Research Report

GERALD P. HUNTER, STEPHANIE WILLIAMSON, ASA WILKS, JANET M. HANLEY, BRIAN M. STECHER

Using Data to Support the Intensive Partnerships for Effective Teaching Initiative

Data Collection, Metric and Dashboard Creation, and Lessons Learned

The Intensive Partnerships for Effective Teaching (IP) initiative, which was funded by the Bill & Melinda Gates Foundation, was a multiyear effort launched in 2009 to improve student outcomes dramatically—particularly high school (HS) graduation and college attendance among low-income minority (LIM) students—by increasing student access to effective teaching. Specifically, the IP initiative focused on reforms aimed at (1) improving teacher evaluation (through a measure of effective teaching); (2) enhancing staffing policies (including recruitment and hiring, placement and transfer, and tenure and dismissal); (3) providing customized professional development (PD); and (4) reforming compensation and career ladders. The intermediate goals of these policy changes were to improve the level and distribution of effective teaching and boost the access of LIM students to higher-quality teachers; the long-term goals were to increase college readiness and attendance, particularly among underrepresented groups.

KEY FINDINGS

- Long-term projects need the flexibility to accommodate change. During this project, there were changes in sites' data collection procedures, data management systems, and research emphasis.
- Concerns about accurate data handling are essential to a successful data-based analytic project, but it might be equally important to invest in building positive working relationships with client data administrators and information technology personnel.
- Different kinds of solutions worked in different contexts, and a successful data-based analytic project will benefit from creative thinking about ways to rescale or reclassify based on existing measures.
- Collaboration with district staff was critical in the deidentification of student data and the substitution of new unique identifiers. Confidentiality requirements necessitate data use agreements and data safeguarding procedures that will add time and complexity to the project.

Seven IP sites—three school districts and four charter management organizations (CMOs)—were selected to implement the reforms over a six-year period. The three school districts were Hillsborough County Public Schools (HCPS) in Florida, Memphis City Schools in Tennessee, and Pittsburgh Public Schools (PPS) in Pennsylvania.[A] The four CMOs were Alliance College-Ready Public Schools, Aspire Public Schools, Green Dot Public Schools, and Partnerships to Uplift Communities (PUC) Schools in California. The seven sites were selected to participate because of the high degree of stakeholder support the administration, teachers, unions, and the community provided. The seven IP sites were encouraged to implement the initiative—including developing measures of teaching effectiveness and modifying personnel policies—in ways that best suited their specific conditions. An eighth site, Denver Public Schools (DPS) in Colorado, joined the project in 2013.

Abbreviations

ACT	American College Test
AIR	American Institutes for Research
CCSS	Common Core State Standards
CMO	charter management organization
CST	California Standards Test
DPS	Denver Public Schools
DUA	data use agreement
EAP	Early Assessment Program
EOC	end-of-course test
ELA	English language arts (reading is sometimes used to refer to English language arts)
FERPA	Family Educational Rights and Privacy Act
FRL	free or reduced-price lunch
HCPS	Hillsborough County Public Schools
HR	human resources
HS	high school
ICPSR	Inter-University Consortium for Political and Social Research
ID	identifier
IP	Intensive Partnerships for Effective Teaching
IT	information technology
LIM	low-income minority
NSC	National Student Clearinghouse
PD	professional development
PII	personally identifiable information
PPS	Pittsburgh Public Schools
PUC	Partnerships to Uplift Communities
SAT	Scholastic Aptitude Test
SCS	Shelby County Schools
SIS	student information system
SME	subject-matter expert
SY	school year
TE	teacher effectiveness
TVA	teacher value added

As part of the initiative, the foundation wanted to help the sites monitor their progress and reflect on their improvement efforts. Although the project data team was not involved in decisions about creating and using the dashboards, such tools are widely used in education, and research suggests that they can be useful in helping districts and schools monitor and improve outcomes for students (Phillips, Reber, and Rothstein, 2018; West, 2012; Marsh, Pane, and Hamilton, 2006). Toward that end, the foundation supported the construction of an annual data dashboard for each site that presented indicators of student and teacher performance along several dimensions. The dashboards were used to inform discussions at the annual meetings the foundation held with each site. The foundation contracted with the RAND Corporation to collect and warehouse data from participating sites and to produce the annual dashboards. RAND and the American Institutes for Research (AIR) also have led efforts to evaluate the implementation, outcomes and impact, and replication and scaling of the broader IP initiative.[B] Some of the data described in this report were key inputs for the RAND/AIR IP site evaluation teams. The evaluation results are presented in

[A] In 2013, Memphis City Schools and the surrounding county merged and became Shelby County Schools (SCS).

[B] The RAND Corporation and AIR conducted an independent evaluation of the IP initiative. To learn more about the initiative and the results of the evaluation, see Stecher et al., 2018; and Stecher et al., 2019.

separate publications, which can be viewed on the RAND website (RAND Corporation, undated).

In this report, we describe the data warehouse and dashboard, including the steps used to develop and implement data collection and dissemination protocols to support the sites' participation in the IP initiative. In addition, we highlight some of the key data-related challenges we had to overcome and draw lessons related to the systematic use of education data for periodic program monitoring. This report provides useful illustrations about the complexities the RAND data team encountered and describes solutions or decisions researchers made. We believe that the lessons learned from this effort will be useful to research teams, data scientists, administrators, software developers, and other analysts who want to use district-level and state-level administrative data to monitor changes in administrative procedures, teacher assignments, and student achievement related to large-scale education interventions.

Role of the RAND Data Team

A RAND data team was formed to collect and warehouse data from participating sites with the goals of facilitating multisite monitoring through the creation of annual dashboards and facilitating a research effort using harmonized data from all sites. That effort involved four main activities: (1) defining the metrics that would be used to monitor and assess annual progress and that would appear in the dashboard, (2) collecting the data from the sites used to compute the metrics and complete the research, (3) managing and standardizing the data, and (4) creating the dashboard and reporting the metrics to the sites and the foundation.

The RAND data team worked with the foundation to define the metrics that could be tracked over time and that would provide accurate, clear, and timely information about the sites' progress toward initiative objectives. These metrics fell into five categories: effectiveness of teaching, equity of access to effective teaching, excellence in student achievement and engagement, endurance of reforms, and execution of key activities. The *effectiveness* metrics include several measures of teacher effectiveness and measures that address managing the teaching workforce, such as tenure decisions and selective retention of newly hired teachers. The *equity* metrics measure equitable access to effective teachers for all students, strategic staffing of math teaching positions, and discipline rates. The *excellence* metrics focus on student achievement and engagement. The *endurance* of the reforms and the *execution* of key activities were measured by responses to teacher and school leader surveys conducted by AIR. The RAND team, the sites, and the foundation held a series of discussions to define the specific metrics. In many cases, the initial definitions were modified to reflect data availability, make the interpretation clearer, or correspond to other published statistics.

Once the metric definitions were established, the RAND data team designed the procedures for obtaining required data from the sites and methods for using the data to compute the metrics. The specifics of the data and the computations varied across sites because of differences in the underlying data the sites' information technology (IT) systems produced. The RAND data team worked directly and iteratively with the sites, sharing preliminary indicators to give them a chance to address data issues. This corrective phase was important to the sites because the dashboard contents were discussed at the annual "stock-taking" meetings held between the sites and the foundation to monitor progress toward achieving the initiative's goals. After the data from the sites were read, validated, cleaned, and standardized (a process that often included requests by the RAND data team to the sites to clarify or modify the files provided), the metrics were calculated and reviewed and the dashboards were generated. The cycle of requesting data, validating data, and computing metrics was done annually over the project.

This report is intended to discuss the challenges of conducting the four activities described earlier and how the RAND data team addressed those challenges. After briefly describing the data warehouse and the dashboard created for context, including the steps used to develop and implement data collection and dissemination protocols to support the sites' participation in the IP initiative, we examine challenges and ways to address those challenges in four areas: (1) issues related to defining metrics used to track system performance; (2) issues related to data collection; (3) issues related to managing and standardizing data across sites; and (4) issues related to data confidentiality, data sensitivity, and partnerships. We also draw overarching lessons related to the systematic use of education data for periodic program monitoring.

Challenges and Recommended Ways to Address Them

In this report, we discuss some of the key data-related challenges and ways to address them for each of the four areas. Some of the recommendations for addressing challenges were implemented in our project, while other recommendations were not used by us but could be used by researchers, data scientists, and practitioners. The former are shaded in Table 1, which presents an example challenge and the way to address it for each of the four areas.

Overarching Lessons Learned

Some general lessons learned to share with other researchers who work with school district data systems to evaluate improvement efforts include the following: long-term projects need the flexibility to accommodate change; success is more likely if you establish good relationships with your counterparts in district offices; to create comparable metrics across sites, think creatively about redefining and rescaling existing measures; and be prepared to collaborate with district staff in the deidentification and substitution of new unique student and teacher identifiers.

