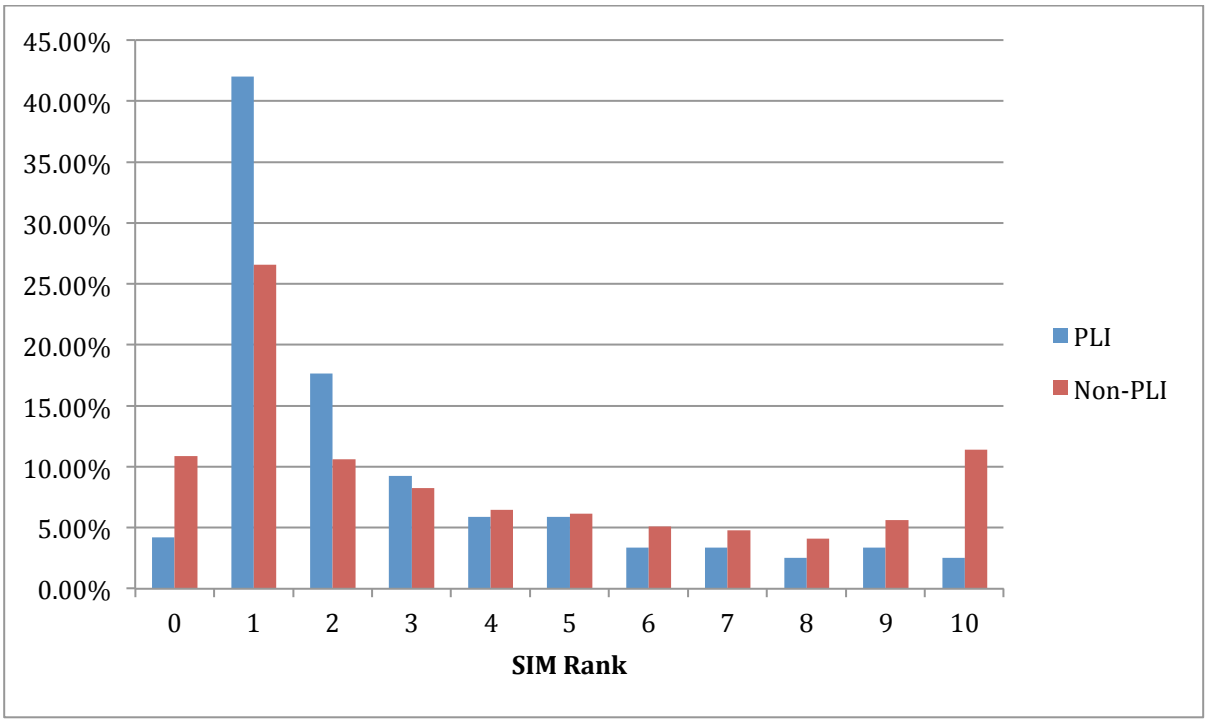


Table 2: SIM Rank 1 CST Scores (Math and Language Arts Combined)

CST Scores - SIM Rank 1 PLI Present - 362 Obs.	2002-03	2003-04	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-12	Total
Below or Far Below Basic	n/a	53.00%	70.44%	66.00%	51.03%	44.97%	41.11%	30.94%	26.79%	29.37%	36.78%
Basic	n/a	31.00%	23.56%	16.83%	28.19%	29.60%	29.77%	30.67%	26.61%	30.42%	28.78%
At or Above Proficient	n/a	16.14%	6.00%	17.00%	20.75%	25.48%	29.17%	38.46%	46.71%	40.25%	34.49%
Number of Observations											
PLI Not Present - 1768 Obs.											
Below or Far Below Basic	62.93%	63.45%	56.66%	56.84%	51.68%	42.89%	42.37%	32.78%	30.11%	33.21%	46.44%
Basic	25.81%	25.77%	26.95%	27.46%	28.58%	28.20%	28.69%	26.73%	27.10%	28.60%	27.45%
At or Above Proficient	11.36%	10.79%	16.40%	15.76%	19.83%	28.95%	29.07%	40.51%	42.85%	38.29%	26.17%

Overall, the analysis of test scores in three forms — in CST scores, in API scores, and in CST scores for schools by Similar Schools Rankings — show that test scores in schools with PLI leaders are higher than those in schools with leaders from other programs. However, the causality of this relationship is difficult to disentangle. When we try to do so in regression analyses that include the effects of various student and teacher variables, the test score differences between PLI and non-PLI schools disappear, and other demographic and teacher factors are much more powerful in explaining API scores. So we must be prepared for finding that the differences among preparation programs are in practice overwhelmed by more powerful differences in student and teacher characteristics.

III. Examining Student Outcomes: Suspension and Truancy Rates

Unfortunately, the state data on which we must rely contains very few other outcome measures. However, the state does collect information on suspension and truancy rates for each school. These are important because student suspensions lead to students being out of school for varying periods of time, which is never good for academic progress. Furthermore, given the patterns of suspension rates (further analyzed below and in Appendix Table 4) — which tend to be higher for low-income and racial minority students — the effects of

suspensions on time spent in school are particularly harmful to those students who can least afford to be out of school.

For this reason, many schools have tried to reduce suspension rates, for example by using Positive Behavior Support (PBS) and Restorative Justice, which are less likely to result in suspension. Along the same lines, the PLI teaches its students about these alternative disciplinary methods, and we would hope that suspension rates are lower in schools headed by PLI graduates.

Along the same lines, high truancy rates — measures of chronic absences[23] — have also been linked to poor levels of preparation for further schooling and then to dropout rates, so their consequences are dire. The question of what is responsible for truancy is a matter of debate: the California State Department of Education guidelines outline penalties for *students* and *parents*, but it fails to recognize that *schools* may be responsible for poor attendance if teachers are demeaning or unsupportive, or if the school is dangerous, or if the school culture is hostile to some students (including racial minority students). As a result strong principals work to reshape schools in order to minimize truancy rates as well as suspension, and again we would hope that suspensions and truancy rates are lower in schools lead by PLI graduates.

We analyzed suspension and truancy rates for 2004-05 (the earliest year available) through 2012 – 2013, with the results presented in Table 3. (As in Section II, we are forced by the need for a larger sample size to use information from OUSD schools regardless of the experience levels of their principals.) From these data, it appears that PLI-led schools have *higher* suspension rates and truancy rates than other schools over the 8-year period studied, though the patterns are somewhat uneven by year. The differences on average are greater for suspension rates than for truancy rates.