Although we encountered several challenges, we were able to produce dashboards for the sites and the foundation in a timely fashion and provide reliable data to the research teams. The flexibility, thoughtfulness, patience, and creativity on the part of the sites, the foundation, and the research teams in handling these issues ultimately led to a sustained, successful effort. We hope that the issues, suggestions, and solutions presented in this report can inform future efforts related to collecting, processing, and analyzing these types of data.

Literature Review

This project supported research and monitoring using site-provided data and followed a data-driven decisionmaking framework. The data were organized and synthesized to give policymakers information and actionable knowledge to use in decisionmaking (Marsh, Pane, and Hamilton, 2006). We used both data repositories and dashboards to facilitate this effort. School districts, CMOs, and states increasingly maintain systems that include data on attendance,

TABLE 1
Example Challenge and Way to Address It in Each Area

Area	Challenge	Recommended Way to Address It
Defining metrics used to track system performance	There are differences in subjective views about how to operationalize a general construct, such as effectiveness	Conduct conversations with stakeholders to come to a shared understanding about key constructs
Data collection	State data release schedules vary, affecting the timeliness of data processing and reporting	Make proposed reporting schedules contingent on the receipt of external data and plan for multiple reporting cycles with increasing levels of completeness so that stakeholders have results as soon as possible for decisionmaking
Managing and standardizing data across sites	Course names and subject-matter categories are not standardized across sites	Work with sites to develop crosswalks between course titles, obtain detailed course descriptions and use them to classify courses into subjects, and work with SMEs to understand course content and make proper classifications
Data confidentiality, data sensitivity, and partnerships	Confidentiality requirements (such as FERPA) necessitate DUAs and data safeguarding procedures that will add time and complexity to the project	Make personal visits to each site to establish working relationships with data providers, develop mutually agreeable procedures, and build sites' confidence in researchers' ability to handle sensitive data

NOTES: DUA = data use agreement. FERPA = Family Educational Rights and Privacy Act. SME = subject-matter expert.

special program participation, course enrollment, and state tests scores and that follow students over time (Phillips, Reber, and Rothstein, 2018). These data repositories can even connect academic data to human services data to help states understand students' mental health and child welfare, as is done in Allegheny County, Pennsylvania (Fraser, 2015). To be most useful, the different data streams often must be combined and synthesized. Data dashboards with key metrics are one method for school officials to visualize how their organizations are performing (West, 2012).

Data dashboards are widely used in education to allow policymakers, administrators, teachers, students, and/or parents to monitor the status of national or state systems, school districts, specific schools, classrooms, or students on a variety of quantifiable measures (West, 2012). For example, the United States Education Dashboard shows annual state progress toward national student outcome goals for 2020 (U.S. Department of Education, undated). Most states have created their own dashboards that track progress on measures adopted at the state level. Educational dashboards exist at many levels, from local—individual school districts have dashboards that display data at the school and even classroom levels—to international—The World Bank is designing a Global Education Policy Dashboard that will monitor educational service delivery, policies, and politics around the world (The World Bank, 2019). Data dashboards offer a way for educators to distill data about their students and schools into "manageable chunks" (Donhost and Anfara, 2010, p. 60). Furthermore, some school districts are currently connecting their data systems to encourage the use of data by educators and administrators by comparing measures across classrooms and schools and reviewing trends over time (Means, Padilla, and Gallagher, 2010).

From the proliferation of dashboards mentioned earlier, one might infer that it is easy to develop a useful educational dashboard, but that is not the case. Many educational institutions have the ability to collect, analyze, and summarize data and display them on a chart, but such displays are not likely to be useful unless sufficient effort has gone into the planning, conceptualization, design, delivery, and interpretation of the information. Measure definition often is the most challenging part of dashboarding because it requires developing consensus across stakeholders on exactly how the measures are to be monitored (Smith, 2013). Educators and data scientists have explored the potential benefits and challenges of dashboards in education and in other policy areas. For example, Matheus, Janssen, and Maheshwari (forthcoming) delineate 13 potential strategic, political, and operational benefits of dashboards, but they also identify 15 risks that threaten the attainment of these potential benefits. Similarly, Raftree (2015) talked with data leaders at private organizations and found that they were "struggling" with designing and operating dashboards that would lead to better decisions. Based on these conversations, Raftree offers suggestions for better dashboard design.

[D]isplays are not likely to be useful unless sufficient effort has gone into the planning, conceptualization, design, delivery, and interpretation of the information.

We did not find any research that empirically tested specific practices for designing or implementing educational dashboards. Most of the publications we identified offered suggestions for dashboard development that grew out of problems the authors had seen in one context or another. These suggestions related to five broad themes: (1) clarifying the purposes of the dashboard; (2) engaging stakeholders; (3) obtaining high-quality, meaningful data; (4) generating effective summary statistics and data displays; and (5) helping people interpret and use the dashboard for decisionmaking. The second, third, and fourth themes are directly relevant to the work of the RAND data team in this project; the foundation established the purposes of the dashboard (first theme) and worked with the districts to make use of the data (fifth theme).

Rothman (2015) found that stakeholder engagement involves setting appropriate targets or benchmarks, choosing the right indicators, and identifying the most-critical indicators for special scrutiny. Also, because dashboards must be responsive to ongoing and changing monitoring needs, occasional reassessment of the usefulness of individual measures is recommended (Smith, 2013). High-quality and meaningful data might depend on a balance between availability and recency. Some users (or metrics) can use data that are several months old, while others require almost real-time access to information (Philips, Reber, and Rothstein, 2018). Fragmented responsibility for data can decrease the quality and timeliness of the data, and failure to maintain or update data can result in a loss of stakeholder trust (Matheus, Janssen, and Maheshwari, forthcoming). Lastly, the use of similar data formats and common metrics across institutions enhances data comparability and adds to the effectiveness of the data displays (West, 2012).

Several issues were identified in the literature that were relevant to the tasks undertaken by the RAND data team, although none of the publications found were comprehensive with respect to dashboard design and presentation. The main suggestions are to (in each case, we are quoting from or paraphrasing the cited work):

- work with users to develop the dashboard; this will create ownership on the part of stakeholders—including staff who supply the data—and that buy-in will improve the quality of the data (Raftree, 2015)
- be clear about your data categories and indicators, particularly if different sites are using different data models (Raftree, 2015)
- constantly change and iterate on the dashboards; they will need continuous upkeep (Raftree, 2015)
- attend to data quality, making sure data are accurate and precise; show error bars so that people make realistic interpretations; and provide appropriate caveats (Raftree, 2015; Matheus, Janssen, and Maheshwari, forthcoming)
- figure out how to connect internal information systems that do not currently interact (West, 2012)
- respect data privacy and properly anonymize results (Matheus, Janssen, and Maheshwari, forthcoming)
- present data in easy-to-understand displays (Matheus, Janssen, and Maheshwari, forthcoming)
- be cautious when there is not a clear consensus on how to measure something; consider using alternative outcomes instead (U.S. Department of Education, undated)
- be aware that dashboard results might not match individual site-produced data summaries; understand the differences and be able to explain them for users (Matheus, Janssen, and Maheshwari, forthcoming)
- organize the data into logical groups; cluster together various measures related to instruction, outcomes, operations, and procedures (Herman, 2016).

In addition to these general recommendations, several obstacles associated with dashboard creation are common to longitudinal data analysis across multiple jurisdictions. Inconsistent terminology (e.g., coding drop-outs), balancing student privacy with data access, and building networks of data sharing (West, 2012) are challenges of creating usable data sets for dashboards or analysis. One of the specific challenges facing the IP initiative was the need to measure teachers' value added and create an indicator of teaching effectiveness. There has been extensive research on ways to measure the effect of teachers on student achievement (Mihaly et al., 2013; Kane and Staiger, 2008) and on the impacts of using teacher value-added estimates for teacher performance (Chetty, Friedman, and Rockoff, 2014; this information also is included in unpublished 2017 research by Michael Dinerstein and Isaac M. Opper).

These studies detail how the estimates are modeled and mention decisions made in data processing only at a high level. Other work has compared value added with other measures of instructional quality, giving more-explicit details on data handling in complex cases (McEachin et al., 2018). As researchers develop these models, school districts have been trying to increase the capability of educators to get timely information about how their districts, schools, and classrooms are performing on key metrics (Means, Padilla, and Gallagher, 2010). The RAND data team did not have to design the value-added measure, but we did have to develop indicators based on the effectiveness measures adopted by the IP sites.

The work of the RAND data team was unique in that it served both the analysis of teacher value-added research and the timelier monitoring needs of school administrators and the foundation. This effort also involved developing metrics across sites rather than at a single site. As a result, the insights we developed about dashboard production and metric development expand the practical recommendations currently found in the literature.

Organization of This Report

In the next section, we describe the final form of the data warehouse and the dashboard created. We begin with the warehouse and dashboard to provide a sense of what was built using the sites' data. Next, we discuss some of the challenges we encountered, roughly in the order in which we faced them, that are the most relevant to other data teams and explain our approaches to solving them. In particular, we focus on defining the metrics used to track system performance; we examine issues related to data collection; we illustrate some of the challenges we encountered in managing and standardizing data across sites; and we focus on the importance of data confidentiality, data sensitivity, and partnerships. In the final section, we summarize our experiences and highlight the most-important lessons learned.