Table 3: Suspension and Truancy Rates

Suspension Rates	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-12	2012-13	Grand Total
With PLI	22.22%	21.47%	15.39%	5.68%	17.86%	11.41%	16.47%	8.63%	3.41%	**12.96%**
W/O PLI	10.34%	15.00%	15.53%	4.38%	14.38%	14.72%	11.25%	7.23%	5.78%	**10.85%**
Truancy Rates	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-12	2012-13	Grand Total
With PLI	34.99%	45.32%	50.09%	47.73%	36.15%	30.09%	47.34%	55.10%	37.84%	**43.33%**
W/O PLI	28.16%	34.72%	49.71%	46.63%	41.73%	26.19%	41.42%	40.33%	48.08%	**39.73%**

Once again, however, the causality behind these data is unclear. It may be that PLI principals are less successful than other leaders in reducing suspension and truancy. On the other hand, we also know that over 40% of PLI graduates were placed at SIM rank 1 schools and 69% are in SIM Rank 1 – 3 schools. We therefore decided to take a closer look at the effects of these rankings by calculating truancy and suspension rates only in schools ranked at the lowest SIM level. The results of these calculations are presented in Table 4. Indeed, this analysis changes the results: now both suspension and truancy rates are slightly *lower* on average in schools with a PLI leader (though again there are irregularities from year to year).

Just as we did for API scores in Appendix Tables 1 and 2, we can estimate regression equations for suspension and truancy rates, including a variety of other factors that might affect them. The results are presented in Appendix Table 4 for suspension rates, and in Appendix Table 5 for truancy rates, in "value-added" specifications that include the suspension and truancy rates three years earlier. Once we consider these other factors, the differences between PLI and non-PLI schools in Appendix Table 4 for suspension rates

Table 4: SIM Rank 1 Suspension and Truancy Rates

Suspension Rates - SIM Rank 1	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-12	Grand Total
With PLI	8.11%	41.31%	12.42%	7.10%	21.61%	11.53%	13.28%	9.10%	**11.94%**
W/O PLI	15.17%	21.03%	19.03%	4.09%	17.45%	17.31%	13.23%	9.41%	**14.75%**
Truancy Rates - SIM Rank 1	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-12	Grand Total
With PLI	54.28%	56.81%	29.87%	46.16%	41.64%	28.22%	28.63%	45.87%	**38.46%**
W/O PLI	35.21%	45.70%	52.79%	48.50%	43.65%	29.92%	34.80%	34.88%	**40.75%**

are insignificant, both statistically and practically. Instead, suspension rates are higher where there are more teachers with full credentials, an odd result we cannot explain; where there are more low-income students; and where there are more African American students, a result consistent with many findings that African American boys in particular are likely to be more severely disciplined, even for offenses for which other students are not. As is generally true in such regressions, the lagged dependent variable has quite high explanatory power.

For truancy rates in Appendix Table 5, the effect of a PLI leader is again insignificant. There are higher truancy rates in schools with more African Americans, more low-income students, and more students re-designated fluent English Proficient. The lagged dependent variable is again highly significant.

Finally, the coefficients on the lagged dependent variable for suspension rates, .40, and for truancy rates, .49, are substantially lower than for API scores in Appendix Table 2, .68. This means that prior patterns have smaller effects for suspension and truancy than for test scores, which are more resistant to change. This finding implies that school policies may be more effective in modifying student behavior, reflected in suspension and truancy rates, than in changing test scores, something that requires difficult changes in teachers and instructional methods.[22]

As with test scores, then, any differences in suspension and truancy rates that might be due to principals (like those in Tables 3 and 4) are in practice overwhelmed by the other characteristics of students and teachers among schools.

IV. Results from the California Healthy Kids Survey

So far the outcome measures we have examined come from the California Department of Education. A different source of information, the California Healthy Kids Survey (CHKS), is described by its distributor WestEd as "the largest statewide survey of

resiliency, protective factors, and risk behaviors in the nation." The CHKS' goal is to better understand students' mental health and overall well-being through a detailed questionnaire regarding students' perceptions of belonging at school, their perceptions of school leadership and its impact on their performance, and various factors regarding their families and home lives. (In many ways the CHKS is the California equivalent of the student perception surveys now being tried out in a number of other places including Colorado, Georgia, Chicago, Denver, Memphis, and Davis, Utah). Evidently schools with negative responses from students might be poor places for them to learn, and may require some intervention by administrators or outsiders, like school improvement teams. In addition, negative responses on family and home life may signal external problems that hamper student learning, though most schools are not well-equipped to handle these issues (except perhaps when a comprehensive School-Based Health Center exists).

For the purposes of our study, the CHKS serves as a qualitative measure of school leadership's impact on students outside of traditional measures. To start, we selected 14 questions that appear consistently on the survey as measures of school culture, something that a principal can significantly affect. We were able to use 32,733 records for students in the Oakland school district. Table 5 shows the results for the questions of interest for grades 5 through 12.

Many responses in Table 5 are quite similar for PLI and non-PLI schools. However, the responses to whether "I plan to graduate from high school" and "I know where to go for help with a problem" are much more positive for PLI-led schools. In addition, PLI-led schools have slightly more positive responses to all the questions in Table 6, though the differences are small in some cases.

Another revealing way to examine the CHKS data is to focus on SIM Rank 1 schools, those with the most challenging demographic compositions and students who are conventionally thought to require more support from teachers and leaders. In Table 7 we see

PLI-led schools faring better across every question, as shown by the higher proportion of students responding "very much true". The differences are particularly striking for the responses about teachers and others who really care about students, and teachers who listen when students have something to say.