Metric Definitions, Data Warehouse, and Dashboard

Throughout this project, we created a data system that allowed the sites and the foundation to monitor the progress of the initiative over time and across sites. The foundation identified five broad performance goals of the initiative to track—effectiveness, equity, excellence, endurance, and execution—and chose specific elements related to each goal to measure. The RAND team worked with the foundation to define metrics that could be calculated using data provided from the sites. The three main recurring steps in this process were defining the specific metrics that would be calculated, gathering information from the sites and assembling it into a data warehouse, and creating an annual dashboard display that summarized the metrics and tracked them over time.

Understanding the discussions in this section requires knowing some details about the IP initiative. As we noted in the previous section, the IP initiative was designed to improve outcomes for LIM students by increasing their access to effective teaching. Central to the initiative was the development of a measure of teaching effectiveness. Each of the eight sites adopted its own quantitative indicator of effectiveness that included information from structured classroom observations and a value-added measure based on student test performance; some sites also included information from student and

[T]he insights we developed about dashboard production and metric development expand the practical recommendations currently found in the literature.

parent surveys. Once the effectiveness measures were calculated, each site set cut points to classify teachers into four or five levels of effectiveness. These levels of effectiveness were important to multiple IP policies; as a result, they were used in several of the metrics and appeared frequently in the dashboard displays.

In this section, we discuss the final form of the data warehouse and the dashboards for the sites, starting with the process of defining metrics. In the next three sections, we discuss the challenges we encountered in the process and the solutions we used to address those challenges, in addition to some challenges that are part of the process and potential solutions to them.

Metric Definitions

The specific metrics developed to monitor each of the five performance areas discussed earlier are shown in Table 2. The table lists the main performance areas, the metrics in each category, and a brief definition; full definitions are provided in the appendix. The *effectiveness* metrics included several measures of teacher effectiveness (TE) and measures that address managing the teaching workforce, such as tenure decisions and selective retention of newly hired teachers. The *equity* metrics measure equitable access to effective teachers for all students, strategic staffing of math teaching positions, and discipline rates. The *excellence* metrics focus on student achievement and engagement. The *endurance* of the reforms and the *execution* of key activities were measured by responses to teacher and school leader surveys conducted by AIR.

The metric development process involved lengthy discussions among the foundation, the sites, the evaluation research teams, and the RAND data team. It also involved exploring each site's data system because final definitions had to account for different data standards, test regimes, and personnel files. It was important to adopt definitions that permitted the RAND data team to calculate a single metric across all sites. We highlight some of the challenges in defining metrics later in this report.

Data Warehouse

The RAND data team made formal requests to the sites to provide data needed both by the research teams and to compute the metrics described earlier to support the IP initiative. These data were assembled into a data warehouse. In the first year of data collection, sites provided available historical data from school year (SY) 2006–2007 to SY 2011–2012; in each subsequent year, the sites provided annual data through SY 2017–2018. Data files and data elements were standardized across sites and across time. In Table 3, we list the data files that constituted the data warehouse. In addition to the data required for the metrics, the data warehouse included data required to support the research teams. The survey data collected by AIR were not included in the warehouse, although AIR used those data to compute relevant metrics that were included on the dashboards.

Dashboard

After the metrics were defined and the data warehouse was populated with the necessary data from each site, foundation leaders provided broad guidelines about the type of static visual display they preferred for the dashboard. Given this input, the RAND data team developed the dashboard as an Excel workbook with several worksheets. The first worksheet tab showed the number of schools, teachers, and students by school year. The second tab contained a matrix of the metrics by school year, along with notes explaining any missing information, differences in computation from prior years, or general cautions because of a change in testing or reporting (see Figure 1). The third tab displayed a chart or graph of each metric (see Figure 2). The fourth tab listed all computed metrics and the number of students and teachers used for each computation. The dashboards were produced annually in the fall for each site.

The development of the metrics, data warehouse, and dashboard was a complicated task, one that involved the foundation, the sites, the evaluation teams, and the RAND data team. We spent about 18 months discussing, testing, and reviewing prototypes before the final form of the dashboard was set. It was first produced in fall 2012 and incorporated

TABLE 2

Metrics for Effectiveness, Equity, Excellence, Endurance, and Execution

Category	Metric	General Definition
Effectiveness	Differentiation of Performance	Percentage of teachers in each effectiveness category
	Strength of Tenure Decisions	Percentage of newly tenured teachers rated effective or higher
	Performance of New Teachers	Average effectiveness percentile of teachers with three or fewer years in district
	Selective Retention	Percentage of teachers in the highest- and lowest-rated categories (and for top and bottom decile)
	Overall Teacher Improvement	Percentage of teachers who improved minus percentage of teachers who declined, continuous score
	Addressing Ineffective Teaching	Percentage of ineffective teachers who exited or improved
Equity	Access to Effective Teachers	Probability of having a top-decile reading teacher, by student LIM status Probability of having a bottom-decile reading teacher, by student LIM status Probability of having a top-decile math teacher, by student LIM status Probability of having a bottom-decile math teacher, by student LIM status Probability of having a highly effective reading teacher, by student LIM status Probability of having an ineffective reading teacher, by student LIM status Probability of having a highly effective math teacher, by student LIM status Probability of having an ineffective math teacher, by student LIM status
	Strategic Staffing of Math Teachers	Difference of percentage proficient in prior year math achievement scores by students taught by novice (less than or equal to one year) versus experienced (three or more years) teachers
	Discipline Data	Percentage of LIM and non-LIM students expelled or suspended
Excellence	Student Achievement	Percentage proficient or above on state standardized exam in reading and mathematics for all students and by LIM status
	Growth Relative to State	Percentage of students whose scores on state achievement tests increase from grade to grade more than the overall increase in state scores for white students in the same grades
	Closure of Achievement Gap	Closure of the achievement gap for LIM students for reading and mathematics
	Four-Year Graduation Rate	Percentage of students who graduate within four years of entering grade 9 for all students and by LIM status
	College Readiness	Percentage of students at college level on standardized tests for all students and by LIM status
	Participation in College Readiness Assessments	Percentage of each cohort participating in standardized test to measure college readiness
	On-Time College Enrollment	Percentage enrolling in college within five years of entering grade 9 by LIM status and for all students

Table 2—Continued

Category	Metric	General Definition
Endurance	Evaluation System Accuracy (Teachers)	Percentage of teacher survey respondents agreeing that the teacher evaluation system does a good job distinguishing effective from ineffective teachers
	Evaluation System Fairness	Percentage of teacher survey respondents agreeing that the consequences tied to teacher evaluation results are reasonable, fair, and appropriate
	Evaluation System Accuracy (Leaders)	Percentage of school leader survey respondents agreeing that the teacher evaluation system does a good job distinguishing effective from ineffective teachers
	Rigorous Tenure Decisions	Percentage of school leader survey respondents agreeing that, over the past few years, it has become more difficult for teachers to earn tenure in their district
	Value of Supports Aligned to Evaluation	Percentage of teacher survey respondents agreeing that supports aligned to evaluation results are appropriate and helpful
Execution	Prevalence of New Measure Usage	Percentage of teachers receiving a performance rating based on the new measures
	Impact of PD on Learning	Percentage of teacher survey respondents agreeing with the statement, "My PD experiences this year have enhanced my ability to improve student learning"
	Culture of Collaboration	Percentage of teacher survey respondents agreeing with the statement, "Teachers at my school support each other in their efforts to improve teaching"
	Quality of Administrator Feedback	Percentage of teacher survey respondents agreeing with the statement, "After my teaching is observed, I receive useful and actionable feedback"
	Intentional PD Alignment	Percentage of school leader respondents agreeing with the statement, "In my school, we use teacher evaluation results to align PD to each teacher's strengths and weaknesses"
	Strategic Staffing Decisions	Percentage of school leaders reporting that to a moderate or large extent, teachers' evaluation results are used to assign teachers to classes or students within the school
	Job-Embedded Supports	Percentage of school leaders reporting that more than 50 percent of the teachers at their school received some type of formal, individualized coaching and/or mentoring this year
	Selective Hiring for Teacher Leader Roles	Percentage of school leader respondents agreeing with the statement, "The teachers who hold higher-level career ladder or specialized positions at my school are effective educators"

NOTES: Further details about the calculation of each metric are provided in the appendix to this report. Endurance and execution metrics are based on the annual teacher survey and school leader survey conducted by AIR.

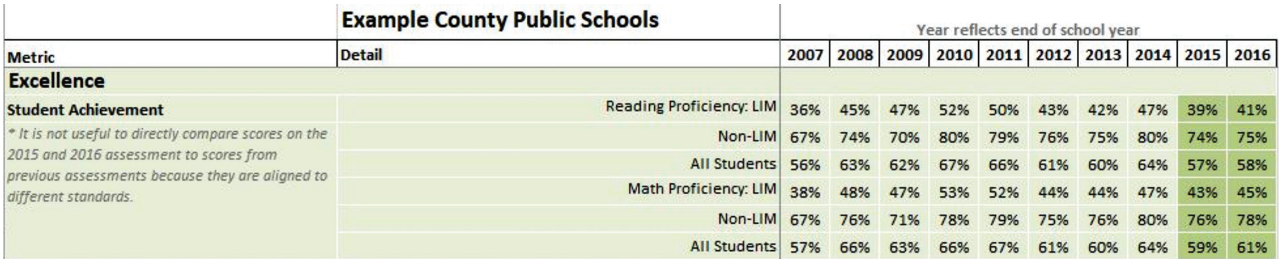
TABLE 3

Content of the Data Warehouse

Data File	Description
Assessment	Student state assessment scores and accommodations
Attendance	Student total days enrolled, present, and absent
College Assessment	Student college assessment scores and accommodations
College Enrollment	College enrollment history
Compensation	Teacher compensation
Course	All courses offered by school, course name, and section, with categorization by type (i.e., mathematics, science, English language arts [ELA])
Course Student	Students' links to course sections
Course Teacher	Teachers' links to course sections
Disability	Student disability
Discipline	Every disciplinary event, with dates, description, and resolution
Enrollment	Student enrollment dates and demographic data for each school attended
Job	Staff job history
School	School location, first and last day of school, grades taught, Title 1 status, and type (i.e., magnet, charter, alternative)
Staff	Staff and teacher demographics and credentials
Teacher Effectiveness	Teacher effectiveness scores and components

FIGURE 1

Snapshot of Dashboard Table



	Example County Public Schools	Year reflects end of school year									
Metric	Detail	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Excellence											
Student Achievement	Reading Proficiency: LIM	36%	45%	47%	52%	50%	43%	42%	47%	39%	41%
** It is not useful to directly compare scores on the 2015 and 2016 assessment to scores from previous assessments because they are aligned to different standards.*	Non-LIM	67%	74%	70%	80%	79%	76%	75%	80%	74%	75%
	All Students	56%	63%	62%	67%	66%	61%	60%	64%	57%	58%
	Math Proficiency: LIM	38%	48%	47%	53%	52%	44%	44%	47%	43%	45%
	Non-LIM	67%	76%	71%	78%	79%	75%	76%	80%	76%	78%
	All Students	57%	66%	63%	66%	67%	61%	60%	64%	59%	61%

NOTES: The darker color in 2015 and 2016 alerts the reader to the change in the state assessments. Scores in 2015 and after are not directly comparable to scores in 2014 and earlier because the assessments align to different state standards.

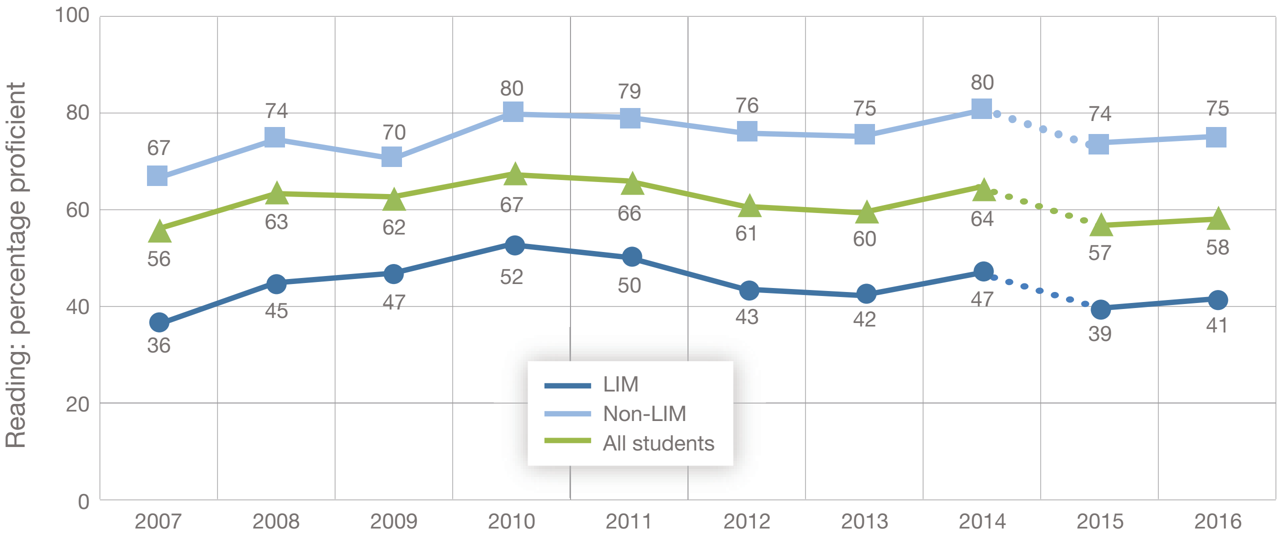
data from SY 2006–2007 through SY 2011–2012; we then produced an annual dashboard for each site for the next six years. For a variety of reasons (and consistent with observations from Raftree, 2015), small changes were required almost every year to accommodate the foundation's needs, site concerns, or site-specific changes in the underlying data. This is an example of one type of challenge we faced: periodic changes in the data collected by the sites or in how such data were provided. These and other challenges are discussed in subsequent sections.

Challenges Related to Defining Metrics and Ways to Address Them

As noted earlier, the metric development process involved collaboration among the foundation, the sites, and the evaluation research teams. The process was complicated because it required taking into account different standards, test regimes, and data formats so that a single metric could be reported across sites. This process generally entailed a series of discussions about which measures should be

FIGURE 2

Snapshot of Dashboard Chart



NOTES: The dashed line between 2014 and 2015 reflects a change in the state assessments. Scores in 2015 are not directly comparable to scores in 2014 because the assessments align to different state standards.

computed and exactly how these computations should be done. On the surface, this might seem straightforward, but there are many complexities in tracking students and teachers over time that had to be addressed for each metric. The negotiations included considering what standards would be appropriate and fair for assessing progress and making the metrics comparable across sites located in different states with different organizational structures (i.e., school districts and CMOs) and consistent with different internal accountability systems. Some metric definitions had to be scaled back because of a lack of high-quality administrative data needed to define those metrics.

As noted earlier, the dashboard metrics provided an organizing framework for the annual stock-taking meetings between the sites and the foundation. The dashboards were intended to be used as a tool to capture progress on key indicators, both as a way to develop a shared understanding of improvement and to prompt collaborative reflection, inquiry, and action. One of the challenges in developing metrics across sites is that stakeholders have different, subjective views on which measures are best for these purposes.

To illustrate how decisions were made during the metric development process, we describe the development of three metrics: *college readiness*, *access to effective teaching*, and *career ladders*. The first two metrics are part of the excellence and equity categories, respectively, that were shown earlier in Table 2. We were unable to develop an effectiveness metric for selective hiring for career path positions (i.e., career ladders).

College Readiness

Context

One way in which schools and districts measure academic success is by assessing how many of their students are prepared for college education. The foundation wanted to create a metric of college readiness that could be used to compare each site's progress over time and compare the sites with one another.[C] Foundation leaders decided to use available standardized college preparation tests to measure readiness across sites. Next, we had to decide on the tests that would be included. The Scholastic Aptitude Test (SAT) and American College Test (ACT) are

[C] For the dashboard, we needed a straightforward indicator that was widely available, although we recognize that college readiness is multidimensional and that there are various methods to measure it (Porter and Polikoff, 2012).

administered nationally, and all sites encourage their students to take at least one of them. Students in California also received a score on the Early Assessment Program (EAP) when they completed the Smarter Balanced Summative Assessments,[D] which were administered statewide. The foundation proposed one set of scores, which received some pushback from the site stakeholders. After discussion among the RAND team, the foundation, and the sites, we agreed that a student would be counted as "college ready" if they obtained a total or composite score above a certain level on at least one of the college readiness assessments: the ACT, SAT, or EAP in California.

Once the general framework was established by working with the stakeholders (Raftree, 2015), we had to decide on the specific scores that would represent college readiness on each test. Decisions had to be made to balance the preferences of different stakeholders and accommodate a variety of data formats and academic testing regimes. It was important to be responsive to the concerns of the site leaders and engage them in the process so that the dashboards would contain indicators they trusted. Site and foundation stakeholders had concerns about which tests would be eligible for inclusion and how high the scoring standards should be. We describe some of the key challenges we encountered and the solutions we arrived at in the following section.

Challenges and Ways to Address Them

Stakeholders in Florida and Tennessee suggested that we use the college admission requirements for the SAT and ACT that are listed by the universities in their respective states to define college readiness. Unfortunately, these two sets of standards were not the same. Foundation stakeholders wanted comparability across sites and thus preferred a single standard for each test. One of the site stakeholders felt that the originally proposed cut points should not be used as a measure of college readiness in their district because the cut points were significantly different than state admission standards. Without a clear consensus on how to measure college readiness (a problem noted by the U.S. Department of Education [undated]), it was difficult to come up with an approach that satisfied all parties. After negotiating with the sites and the foundation, two different standards for readiness for each test were created: basic and substantial.