Table 5: Student Perceptions of Belonging and Support

	PLI			
Question	Not at all true	A little true	Pretty much true	Very much true
There is a teacher or some other adult who really cares about me.	8.85%	27.84%	28.77%	**34.53%**
There is a teacher or some other adult who tells me when I do a good job.	22.37%	36.68%	5.87%	35.08%
There is a teacher or some other adult who notices when I'm not there.	12.81%	28.91%	32.06%	26.22%
There is a teacher or some other adult who always wants me to do my best.	5.60%	16.28%	33.46%	44.66%
There is a teacher or some other adult who listens to me when I have something to say.	8.70%	25.36%	33.47%	32.47%
There is a teacher or some other adults who believes that I will be a success.	9.56%	20.79%	31.90%	37.75%
At school I do things that make a difference.	29.26%	34.52%	21.49%	14.73%
I plan to graduate from high school.	3.01%	4.91%	8.40%	**83.68%**
I know where to go for help with a problem.	5.63%	11.82%	23.07%	**59.49%**
	Non-PLI			
	Not at all true	A little true	Pretty much true	Very much true
There is a teacher or some other adult who really cares about me.	9.45%	27.45%	31.27%	**31.83%**
There is a teacher or some other adult who tells me when I do a good job.	7.03%	21.09%	35.74%	36.13%
There is a teacher or some other adult who notices when I'm not there.	12.07%	25.10%	32.78%	30.05%
There is a teacher or some other adult who always wants me to do my best.	6.76%	15.98%	31.21%	46.04%
There is a teacher or some other adult who listens to me when I have something to say.	9.23%	23.86%	34.05%	32.86%
There is a teacher or some other adults who believes that I will be a success.	9.18%	19.28%	31.99%	39.56%
At school I do things that make a difference.	27.80%	33.89%	23.26%	15.05%
I plan to graduate from high school.	5.23%	5.58%	11.58%	**77.61%**
I know where to go for help with a problem.	7.65%	16.03%	25.14%	**51.17%**

Table 6: Student Feelings about their Schools

	PLI		
Question	Agree or Strongly Agree	Neither agree nor disagree	Strongly Disagree or Disagree
I feel close to people at this school.	56.25%	28.96%	14.78%
I am happy to be at this school.	59.68%	26.97%	13.35%
I feel like I am part of this school.	57.20%	26.91%	15.89%
The teachers at this school treat students fairly.	52.36%	29.31%	18.33%
I feel safe in my school.	50.28%	29.60%	20.13%
	Non-PLI		
	Agree or Strongly Agree	Neither agree nor disagree	Strongly Disagree or Disagree
I feel close to people at this school.	54.95%	28.30%	16.75%
I am happy to be at this school.	56.55%	27.51%	15.94%
I feel like I am part of this school.	54.28%	28.04%	17.68%
The teachers at this school treat students fairly.	52.06%	28.16%	19.78%
I feel safe in my school.	47.92%	31.61%	20.47%

Similarly, in Table 8 reporting on student feelings in SIM Rank 1 schools, the responses are again more positive for PLI-led schools, with especially large differences for being happy at school, for feeling part of the school, and for feeling safe. (The differences in both Table 7 and 8 are statistically significant according to Chi-square tests.) Unlike the smaller differences in Tables 5 and 6, these results cannot be due to the greater assignment of PLI leaders to SIM Rank 1 schools, as we saw in Figure 5.

Table 7: Student Perceptions of Belonging and Support, for SIM Rank 1 Schools

SIM Rank 1	PLI			
Question	Not at all true	A little true	Pretty much true	Very much true
There is a teacher or some other adult who really cares about me.	8.30%	22.11%	27.45%	**42.14%**
There is a teacher or some other adult who tells me when I do a good job.	5.72%	18.82%	33.60%	**41.86%**
There is a teacher or some other adult who notices when I'm not there.	12.01%	27.30%	29.80%	30.89%
There is a teacher or some other adult who always wants me to do my best.	5.35%	13.21%	27.20%	**54.25%**
There is a teacher or some other adult who listens to me when I have something to say.	8.27%	23.63%	31.11%	36.99%
There is a teacher or some other adults who believes that I will be a success.	9.61%	18.43%	30.08%	41.89%
At school I do things that make a difference.	27.81%	33.13%	21.25%	17.81%
I plan to graduate from high school.	1.73%	8.23%	12.12%	**77.92%**
I know where to go for help with a problem.	4.65%	11.70%	21.31%	**62.34%**
	Non-PLI			
	Not at all true	A little true	Pretty much true	Very much true
There is a teacher or some other adult who really cares about me.	9.85%	26.76%	28.42%	**34.97%**
There is a teacher or some other adult who tells me when I do a good job.	7.66%	21.43%	33.33%	**37.58%**
There is a teacher or some other adult who notices when I'm not there.	12.96%	24.59%	32.70%	29.75%
There is a teacher or some other adult who always wants me to do my best.	7.81%	16.47%	30.56%	**45.15%**
There is a teacher or some other adult who listens to me when I have something to say.	9.23%	25.57%	31.87%	33.32%
There is a teacher or some other adults who believes that I will be a success.	10.77%	19.87%	31.77%	37.58%
At school I do things that make a difference.	29.85%	33.45%	21.68%	15.02%
I plan to graduate from high school.	6.98%	6.59%	13.18%	**73.26%**
I know where to go for help with a problem.	6.39%	16.20%	19.90%	**57.51%**

Table 8: Student Feelings about their Schools, for SIM Rank 1 Schools

SIM Rank 1	**PLI**		
Question	Agree or Strongly Agree	Neither agree nor disagree	Strongly Disagree or Disagree
I feel close to people at this school.	58.78%	28.65%	12.57%
I am happy to be at this school.	66.06%	23.33%	10.60%
I feel like I am part of this school.	63.98%	21.72%	14.30%
The teachers at this school treat students fairly.	58.94%	26.23%	14.83%
I feel safe in my school.	57.12%	25.20%	17.68%
	Non-PLI		
	Agree or Strongly Agree	Neither agree nor disagree	Strongly Disagree or Disagree
I feel close to people at this school.	53.96%	29.45%	16.60%
I am happy to be at this school.	56.88%	27.13%	15.99%
I feel like I am part of this school.	54.79%	26.61%	18.60%
The teachers at this school treat students fairly.	53.78%	27.03%	19.19%
I feel safe in my school.	46.68%	31.34%	21.98%

The Principal Leadership Institute emphasizes preparing its graduates for work in urban schools. To further equity and social justice, it stresses that leaders must help correct the conditions that are so common in urban schools — inexperienced and demeaning teachers, harsh discipline policies, forms of class and racial discrimination. Furthermore, our graduates agree that these dimensions of the PLI are important in the positions they fill.[23] We like to think, therefore, that the differences in student responses apparent in Tables 7 and 8 are due to the preparation students receive in the PLI.

We should say something about the magnitudes of student responses in these four tables, since they reflect some discouraging aspects of schools in this district. Consistently, about one-third of students respond very positively to the questions in Table 5, and another third respond that these questions are "pretty much true". But this leaves a third of students whose agreement is tepid at best, and often negative. The agreement is particularly low for the statement "At school I do things that make a difference", where only 15% of all students respond most positively. Evidently, there is room for teachers and leaders, as well as parents and community leaders, to convince students about the importance of schooling.