In addition to different readiness standards for the SAT, ACT, and EAP, students in different states were more likely to take one test than another. In some sites, the SAT was more prevalent; in other cases, the students were encouraged to take the ACT (Saget, 2013). When reading and math were combined, the SAT scores ranged from 400 to 1,600. The ACT score was given as both a composite score and scores for each subject area, each with a maximum of 36. EAP scores were given in terms of three possible levels. Given the different scales and scoring distributions of each test, we had concerns about the comparability of readiness across sites that used different tests.

Foundation leaders and the RAND team decided to report multiple cut points for each assessment and to use whatever testing data the sites provided. This allowed us to observe site trends at multiple levels of readiness while allowing the site stakeholders to focus on the readiness levels they thought were most appropriate. However, the issue of cross-site comparability because of different tests was not resolved. College readiness comparisons across sites were

It was important to be responsive to the concerns of the site leaders and engage them in the process so that the dashboards would contain indicators they trusted.

[D] The Smarter Balanced Summative Assessments are end-of-year assessments for ELA and math that measure progress toward college and career readiness.

restricted to comparisons of site trends rather than absolute readiness levels.

Access to Effective Teaching

Context

The metric *access to effective teaching* was designed to indicate whether effective teachers were distributed equitably between LIM students and the rest of the student population. There are many ways this might be expressed, but because the sites all developed similar (although not identical) measures of effective teaching and used the TE measures in their internal teacher evaluation systems, this measure seemed like the logical place to start. The metric we defined compares the proportion of LIM students with non-LIM students who have at least one teacher who was deemed highly effective in the preceding school year; it is calculated separately based on the effectiveness of teachers who taught mathematics and reading/ELA. For all student-teacher links in a subject, the procedure identifies the prior-year effectiveness of each teacher and assigns a 1 to a student if any of their teachers were highly effective or a zero if all teachers were not highly effective. A similar process is used to calculate a metric that measures access to ineffective teachers.

Challenges and Ways to Address Them

Access to effective teaching is one of the most important metrics in the dashboard because it is a key IP initiative goal. As a result, it was essential that the metric we developed supported valid cross-site comparisons. However, it was difficult to achieve comparability because of differences in the academic systems and data-collection procedures in each site. Two conditions were most problematic. First, there were important differences in the way sites kept track of students who were enrolled in a class for less than a full term (e.g., they transferred to another teacher), which forced us to come up with a way to adjust the computation when students were associated with teachers for different lengths of time. Second, (and reminiscent of Raftree, 2015) although all sites developed similar TE measures and classified teachers into levels based on effectiveness, they did not all use the same number of levels or comparable cut points for assigning teachers to those levels.

Our first challenge, as discussed above, was that the sites had different systems for maintaining information about which students were taught by which teachers. Some sites kept track of every time a student changed sections; others only kept student-teacher links that remained after an established validation date for each semester (e.g., five weeks from the beginning of the semester). We restricted student-teacher matches to those involving at least four months of instruction. Four months was chosen because it accounts for a significant portion of one semester. If we had included all links, those with schedule changes at the beginning of the semester would have been linked to more teachers; therefore, they would be more likely to be associated with at least one ineffective or highly effective teacher.

The second challenge related to differences in effectiveness levels. As we described earlier, administrators and staff at each site were expected to develop a shared understanding of effective teaching and a way to measure it that included at least two elements: a direct measure of teaching practice (e.g.,

[A]lthough all sites developed similar TE measures and classified teachers into levels based on effectiveness, they did not all use the same number of levels or comparable cut points for assigning teachers to those levels.

administrator rating of teaching using a structured observation rubric) and a measure of each teacher's contribution to student achievement growth (e.g., a value-added measure based on standardized achievement test scores). The TE measure could include other elements, such as feedback from student and parents. There was considerable variation in the approaches the sites used to collect and combine elements into an overall TE score, although all TE measures included the required two components (Stecher et al., 2018; Stecher et al., 2019).

Once the local TE measures were calculated, the sites designated different numbers of levels of effectiveness (i.e., different numbers of cut points on their TE measure). Some sites created four levels (e.g., novice, intermediate, effective, advanced), while other sites used five levels to categorize effectiveness. (For additional details, see the appendix.) Although these differences might not seem that important, the TE measures were used for teacher evaluation and thus had real consequences for teachers in terms of job security and salary. The levels had practical meanings in each site, and the sites were protective of their designations. One could not equate "novice" in one site with "beginner" in another without careful discussion and agreement.

Further complicating efforts to create this metric was the fact that there was considerable variation across sites in the distribution of teachers among the effectiveness categories. In one site, by design, the highest category contained just 1–2 percent of teachers. In another site, 30 percent of teachers were in the top category. Although everyone agreed that access to effective teaching should be based on each teacher's site-determined effectiveness level, it was not at all obvious how to use the available information on effectiveness levels to create a comparable metric across sites. We held several conversations with the sites and the foundation during which we compared the descriptions of the effectiveness levels and examined the distributions of effectiveness across sites. Eventually, we reached agreement on two metrics, one of which focused on highly effective teachers and the other of which focused on ineffective teachers. We also agreed on which categories in each site would be included in the calculation of the access to effective teaching metrics. Although the definition we adopted had some good features (e.g., the effectiveness designations within sites were generally consistent over time), it had some problems. Most problematic among these challenges was the fact that these two metrics were not directly comparable across sites.

We also noticed that the effectiveness ratings tended to improve over the years in some sites far more than in others, as more and more teachers were classified into the top categories. This raised some doubts about the comparability of the two metrics. To achieve comparability across sites and better relative comparisons within sites, we began using deciles in addition to categories. Given that the categorical effectiveness levels are not comparable across sites, we used the underlying distribution of the effectiveness scores as a way to develop the highly effective or ineffective metric. Specifically, instead of looking at the top and bottom TE categories, we used the top 10 percent and the bottom 10 percent of all teachers receiving a TE score on the underlying effectiveness scale to determine highly effective and ineffective teachers, respectively. This method resulted in a metric that was more comparable across sites and that was not subject to changes in the overall TE scores over time (e.g., scores increasing substantially). It also provided a more refined measure of the relative performance of teachers within a site, allowing for more-precise differentiation between the higher- and lower-performing teachers. Finally, it was easier to interpret: On average, about 10 percent of students are expected to have had a top-decile math teacher. Anything above 10 percent for a given group of students suggested that that group had greater exposure to highly effective teachers compared with another group of students. However, the decile-based measure of access was less meaningful to the sites because it did not align with any site's internal teacher evaluation system based on the four or five effectiveness levels.

The next step in the evolution of this metric was to use subject-specific distributions instead of all TE scores. It is possible that, on the whole, math teachers had higher average effectiveness scores than the other teachers. This was more likely to occur in smaller sites with fewer teachers where the distribution of effectiveness was not as smooth. Using effective math teachers as an example, we used a method where only those teachers who taught a math course were used for the math metric. In this method, the top 10 percent of math teachers were used for the top decile. Using the subject-specific restriction did not substantially change the results.

We were not able to solve all the challenges associated with the access to effective teaching metric. For example, the number of courses per student varied by site because in some sites, students change courses each semester, while in other sites, they change courses annually. In the former case, a student might have had 16 one-semester courses; in the latter case, they might have had eight year-long courses. This suggests that the former student might be taught by a higher number of teachers in a given school year. In this case, any metric that is defined in terms of having at least one teacher of a particular type might be inflated for sites that tended to have shorter course lengths.

In Table 4, we show the average number of courses and teachers per subject for 7th grade students across all sites from 2011 to 2016. The numbers range from about one teacher and one course per student in site 2 for both ELA and math to 3.5 courses and about two teachers in site 7 for ELA. There also were differences between subjects within a site in some cases. Barring major scheduling differences from one year to the next, the access to effective teaching metric was measured consistently within sites.

As this analysis shows, any metric based on exposure to a particular type of teacher will be affected by school scheduling and teacher assignment practices. Despite a great deal of discussion with site leaders, we were unable to resolve this challenge entirely. As a result, the access to effective teaching metric was more valid for showing within-site trends over time than for comparing equity across sites.

Career Ladders

Context

One of the goals of the IP initiative was that sites would create career ladders to give more-effective teachers greater leadership opportunities; for example, teachers might serve as mentors for part of the day to help other teachers improve their practice. In a career ladder, there is a progression of jobs with increasing responsibility (and increasing pay), which would help sites retain their most-effective teachers. Foundation stakeholders hoped to be able to define a metric that would assess the degree to which more-effective teachers were being provided such opportunities; specifically, the percentage of teachers advancing to career ladder or teacher leader positions with highly effective ratings, with the rationale of measuring the selectiveness of hiring for career ladder or teacher leader positions. However, as we examined the personnel data more closely, we could not identify sites where career ladders existed. A career ladder was envisioned as a job path of increasing responsibility that would be clearly indicated by job descriptions and information on promotions or compensation increases, much as government positions of relative seniority are indicated by levels I, II, or III. However, the personnel data we received from sites did not reveal clear linear paths. Although some teachers moved from classroom teaching to administrative roles, there were no clear job titles that indicated seniority or experience among teachers. We tried to compare job descriptions with compensation files to see whether that would reveal a link between a path of greater responsibility and a higher salary, but

TABLE 4

Average Number of ELA and Math Courses and Teachers Among 7th Graders Across Sites, 2011–2016

Type of Course or Teacher	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8
ELA courses	1.97	1.00	1.72	1.46	1.11	1.43	3.49	1.00
ELA teachers	1.02	1.04	1.97	1.18	1.15	1.45	1.99	1.17
Math courses	2.25	1.00	1.34	1.28	1.09	1.37	2.91	1.08
Math teachers	1.18	1.05	1.65	1.08	1.08	1.39	1.70	1.36

that also failed to get us closer to a reasonable way to define a career path metric.