Finally, all educators should address one particular response in these CHKS results. In Table 5, 84% of students in PLI schools and 78% of those in other schools say they plan to graduate from high school; the figures are only slightly lower (78% and 73%) for those in SIM 1 schools. It's wonderful to see students with high aspirations, especially since almost no one receives more schooling than they aspire to achieve. Unfortunately, these aspirations are not being met: the most recent figures indicate that 81% of all students complete high school, but this figure is much lower for Latinos at 76%, African Americans at 68%, low-income students at under 72%, English learners at under 59%, and special education students at 61%.[24] The Oakland Unified School District has high proportions of those students who have historically had low graduation rates, and the aspirations reflected in Table 5 and 6 are likely

to be overly optimistic. The challenge, to all teachers, leaders, districts, state and federal officials, and to preparation programs like PLI is to make that these dreams can come true.

V. Some Conclusions: The Limited Possibilities for Evaluating Principal Preparation with Student Outcomes

At the end of our investigations with student outcomes, we have found only a little evidence that the Principal Leadership Institute does a better job of preparing school leaders than other programs serving the Oakland Unified School District. Instead, when we examined tests and their variants calculated by the California Department of Education, we found that other factors usually associated with academic success — parental income, the race/ethnicity of students, the concentration of English learners and disabled students, the credentials teachers have — account for all of the variation among schools we could explain. These results are similar to those for the other outcome studies we know about.[25]

To be sure, data from the California Healthy Kids Survey, with very different information about students' perceptions of their schools, provide some evidence that students in schools led by PLI graduates have more positive conceptions of their schools as safe and supportive places. Whether these attitudes translate into better outcomes remains unclear, however, in the absence of demonstrated links between student perceptions and outcomes.

Overall, it's possible that principals, and therefore principal preparation programs, are not powerful enough to prevail over the strong influences of family background, race/ethnicity, other student characteristics like being an English learner or disabled, and simple measures of school resources like teacher credentials. We doubt this, however: analyses of student outcomes with much more detailed data reveal that dozens of school resources affect a variety of outcomes, and many of these resources can in theory be influenced by the actions of principals and leadership teams.[26]

Instead, we suspect that data and methodological problems make it impossible to carry out the evaluations necessary to ascertain whether leaders, and then leadership preparation programs, have powerful effects. In particular, there are at least eleven problems that cannot be resolved, only some of which we have mentioned so far:[27]

• The need to confine analysis to leaders within one district, and to principals with several years or more of experience, reduces the sample size dramatically, making it more difficult to find significant effects. This means that all programs like the PLI, which prepare graduates for a variety of local districts, are difficult to evaluate. Only where a program works with one large district — for example, programs preparing leaders for New York, Chicago, or Miami-Dade — can there be large sample sizes.

• Just as principals need some experience in a school to have any effect, students need some experience with a principal for him or her to have any effect. Even an experienced principal is unlikely to have an effect on a student who has just moved into the school, since the influences of prior schools will be much greater. We might surmise, then, that it requires principals with experience in a school *and* students with tenure under that principal for there to be much effect. Showing this requires data that link principals and their experience with students and their tenure in schools — data that are almost impossible to get, and that would require an enormous sample. The only study to examine both principal experience and student tenure found different results for each. It seems to be more important to measure student tenure with a principal than principal experience in the school,[28] implying that both forms of experience may be important.

• The outcome measures available are typically highly restricted, usually to test scores. Furthermore, in California and most other states, test score measures are complex, unintuitive, and opaque, even though they are widely cited as evidence about outcomes; and test scores are rarely vertically equated (that is, adjusted so that one point on a test in one year means the same as one point on a test in another year). Therefore test scores cannot

legitimately be used to measure changes over time or construct value-added measures. Sometimes other outcomes measures are available — like the suspension and truancy rates examined in Section III, or a few other measures like dropout rates in some states, or rates of passing high school exit exams. The lack of variables measuring normal passage through school is particularly serious because test scores and progress are to some extent different, and are affected by quite different school resources.[29]

In the absence of good state outcome measures, the only alternative is for evaluators to collect data themselves from principals and schools, and this task is both overwhelming and expensive. Most leadership programs are small and cannot possibly afford to pay for such data collection; although there are some expensive studies, they are rare and require external funding.

• There are many, many influences on student outcomes in addition to school leaders and the ways they can influence instruction and climate, and these need to be considered (or controlled) before concluding that principals and other leaders are responsible for outcomes. In particular, every study of student outcomes has found that family background is among the most powerful predictors of virtually every outcome. But the only family background measure usually available is eligibility for free and reduced-price school lunch, an extremely crude measure of family income; other dimensions of family background, like parental education, occupation, and aspirations for their children, are unavailable except in large longitudinal data sets. Without these measures, principals in high-SES schools will always look more effective than those in low-SES schools, and accountability systems will tend to drive principals from low-performing to high-performing schools.

• Schools have different levels of district support. In Oakland (and many other districts), low-performing schools are identified with a system of tiers, rating and then color-coding schools according to test scores, growth in test scores, and progress in closing the achievement gap. Those schools with red, orange or yellow codes are then targeted for

special district support and monitoring. This implies that school-level outcome evaluations must have information on district practices, in addition to teacher, student, and school characteristics.

• Just as value-added measures for teachers vary substantially 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. For example, PLI selects students on the basis of their experience in and commitment to urban schools and equity, while others are non-selective. But no evaluation has managed to account for selection and self-selection, and so they cannot distinguish between a program's *effectiveness* and its *selectiveness*.

• Probably the most serious problem is that the model we have used in the Appendix Tables, and also employed by all other efforts to evaluate principals and preparation programs, is the wrong one. Principals don't affect students directly; teachers do that through their instruction and relations with students. Instead, principals and other leaders affect students *indirectly*,[30] through their effects on teachers including teacher selection and firing, evaluations, professional development, coaching, work on data analysis and its use in improving instruction; through their effects on school climate; sometimes in their influences on the curriculum used and reforms adopted (though these are highly restricted in many districts); in efforts to create linkages to parents and community organizations, and sometimes to ancillary services like School-Based Health Centers; and through their success in capturing additional resources from districts (a zero-sum game that may reduce the resources available to other schools) or from foundations (usually short-term). Then these kinds of influences, or resources broadly defined, might in turn affect student outcomes.

This means that the right way to model the effects of principals is to estimate two sets of equations, one reflecting the influences of principals on school processes and resources, and the second to analyze the effects of these resources plus student characteristics on student outcomes. But if instead the two equations are combined (as in our appendix tables and all other analyses) and an evaluator simply estimates the effects of some school resources including principals on outcomes, then the coefficients (like those in our appendix tables) are almost certain to be biased.[31] The implication is that we can't use the conventional model to ascertain the effects of principals or principal preparation programs on student outcomes.

• In some (but not all) schools, the conventional use of outcome measures to determine the influence of principals mistakes the very nature of leadership. The conventional approach is to assume that leadership is a function of the principal only, who is the director of a school in a conventional top-down hierarchy. As an example, the VALED approach to principal evaluation asks various stakeholders about the characteristics and actions of the principal only. But in many schools the model of a single leader, or hero-principal, is inappropriate. This happens, for example, in high schools with a leadership team; or schools following the approach of distributed leadership with leadership responsibilities shared among principals, assistant principals, teacher-leaders, and coaches; or schools with collective leadership based on the similar backgrounds and values of several leaders.[32] Under these conditions it becomes necessary to evaluate the *leadership team* composed of several individuals, or the *leadership environment* shaped by these teams. For example, the CALL system of evaluation asks stakeholders about collective characteristics of the schools — about how often leaders (plural) in a school carry out specific actions, or the school-wide professional development activities, or the feedback practices that teachers typically experience.[33] But existing evaluations are far from being able to measure such forms of distributed or collaborative leadership.

• The analysis of student outcomes can tell only *whether* leaders, or principal preparation programs affect outcomes, not *how* they do so.[34] So outcome evaluations are not useful for improving leaders or leadership preparation programs, but only for threatening principals with dismissal or leadership programs with closure, or for such peripheral activities as providing awards to principals or schools. Of course, there are ineffective leaders who should be fired and weak preparation programs that should be closed, but an emphasis on punishment rather than improvement cannot possibly 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, none of which came from the program's budget.) Very few leadership programs have such resources, and most have to hire statistical expertise from outside. State and federal governments, and foundations, should understand the enormous effort and resources required to find "proof" of effectiveness.

So where does this leave us? Unfortunately, we conclude that at this stage it is impossible to use student outcomes to evaluate principals and leadership preparation programs — unfortunately because improved student outcomes are the goals most worth pursuing. Instead we need to fall back on more conventional forms of evaluation, forms that collect information about practices that presumptively have some influence on outcomes. These include, for example, content evaluations that examine the courses, syllabuses, and practicums that a program provides; satisfaction surveys asking graduates how well they have been prepared for their positions; stakeholder surveys, asking teachers, district officials, other administrators, and perhaps parents and students about the effectiveness of leaders (as the VALED and CALL systems do). Some of these even have distinct advantages over

outcome evaluations: They can more readily identify what dimensions of leadership need to be improved, and whether there are disagreements among stakeholders about the nature and practice of leadership. Their results are more intuitively comprehensible and therefore more transparent than are complex statistical analyses of outcome data. In the absence of much better data and methods, these traditional forms of evaluation will have to do, despite the pressures of this age of accountability.

Appendix Tables

Appendix Table 1: Regression for Base API Score without lagged scores

Source	SS	df	MS		
Model	5955529.03	15	397035.269	Number of obs	654
Residual	5358524.68	638	8398.94151	F(15, 638)	47.27
				Prob > F	0.0000
Total	11314053.7	653	17326.2691	R-squared	0.5264
				Adj R-squared	0.5152
				Root MSE	91.646

BaseAPI	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
PLIMark	-19.66863	12.56081	-1.57	0.118	-44.33415	4.996889
FULL	.7883919	.2384293	3.31	0.001	.3201907	1.256593
EMER	1.164811	.5099281	2.28	0.023	.1634709	2.166152
PCT_AA	-4.090077	.4019497	-10.18	0.000	-4.879381	-3.300773
PCT_AI	-1.344655	2.79361	-0.48	0.630	-6.830436	4.141127
PCT_AS	.4614882	.5651319	0.82	0.414	-.6482552	1.571232
PCT_HI	-1.293612	.543471	-2.38	0.018	-2.36082	-.2264039
PCT_FI	6.673172	3.863507	1.73	0.085	-.9135555	14.2599
PCT_PI	-12.14465	2.766136	-4.39	0.000	-17.57648	-6.712817
P_GATE	2.744971	.3303164	8.31	0.000	2.096332	3.39361
P_EL	-2.405596	.3300119	-7.29	0.000	-3.053637	-1.757555
P_MIGED	-1.503783	1.253426	-1.20	0.231	-3.965122	.9575563
P_RFEP	-3.553812	.4472803	-7.95	0.000	-4.432131	-2.675493
P_DI	-1.418505	.4531305	-3.13	0.002	-2.308313	-.528698
MEALS	.7028087	.2419196	2.91	0.004	.2277539	1.177864
_cons	857.89	39.37993	21.78	0.000	780.56	935.2199

PLIMark = 1 for schools with PLI leaders, 0 otherwise; FULL and EMERG describe the proportion of teachers with full and emergency credentials rather than no credentials; %AA is the percent of students who are African American; %AI American Indian; AS Asian-American; HI Hispanic; FI Filipino; PI Pacific Islander; GATE gifted and talented program; EL English learners; MIGED migrant students; RFEP students re-designated as fluent English proficient; DI disabled; and MEALS is the percent low income as measured imperfectly by eligibility for free and reduced-priced meals.

Appendix Table 2: Regression for Base API Score with lagged scores

Source	SS	df	MS
Model	5466383.34	16	341648.959
Residual	1414662.34	445	3179.0165
Total	6881045.68	461	14926.3464

Number of obs	462
F(16, 445)	107.47
Prob > F	0.0000
R-squared	0.7944
Adj R-squared	0.7870
Root MSE	56.383

BaseAPI	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
PLIMark	-8.143742	8.557445	-0.95	0.342	-24.96177	8.674284
FULL	.2381187	.2111031	1.13	0.260	-.1767641	.6530015
EMER	.9934955	.4551074	2.18	0.030	.0990686	1.887922
PCT_AA	-1.643801	.3499845	-4.70	0.000	-2.331629	-.955973
PCT_AI	2.318424	2.038792	1.14	0.256	-1.688432	6.325281
PCT_AS	.6721534	.4600532	1.46	0.145	-.2319935	1.5763
PCT_HI	.2852688	.4758937	0.60	0.549	-.6500095	1.220547
PCT_FI	7.121747	2.935713	2.43	0.016	1.352162	12.89133
PCT_PI	-4.412109	2.02511	-2.18	0.030	-8.392077	-.4321412
P_GATE	.76622	.2371595	3.23	0.001	.3001283	1.232312
P_EL	-1.436816	.2983016	-4.82	0.000	-2.023071	-.8505612
P_MIGED	-.9271431	.8055539	-1.15	0.250	-2.510306	.6560194
P_RFEP	-2.10445	.3591017	-5.86	0.000	-2.810196	-1.398704
P_DI	-.8594588	.4060793	-2.12	0.035	-1.65753	-.0613875
MEALS	.6162786	.2315845	2.66	0.008	.1611434	1.071414
YearPriorBase	.6819891	.0353214	19.31	0.000	.6125717	.7514065
_cons	293.6723	43.97321	6.68	0.000	207.2514	380.0933