Challenges and Ways to Address Them

Despite the availability of detailed personnel data, we had difficulty defining and measuring this metric because of the way the relevant data were generated and stored. This situation occurred in all the sites; none of the personnel files had an identifiable hierarchy of teaching job titles, roles, and responsibilities. Although career ladders were part of the IP initiative, they were not one of the levers that sites implemented at the beginning of the initiative. After discussion with the site leaders and a review of existing data, we reached the empirical decision that the personnel data were insufficient to identify existing career ladders. This is not surprising given that hierarchical career ladders were never fully realized by the sites. In fact, after three or four years, the evaluators learned that most of the sites had decided not to implement complete career ladders. This metric seemed calculable when the metrics were outlined originally. Only after reviewing the data did we determine that sites either were not creating such positions or were not tracking career progression among teachers in a way that was conducive to the metric's calculation.

Summary

In Table 5, we summarize challenges that arose in developing metrics and present some recommended ways to address them; these detailed recommendations expand on the general suggestions made by, among others, Raftree, 2015; Matheus, Janssen, and Maheshwari, forthcoming; and West, 2012. Some of the recommendations reflect things we did in our project, and those items are shaded in the table. The unshaded recommendations reflect suggestions that we did not use in our project but that could be tried by other researchers, data scientists, and practitioners.

Challenges Related to Collecting Data and Ways to Address Them

In addition to the challenges related to defining metrics, we encountered inherent data-collection challenges when coordinating data transfers between

TABLE 5

Challenges Related to Developing Metrics and Recommended Ways to Address Them

Challenges	Recommended Ways to Address Them
There are differences in subjective views about how to operationalize a general construct, such as effectiveness	Conduct conversations with stakeholders to come to a shared understanding about key constructs
There are differences in existing locally adopted definitions (e.g., the definition of college ready)	Create and report multiple designations that correspond to familiar policies
Sites measure a similar construct (e.g., effective teaching) in different ways, making comparisons between sites challenging	Balance the metrics between those that allow for comparing across sites and those that are more valid for within-site trends
Different tests and scales perform the same function (e.g., SAT, ACT)	Look for research or a practical basis for equating across scales; look for research-based correspondences
There are differences in the distribution of performance levels based on differences in the strictness of cut points (e.g., percentage of highly effective teachers)	Use percentiles or other relative designations rather than cut points that cannot be compared
There are different standards for course length, length of the school day, number of courses per student, and other underlying educational elements	Work carefully with local administrators to understand conditions in the schools; just because you can compute the same indicator in two settings does not mean that the two versions will be comparable
There are insufficient levels of detail in files (e.g., teacher job classifications)	Realize that not all constructs might be present and that it might be better to obtain descriptive information from site leaders about the implemented program than to try to extract indicators from extensive data files

a variety of different organizations. These challenges were amplified when collecting data continuously over a period of years. We collected primarily administrative data, including fiscal information, personnel data (e.g., employment history, compensation, ratings of effectiveness), and student information (e.g., demographics, enrollment history, standardized test scores). Most of these data were collected by the sites, but some data, such as standardized test results, were collected by states. Other important sources of data that were relevant to the dashboard included teacher and school leader surveys, which were collected by AIR; classroom observations, which were conducted by trained teachers and administrators in the sites; and student and parent satisfaction surveys, which were developed by outside organizations and collected by the sites.

In this section, we discuss how the timing of data releases and changes in school district data systems and data-collection procedures can affect researchers' access to data. We also describe challenges related to using national data sources. Information on college enrollment and scores from college admission tests are examples of these national-level challenges and are discussed in detail.

Timeliness and Format of Aggregate State Data

Besides the foundation, stakeholders at the sites were the main consumers of the dashboard. Their review of the dashboard occurred at annual stock-taking meetings, which typically occurred in the fall of each school year. Three of the dashboard metrics were based on results from state achievement tests, and two metrics involved comparisons of proficiency rates by subgroups at the site and state levels. We relied on the timely processing and reporting of state achievement test results to be able to compute these metrics in time for the stock-taking meetings. Unfortunately, even for a given state, both the timing and formatting of aggregate data releases changed from year to year in ways that made it hard to produce consistent and timely dashboards. These changes were particularly difficult for making cross-state comparisons. For instance, we wanted to compare the same subgroups (e.g., grade, sex, ethnicity) at the same point in time in four states, but this was not always possible because of the timing and reporting structure of state data.

Idiosyncratic uncertainties inherent in the state-provided testing data led to variation in the timeliness of the production of dashboards. In some years, we delayed the production of the dashboards, while in other years, we produced incomplete dashboards and issued updates when the data became available. This meant that in some years, the foundation and the sites had all the metrics for discussion during their stock-taking meetings, while in other years, the excellence-based metrics were not available. It was a challenge to balance the flexibility to respond to delays in the availability of state data with timeliness to make the metrics available for decisionmaking.

We suggest that other data analysts be cognizant of this potential problem and do their best to plan around state data release dates. Research projects that rely on current state assessment data should give thought to the timing of the tests and the release schedule of the aggregate data. Creating dynamic dashboards with regular updates based on new state or district data could be a potential solution, but one consequence of doing so might be initial errors because data tend to improve in subsequent releases.

Changes in Longitudinal Data, Systems, and Needs

The metrics developed for the project are tracked over time to help the foundation and the sites monitor progress. We calculated indicators annually over a 12-year period, from SY 2006–2007 to SY 2017–2018.

Over this long period, we encountered several changes that affected our ability to construct indicators that are defined in the same way and that made it harder to interpret trends. For instance, gradual or abrupt changes made by a district in the policies and procedures (e.g., changes in recordkeeping, test administration, privacy requirements, access to TE data) affect how indicators are measured over time. Another example is the adoption of Common Core State Standards (CCSS), which led to changes in state assessments that were used as the basis for some metrics (National Governors Association Center for Best Practices and the Council of Chief State School Officers, 2010). As a result, trend data are discontinuous in the year the underlying measure changes. Furthermore, states altered their respective testing regimens at different points in time, making cross-state comparisons difficult. Indeed, as shown in Table 6, cross-site comparisons of trends are discontinuous for more than one year (Achieve, 2013; CCSS Initiative, undated). Likewise, longitudinal analyses that were built on year-over-year changes were disrupted any time the underlying test changed.

We considered several alternatives: For example, crafting metrics around percentiles instead of absolute scoring levels was one way to address the issue. However, although percentiles might partially address the problem, they do not deal with the fact that the new and old tests might not be measuring exactly the same content, which leads to an irreconcilable discontinuity.

Using National Data Sources

There are distinct challenges associated with the use of national data sources. We encountered two such challenges when we tried to develop measures related to college enrollment and college readiness.

College Enrollment

The *on-time college enrollment* metric reports the percentage of students enrolling in college within five years of entering grade 9. For this metric, sites obtain college enrollment data either directly from the National Student Clearinghouse (NSC) or from a vendor (e.g., Beyond12). NSC is an independent, nonprofit organization that maintains a repository of student-level college enrollment and degree attainment data. However, there are several limitations to using the NSC data for measuring college enrollment that might lead to underreporting. These problems relate to incomplete participation by colleges, matching errors, and privacy concerns. Although 97 percent of all students are enrolled in degree-granting institutions in the United States that submit data to NSC, participation is voluntary and not all degree-granting institutions share their data (Dundar and Shapiro, 2016). Additionally, the algorithm the NSC uses to report enrollment back to districts matches records primarily on student name and date of birth. Matching errors because of name changes and typographical mistakes can introduce inaccuracy. Finally, FERPA is a federal law that protects the privacy of student education records, allowing both students and schools to block the sharing of enrollment information. Researchers have found that suppression of student records because of privacy laws leads to the underreporting of college enrollment (Dynarski et al., 2015).

SAT Revision

As noted earlier, the *college readiness* metric reports the percentage of students who achieved a certain level of proficiency on either the SAT or ACT (or, in California, on the EAP). In March 2016, the College Board changed its scoring system from one that

TABLE 6

Full Implementation of Common Core State Standards, by State

State	School Year of Full Implementation
Colorado	2013–2014
New York	2013–2014
Pennsylvania	2013–2014
Tennessee	2013–2014
California	2014–2015
Washington	2014–2015
Florida	Not adopted

SOURCES: Achieve, 2013; CCSS Initiative, undated.

ranged from 200 to 800 for both math and reading to a system that ranges from ten to 40 for the reading section. It also made modifications to both the math and writing tests. Fortunately, the College Board provided concordance tables to convert scores from one system to the other. However, because the cut points we used to determine college readiness for the dashboard were fixed on the basis of the old scale, we chose to convert the new scores to the pre-2016 system to maintain consistency within the dashboard.