Appendix Table 3: Regression for API Growth Rates

Source	SS	df	MS		
Model	6145639.36	15	409709.29	Prob > F	0.0000
Residual	5116233.32	635	8057.06035	R-squared	0.5457
Total	11261872.7	650	17325.958	Adj R-squared	0.5350
				Root MSE	89.761

Number of obs =	651					
F(15, 635) =	50.85					
Prob > F =	0.0000					
R-squared =	0.5457					
Adj R-squared =	0.5350					
Root MSE =	89.761					

GrowthAPI	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
PLIMark	-13.82785	12.07574	-1.15	0.253	-37.54107	9.885373
FULL	1.215846	.2170806	5.60	0.000	.7895636	1.642129
EMER	1.359373	.4324512	3.14	0.002	.5101659	2.208581
PCT_AA	-3.881908	.3900639	-9.95	0.000	-4.647879	-3.115937
PCT_AI	-2.702245	3.073608	-0.88	0.380	-8.737911	3.333421
PCT_AS	1.291824	.5470822	2.36	0.019	.2175154	2.366133
PCT_FI	6.109265	3.895687	1.57	0.117	-1.540722	13.75925
PCT_HI	-.288913	.5358022	-0.54	0.590	-1.341071	.7632455
PCT_PI	-14.09387	2.643266	-5.33	0.000	-19.28447	-8.903273
MEALS	.4552684	.2303283	1.98	0.049	.0029712	.9075657
P_GATE	1.765096	.2481132	7.11	0.000	1.277875	2.252318
P_MIGED	-1.426329	1.744381	-0.82	0.414	-4.851782	1.999124
P_EL	-3.249828	.3290199	-9.88	0.000	-3.895927	-2.603729
P_RFEP	-3.370741	.4179149	-8.07	0.000	-4.191403	-2.550078
P_DI	-1.030461	.47236	-2.18	0.030	-1.958038	-.1028848
_cons	822.8076	36.18518	22.74	0.000	751.7505	893.8646

Appendix Table 4: Regression for Suspension Rates with Lagged Rates

Source	SS	df	MS
Model	33.7809623	16	2.11131014
Residual	43.9422094	486	.090416069
Total	77.7231717	502	.154827035

Number of obs	= 503
F(16, 486)	= 23.35
Prob > F	= 0.0000
R-squared	= 0.4346
Adj R-squared	= 0.4160
Root MSE	= .30069

SuspensionR~e	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
PLIPresent	.0212274	.0397757	0.53	0.594	-.0569262	.099381
FULL	-.0009727	.0003901	-2.49	0.013	-.0017392	-.0002063
EMER	-.002214	.0039745	-0.56	0.578	-.0100234	.0055954
PCT_AA	.0031452	.001205	2.61	0.009	.0007775	.0055129
PCT_AI	-.0129254	.0178219	-0.73	0.469	-.047943	.0220921
PCT_AS	.001575	.0020101	0.78	0.434	-.0023744	.0055245
PCT_FI	.0020167	.0152507	0.13	0.895	-.0279489	.0319822
PCT_HI	.0016431	.0019453	0.84	0.399	-.0021791	.0054653
PCT_PI	.002028	.0088042	0.23	0.818	-.015271	.0193269
MEALS	-.0017906	.0005527	-3.24	0.001	-.0028767	-.0007046
P_GATE	-.001158	.0009314	-1.24	0.214	-.002988	.000672
P_MIGED	.0019896	.0064899	0.31	0.759	-.0107622	.0147413
P_EL	-.0001529	.00148	-0.10	0.918	-.0030609	.0027552
P_RFEP	.0014953	.0019496	0.77	0.443	-.0023354	.005326
P_DI	-.00279	.0030262	-0.92	0.357	-.0087361	.0031561
YearPriorSusp	.4026146	.024389	16.51	0.000	.3546936	.4505356
_cons	.0236676	.0977814	0.24	0.809	-.168459	.2157941

.

Appendix Table 5: Regression for Truancy Rates with Lagged Rates

Source	SS	df	MS
Model	24.5228816	17	1.44252244
Residual	28.5529498	485	.058872061
Total	53.0758313	502	.105728748

Number of obs	503
F(17, 485)	24.50
Prob > F	0.0000
R-squared	0.4620
Adj R-squared	0.4432
Root MSE	.24264

TruancyRate	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
PLIPresent	-.0102883	.0321225	-0.32	0.749	-.0734048	.0528282
FULL	.0008364	.000323	2.59	0.010	.0002019	.001471
EMER	.0047418	.0032965	1.44	0.151	-.0017354	.011219
PCT_AA	.005783	.0010362	5.58	0.000	.0037469	.007819
PCT_AI	-.0047943	.0144572	-0.33	0.740	-.0332007	.0236121
PCT_AS	.0013666	.001651	0.83	0.408	-.0018774	.0046106
PCT_FI	.0134356	.0122939	1.09	0.275	-.0107203	.0375915
PCT_HI	.0019292	.0015852	1.22	0.224	-.0011855	.0050438
PCT_PI	.0087094	.0073291	1.19	0.235	-.0056913	.0231102
PCT_MR	-.0038422	.0007841	-4.90	0.000	-.0053828	-.0023015
MEALS	-.0033691	.0006696	-5.03	0.000	-.0046847	-.0020534
P_GATE	-.00033	.0007935	-0.42	0.678	-.001889	.0012291
P_MIGED	-.0010775	.0052418	-0.21	0.837	-.011377	.009222
P_EL	.0008151	.0011967	0.68	0.496	-.0015363	.0031665
P_RFEP	.006002	.0015914	3.77	0.000	.0028752	.0091288
P_DI	-.0007738	.0024451	-0.32	0.752	-.005578	.0040304
YearPriorTruancy	.4876175	.0374972	13.00	0.000	.4139406	.5612945
_cons	-.0392224	.0795725	-0.49	0.622	-.1955719	.1171271

Endnotes

[1] The obligatory reference is Kenneth Leithwood, Karen Seashore Louis, Stephen Anderson, and Kyla Wahlstron, *How Leadership Influences Students Learning*. University of Minnesota, Center for Applied Research and Education Improvement; and University of Toronto, Ontario Institute for Studies in Education, 2004. This is the basis for the common statement that one quarter of the outcomes of schooling can be attributed to principals, though this apparently precise number is really a guesstimate based on the qualitative literature.

[2] Indeed, these characteristics may vary among schools, among districts, and among groups of students (like class or racial groups, or students at the elementary, middle school, and high school levels), so there is no reason to think that a consensus can develop. For a recent example reviewing other studies in this area see Jason Grissom, Demetra Kalogrides, and Susanna Loeb, "Using Student Test Scores to Measure Principal Performance", *Educational Evaluation and Policy Analysis* March 2014.

[3] See the National Association of Elementary School Principals and the National Association of Secondary School Principals, *Rethinking Principal Evaluation: A New Paradigm Informed by Research and Practice* (undated). See also Matthew Clifford, Ulcca Hanse, and Sara Wright, *Practical Guide to Designing Comprehensive Principal Evaluation System*, American Institutes for Research, Feb. 2014.

[4] See Joseph Murphy, Ellen Goodring, and Andrew Porter, "Principal Evaluation Takes Center Stage", *Principal*, January/February 2014. The authors are the developers of the VALED system,

[5] There's still no consensus on the usefulness of test scores in evaluating teachers, and virtually every researcher who has examined value-added measures has identified scores of conceptual and statistical problems.

[6] By school-level leaders we mean principals, assistant principals, teacher-leaders, and any others who serve on school leadership teams. In practice, however, most evaluations have focused solely on the principal. For the serious limitations in doing so, see our point about distributed leadership in Section V.

[7] See Tredway, L., Stephens, D., Hedgspeth, C., Jimes, C., & Rubio, R. (2012**),** *A tripartite framework for leadership evaluation.* Berkeley, CA: The Leadership Connection, Graduate School of Education.

[8] One evaluation, for the New York City Aspiring Principals Program (APP), found that that APP principals were more likely than others to transfer schools and leave the district or the profession, thereby exacerbating the instability of principals. Student test scores in APP schools were about the same as in other schools, after controlling for the fact that APP principals were more likely to be in low-performing schools, though the dynamic pattern was complex. However, the study had the usual weak controls for student background, failed to control for the selectivity of APP, and suffered from the flaws we identify in Section V; see Sean Corcoran, Amy Ellen Schwarz, and Meryle Weinstein, "Training Your Own: The Impact of New York City's Aspiring Principals Program on Student Achievement", *Educational Evaluation and Policy Analysis,* April 23, 2012. Another study has examined New Leaders for New Schools, which hired the RAND Corporation to carry out several analyses of the program, at a cost of about $3.5 million; for the final outcome report, see Susan Gates et al., *Preparing Principals to Raise Student Achievement,* RAND, 2014. The study reports that students in schools led by New Leaders have slightly higher test scores, from .7 to 1.3 percentile points in lower grades (but only for students with 2 or 3 years' tenure at a New Leaders school), with an advantage of 3 percentile points at the high school level (but in math only, and only for principals with 3 or more years of experience at the school). These differences are statistically significant but practically negligible; in addition, in several specifications the differences are statistically insignificant — overall there are 4 significant effects and 14 insignificant effects (in Tables 5.1 and 5.2). Furthermore, the effects vary enormously from city to city, with negative results in several districts (Milwaukee, New Orleans, and Prince George's County) for elementary schools, and in Memphis and New York for high school math (in Figures 5.1 and 5.2). Therefore

there is not a New Leaders effect overall, but one that depends critically on how the program is implemented in different districts and different states, with potentially different policies for principals, schools, and test scores. Furthermore, there are several flaws in their methodology, including the use of value-added models with test score data from different states, as well as the ones we outline in Section V. See in addition Edward Fuller, "Review of Preparing Principals to Raise Student Achievement", National Education Policy Center, University of Colorado, Boulder, June 2014. Overall, the confident assertion on the New Leaders website — "students in elementary and middle schools led by New Leader Principals are achieving at significantly higher levels than their peers in other schools" — is simply incorrect.

[9] For more information on the PLI, visit its website at http://gse.berkeley.edu/policy-organization-measurement-evaluation/pli.

[10] See especially the study of the New Leaders for New Schools program noted in footnote 8, which cost roughly $3.5 million — much, much more than leadership programs that are not "donor darlings" can afford.

[11] For such an analysis using the National Longitudinal Survey of the Class of 1988, see W. Norton Grubb, *The Money Myth: School Resources, Outcome, and Equity* (Russell Sage Foundation, 2009). This study used as outcomes 4 test score measures, 3 measures of student aspirations, and 5 measures of academic progress through high school.

[12] See the RAND analysis of the New Leaders for New Schools program, described in footnote 8 above, with the finding of enormous variation among districts.

[13] In addition, we included four instances of a PLI graduate who first served as a vice or assistant principal at a school before progressing to become principal at the same school.

[14] For each year, the total percentages add up to over 100%. The reason for this slight error is that the state reports these data in whole numbers with no decimal places.

[15] Part of the reason was our desire to compare similar schools, or schools with similar SIM rank numbers as assigned by the state. Unfortunately SIM rank data is available only for base API scores and not for API growth scores.

[16] For more information on the construction of the API see the California Ed-Data system at https://www.ed-data.k12.ca.us/pages/understandingtheapi.aspx.

[17] Again, rich data are usually available only with the longitudinal surveys conducted by the U.S. Department of Education. See again see W. Norton Grubb, *The Money Myth: School Resources, Outcome, and Equity* (Russell Sage Foundation, 2009), with 14 measures of family background, 35 measures of school characteristics including the instructional practices of teachers, and 19 characteristics of students.

[18] The reason, from Stat 101, is this: A regression with a value-added measure as the dependent variable is $TS_t - TS_{t-1} = bX + e$, where X is a vector of independent variables including descriptors of principals, *b* is a vector of coefficients, and e is an error term. But this regression is equivalent to $TS_t = TS_{t-1} + \boldsymbol{b}X + e$, a regression where the coefficient on TS_{t-1} is constrained to be one — that is, one point higher on a previous year's test score leads to precisely one point increase on the current year's test score. But there's no reason for this coefficient to be one; indeed, in equations of this kind (which are difference equations), the coefficient on the lagged dependent variable is usually between zero and one. The correct specification, as reported in Appendix Table 2, is $TS_t = aTS_{t-1} + bX + e$, where the coefficient *a* on the lagged dependent variable TS_{t-1} is estimated rather than constrained to be one.