Changes in Data-Collection Procedures Used by Sites

Sites maintain information on both teacher and student demographics, and we relied on these data for comparisons between students from different population subgroups. However, like many data sets, these databases evolved over time, and there were changes in how certain attributes were recorded or interpreted (consistent with Raftree, 2015). As these changes occur, researchers might need to adjust the context of longitudinal trends that compare subgroups of students. We provide several examples of changes in sites' data-collection procedures that affected our analyses and the solutions we implemented in the following sections.

Free or Reduced-Price Lunch Status

Each site maintained an indicator in the student data file for students who were eligible for free or reduced-price lunch (FRL). FRL status was determined based on information a student provided at the outset of the school year related to their household's income (U.S. Department of Agriculture, Food and Nutrition Service, undated). Many researchers used this indicator as a proxy for low-income status, as we did for many of the metrics (Baird et al., 2016). One site stakeholder told us that reported FRL rates among secondary students were lower than they should be because some secondary students were not filling out and returning the forms. This, along with high percentages of free lunches already provided, led to a proposal that entire schools be considered for FRL regardless of individual student information. This reduced researchers' ability to differentiate students by socioeconomic status within schools. Also, it made trends before and after the change related to low-income status more difficult to interpret because changes in FRL percentages possibly were the result of both how this variable is reported and demographic shifts in the student population.

To address this shift in reporting policy, we considered using a student's FRL status from the previous year. One downside to this approach is that it would miss a change in FRL eligibility based on an actual change in the financial situation of the student's family. In addition, this retrospective approach will be less useful in subsequent years as new students enroll in the site. If sites begin collecting economic data other than FRL status, it would provide a better solution. During this project, one site began collecting an economic disadvantage indicator. This indicator is based on a survey administered at registration and on the presence of a student's family on the administrative records of public assistance. Furthermore, this indicator is more comprehensive and measures a student's low-income status more directly.

Race and Ethnicity Designations

The U.S. Census uses broad racial categories that include white, black or African American, American Indian or Alaska Native, Asian, and Native Hawaiian or Other Pacific Islander (Humes, Jones, and Ramirez, 2011; Office of Management and Budget, 1997). The census also records Hispanic or Latino origin regardless of race. For this initiative, *minority students* are defined as those of Hispanic or Latino origin or those who indicated a race of either black or African American, or American Indian or Alaska Native. All the sites kept track of the race or ethnicity of all students, and they typically used the same categories as the census. However, in some sites, Hispanic or Latino origin was treated as another ethnic category, so, for example, students could not identify as both Hispanic and black. Some sites changed their reporting rules over time. For example, in one site, prior to 2010, students could report only a single ethnicity because that is all the data system would allow. After 2010, the data system allowed students to indicate all applicable ethnicities. As a result, the

Managing a diverse stream of education data over multiple years presents many challenges . . . .

percentages of minority students were higher after 2010.

When we encountered cases in which a student's race or ethnicity changed over time, we used the designation that was likely to be the most accurate and comparable and applied it to the other years. For example, in the case mentioned above, we applied students' ethnicities as reported in 2011—when the data system was more consistent with the census designations—to their records from earlier years. Of course, this approach is useful only if a student has a record in the data before and after the change in collection procedures.

Changes in Score Reporting

Data measuring student progress through such achievement and assessment tests as the ACT, SAT, or end-of-course tests (EOCs) came from sources outside the sites. It was not uncommon for a student to take the same test multiple times and possibly in different grades. This testing pattern resulted in variation in how the scores were reported for both types of tests. In some years, the data we received were inadequate to compute the metrics given the change in reporting.

In two cases, site stakeholders told us that in consecutive years they received data from the testing organization that contained different information. When requesting test scores from either the ACT or SAT, the site submitted a list of students (normally its graduating HS seniors) and received students' highest scores as of that date for that year. In both the prior and subsequent years, the site received all scores corresponding to all tests students took during that school year. Under the first reporting method, we received one score per student but were assured that it was their best score. Unfortunately, with this method, some students who took a college readiness test in that year would not have been included in that year's data set unless they were graduating seniors. In the second method, we received multiple scores per student, irrespective of their grade, but we only received test scores from one school year. For our purposes, we analyzed students' best scores from college readiness tests over a four-year period.

Researchers must verify that the scoring report method fits within the aims of the study. At the outset of the project, we did not know the exact reporting method a site would use or that the method might change over time. Had we known, we would have advocated for receiving all scores from all tests a student took in a given year, which was the preferred method for our purposes. Explicit discussions on how standardized data, such as tests, are reported at the outset of the project can reduce the possibilities for inconsistent reporting through the project's life.

Summary

The examples above highlight key challenges related to the collection of data. How data analysts might address these challenges are summarized in Table 7 (which elaborates on suggestions from Raftree, 2015; Matheus, Janssen, and Maheshwari, forthcoming; and others). The shading reflects the fact that the recommended changes were things that we tried as part of the project.

Challenges Related to Managing and Standardizing Data and Ways to Address Them

Managing a diverse stream of education data over multiple years presents many challenges, and the bulk of our examples of challenges and solutions fall into this category. As with metric definition and data collection, metric comparability across sites remains a challenge for data management. In this section, we discuss approaches we used to accommodate missing

TABLE 7

Challenges Related to Data Collection and Recommended Ways to Address Them

Challenges	Recommended Ways to Address Them
State data release schedules vary, affecting the timeliness of data processing and reporting	Make proposed reporting schedules contingent on the receipt of external data and plan for multiple reporting cycles with increasing levels of completeness so that stakeholders have results as soon as possible for decisionmaking
The database structures and procedures used by national organizations might change	Inform stakeholders initially that there are uncertainties about the consistency of external data and consult with them if important changes occur
Sites might change their internal data recording policies or procedures (such as demographic categories)	Communicate often with site administrators to increase awareness of potential changes that might influence analysis or reporting; inform stakeholders about how such changes will affect reporting

data, handle student identifier (ID) problems, and navigate issues of data reliability and validity.

Data Comparability Across Sites

Different data were available in different forms from each site; therefore, to compare progress across sites, we needed methods to standardize data across these disparate systems. For example, each site had its own student information system (SIS), each reflected local and statewide norms in terms of curriculum and standardized tests, and each independently created a method for measuring TE and assigning those measurements to rating categories. In the following sections, we provide examples of procedures we used to standardize information pertaining to course identification and student enrollment or withdrawal from school.

Standardizing Course Categories

As we discussed earlier, *equity in access to effective teaching* was one of the primary project metrics. We computed this metric in both mathematics and reading/ELA, which required identifying all courses and sections that fell into these subject categories. For this calculation, each student was linked with one or more teachers to determine whether they were taught by a highly effective teacher for a given subject. We also needed to classify courses into subjects as part of the sampling of teachers to complete surveys; for survey sampling, we wanted representativeness across several dimensions, including the subject matter taught. Furthermore, the teacher value-added (TVA) models needed to link teachers to students' state assessment scores based on the subjects the teachers taught.

Sites provided a course category (e.g., science, reading, world languages) in the course files. Normally, mathematics and ELA were identified as primary course categories; they are the subjects most-frequently examined for student progress and teacher assignments. Unfortunately, how sites arrived at those course category designations were neither straightforward nor consistent. Course files included records for all sections and students, and were exported directly from SISs, including some hard-to-categorize sections. For instance, some courses were attached to labs (e.g., Algebra I might have an associated Algebra I Lab). It was unclear whether instruction occurred in the lab section or whether it was effectively a study hall where students performed the work assigned in the main section. Depending on the site, both sections might be assigned to the math course category or just the main section. Without more information on the course, we could not determine which attribution was correct. As noted earlier, if two teachers were listed for each section, metrics capturing access to at least one particular type of teacher (e.g., highly effective) could be inflated. In modeling that is based on the amount of instruction, or time spent in the classroom, *dosage* depends on the decisions of how to handle ancillary course sections.

Sites varied in the degree of granularity used to identify course categories. Some sites identified only a handful of categories (e.g., reading, mathematics, science, social studies), while others had as many as 30 categories. This was a concern primarily when identifying ELA courses or teachers. For one site, spelling, composition, grammar, reading, English, and/or literature courses were separate course categories, while another site used ELA for all these courses. We found that, for some research questions, more attention to the categorization based on course names was required to ensure comparability across sites.

One would think that the "other" category would include courses that did not fit neatly into a categorization, but course sections (based on their names) seemed to be suited for broad course categories. For example, in one site, elementary courses were not given the same course categories as secondary courses; instead, they were given a general K–5 catch-all category. We had to make judgments about which courses and teachers should be considered for math and ELA based on the course names and knowledge of the site.