[19] We have also run a regression for API growth scores, adding to the specification in Appendix Table 3 a lagged growth score, to parallel the specification in Appendix Table 2. However, the results are almost precisely the same as in Table 3, and this specification makes no sense given the way API growth scores are constructed.

[20] The calculation of the SIM ranks can be found at https://www.ed-data.k12.ca.us/pages/understandingtheapi.aspx.

[21] For more information on the calculation of truancy, see www.cde.ca.gov/ls/ai/tr/. The state's measure counts those with unexcused absence three times or more as truant, so it seems a very stringent measure of truancy.
[22] The equations with lagged dependent variables take the form of difference equations. The coefficient on the lagged dependent variable implies an equilibrium level of the dependent variable, which is due more to the other independent variables when coefficient is low, and due more to past values of the dependent variable when the coefficient is high.
[23] See the report, mentioned in the introduction, by Norton Grubb and Rebecca Cheung "Collective and Team Leadership: Preparation for Urban Schools" (2014), which interviewed graduates working in schools with two or more PLI alumni. The commitment of the PLI to equity and social justice has been documented in Tina Trujillo and Robert Cooper, "Framing Social Justice Leadership in a University-Based Preparation Program: The University of California's Principal Leadership Institute", *Journal of Research in Educational Leadership* 2014.
[24] The first three figures are calculated by the averaged freshman graduation rate, and the last three figures use the adjusted cohort graduation rate. See Marie Stetser and Robert Stillwell, "Public Four Year On-time Graduation Rates and Even Dropout Rates: School Years 2010-11 and 2011-12", National Center for Education Statistics (NCES 2014-391), U.S. Department of Education, April 2014.
[25] See again the study of the New York Aspiring Principals Program and the RAND evaluation of the New Leaders for New Schools program, Described in footnote 8.
[26] Again, see W. Norton Grubb, *The Money Myth: School Resources, Outcome, and Equity* (Russell Sage Foundation, 2009), especially Chapters 2, 4, and 5 with the accompanying tables of regression coefficients.
[27] In addition to the difficulties we examine in this section, see also Edward Fuller, Liz Hollingsworth, and Michelle Young, "Problems and Prospects of Using Outcome Measures to Evaluate Principal Preparation Programs", paper prepared for the annual meeting of the University Council of Educational Administration, Nov, 2013; Susanna Loeb and Jason Grissom, "What Do We Know about Value-added Measures for Principal Evaluation?", Carnegie Foundation for the Advancement of Teaching; and Hanley Chiang, Stephen Lipscomb, and Brian Gill, "Is School Value-added Indicative of Principal Quality?", Working Paper, Mathematica Policy Research, November 2012. All these papers conclude as we do that existing methods and data, especially those using value-added measures, are inadequate and often misleading.
[28] In the RAND analysis of New Leaders cited in footnote 8, students with one year in a New Leaders school experienced no effect for either math or reading, as we might expect; the effects on math only were small and positive for students in their second year; and effects on both math and reading were larger their third year. The effects of principals with 1, 2, and 3 years of experience were all insignificant in elementary schools, and significant only for reading scores for principals with three or more years of experience (Tables 5.1 and 5.2). One implication of these results is that the experience of principals in a school doesn't matter as much as student tenure, though the enormous variation of outcomes among districts makes this conclusion highly uncertain.
[29] Grubb, *The Money Myth*, presents evidence that increasing test scores requires improving the quality of teachers and instruction, while progress through high school is affected more by factors that enhance student support.
[30] There's a long literature recognizing that the influences of principals on students is indirect; see for example, Phillip Hallinger and Ronald Heck, "Exploring the Principal's Contribution", *School Effectiveness and School Improvement* 2013, Vol. 9(2): 157 – 191. However, this insight has not carried over to evaluation methods.
[31] Again from Stat 101, the first equation describing effects on school resources broadly defined SR is

$$SR = a + bX + \epsilon$$

where X is a vector of variables affecting school resources including the characteristics of leaders, b is a vector of coefficients describing their effects of on school resources, and ϵ is an error term. Then the second equation for student outcomes as a function of school resources plus other variables influencing outcomes (including student characteristics) is

$$OUT = \alpha + \beta \mathrm{Y} + \gamma \mathrm{SR} + \mu$$

where OUT represents many outcomes (including but not limited to test scores), Y includes the other variables influencing outcomes including student characteristics; school resources SR affect outcomes; and μ is another error term. (The second of these equations is an educational production function.) But these two equations imply, substituting the first into the second, that

$$OUT = \alpha + \beta \mathrm{Y} + \gamma(a + bX + \epsilon) + \mu$$

or

$$OUT = \alpha + \beta Y + \gamma a + \gamma bX + \gamma \epsilon + \mu$$

This is the reduced-form equation that everyone (including us) has estimated, with student outcomes regressed on Y and X, and a complex error term $\gamma\epsilon + \mu$. What we want to know are the vectors of coefficients b, β, and γ. However, if X and μ are correlated (schools with exceptional principals and other influences on school resources have higher than expected outcomes), or Y and ϵ are correlated (student characteristics and other influences are related to unexpectedly high or low levels of resources, as happens when low-income students are placed in lower tracks) then the coefficients on Y and X will be biased. In addition, the reduced form equation cannot separately identify γ and b, even through their meanings are quite different. If these coefficients are biased and/or under-identified, then a single reduced-form equation cannot accurately estimate the effects of principals reflected by parameters in b.

[32] On distributed leadership see especially James Spillane, *Distributed Leadership* (San Francisco: Jossey-Bass, 2006). On the concept of collective leadership and the way it emerges in the PLI, see W. Norton Grubb and Rebecca Cheung, "Collective and Team Leadership: Preparation for Urban Schools" (2014), based on interviews with PLI graduates working in schools with two or more PLI alumni.

[33] For a brief introduction to the CALL system, see Richard Halvorsen, Carolyn Kelly, and James Shaw, "A CALL for Improved School Leadership", *Phi Delta Kappan*, March 2014.

[34] A partial exception is the empirical literature on the effects of principals cited in footnote 2 above, where data can sometimes be collected on the emphasis or the amounts of time principals spend on various activities. Such data are available only under special conditions.