As much as possible, we used the course categories provided by the sites. If that was insufficient, we used the more detailed description in the course names. In some cases, we observed course enrollment patterns to better understand how to categorize courses (e.g., Was Algebra I Lab offered as a stand-alone course, or did it appear only when a student was enrolled in Algebra I?). In addition to the steps we took to standardize courses, we iterated with the sites to obtain more detail on the course information when necessary. As the capabilities of SISs are enhanced, more-detailed course information might become available. For instance, if information on assigned textbooks associated with each course were available, we would have been even more specific in identifying comparable courses across sites. We also could have used more-specific course descriptions that detailed course meeting schedules, content, and related sections—such as labs—that usually are paired with a main course.

Standardizing Enrollment and Withdrawals

Several metrics related to achievement and educational attainment were based on a ninth-grade cohort of students (e.g., four-year graduation rate). The definition of these student cohorts depended on accurately identifying students who were currently enrolled and removing students who transferred out of the site for reasons that were outside the site's control.

Student transfers were identified using withdrawal codes and reasons contained in student administrative data. Sites developed withdrawal codes that met their administrative needs, but these codes had slightly different meanings across sites. For example, a transfer because of a disciplinary reason was counted as a transfer by one site but was considered an expulsion by another site. In addition, within-site changes in withdrawal codes across time required standardization. One site had a text field describing withdrawals that was unique for each entry (and likely was filled out by each school's administrator). This site eventually changed from free-form text entry to specific withdrawal codes.

We conferred with the sites several times initially and then on an as-needed basis as codes changed on how we planned to categorize each withdrawal code, giving them the opportunity to correct our understanding of how a particular code was used in practice.

Missing Data Patterns

We anticipated that there would be challenges with categorization and standardization at the outset of the project, but we encountered some missing data problems that we did not foresee. Some unforeseen patterns in data missingness emerged that led to changes in analysis and reporting. We discuss one pattern below, where the missingness appears to be related to characteristics of interest. These are not problems with data collection or reporting; rather, they are idiosyncratic patterns of missingness in a site or for specific subgroups.

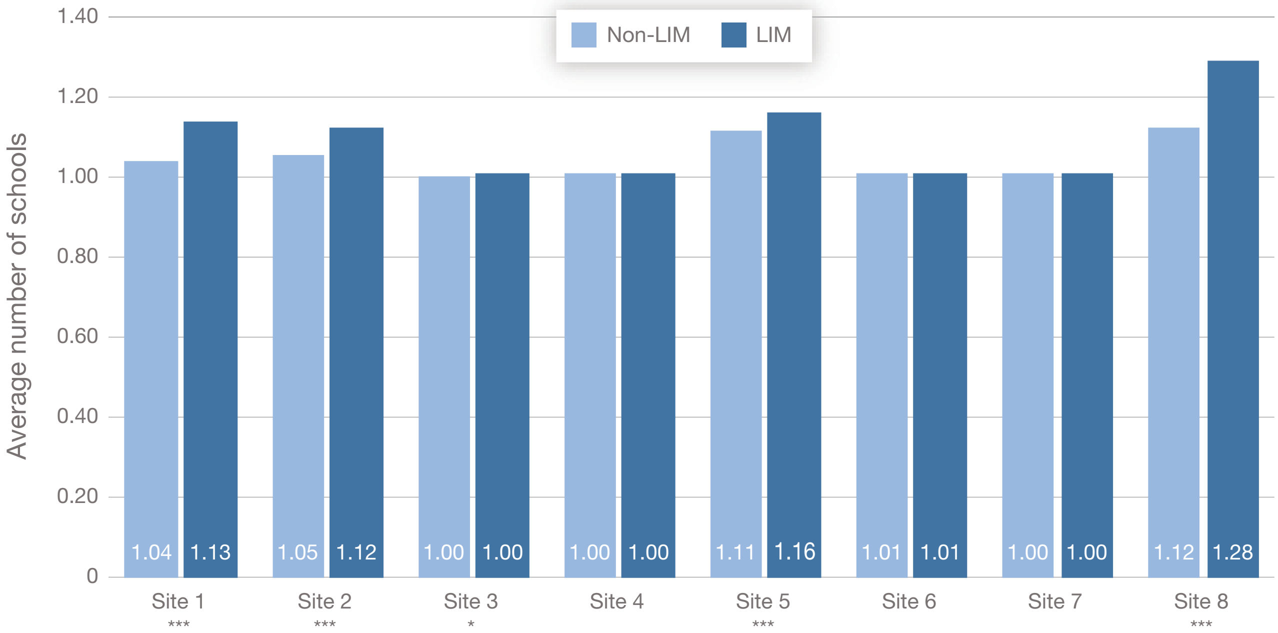
Understanding Missingness Patterns

Research involving student growth and graduation is inherently longitudinal, and analyses depend on students being tracked from one year to the next. However, transient students and/or key subgroups were less likely to be included in the calculations for cohort-based metrics we calculated (such as college enrollment within five years of entering ninth grade). In our case, LIM students were a key subgroup because several metrics in the dashboard used LIM status to disaggregate measures of student achievement and progress. If a key subgroup disappeared from longitudinal data sets, it could have led to systemically underrepresenting groups in panel analyses. These were not cases where a student took a test but for some reason the test score was missing, which is a standard kind of missingness. Rather, these previous year scores were not present because the student was not enrolled in the school district or CMO in the previous year. All similar students would have missing values for the previous year's test score, which is commonly used in TVA models (Mihaly et al., 2013; Hock and Isenberg, 2017; McEachin et al., 2018).

To understand the extent of this issue, we examined several measures for seventh and eighth graders who were identified as LIM students in the sites' enrollment files and for their non-LIM peers. Specifically, we examined the average number of schools seventh graders attended, school enrollment periods for seventh graders, the number of ELA teachers for seventh grade, and the percentage of eighth graders who were enrolled at the site in the previous year (see Figures 3 through 6). We compared the averages of each measure by LIM status using *t*-tests.

Other studies on student mobility (i.e., transferring schools) have found that it is higher among economically disadvantaged students (Welsh, 2017; Colorado Department of Education, 2016). Figures 3 and 4 suggest that LIM students attended more schools and had more transfers within sites than non-LIM students. We counted the number of distinct schools for each student's enrollment record and the number of distinct enrollment periods. In sites 1, 2, 3, 5, and 8, LIM students had statistically significantly higher numbers of schools and enrollment periods than non-LIM students. In all cases except sites 3 and 6, LIM students also were taught by more teachers on average.

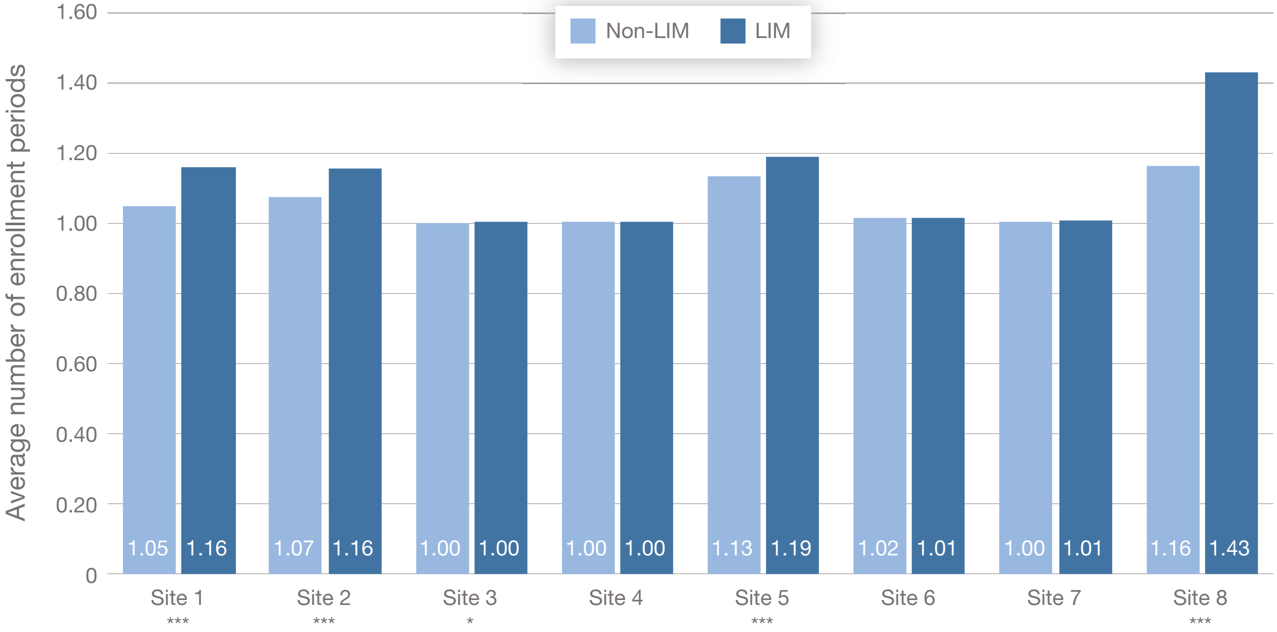
FIGURE 3
Average Number of Schools Attended, by LIM Status



NOTES: * $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.

FIGURE 4

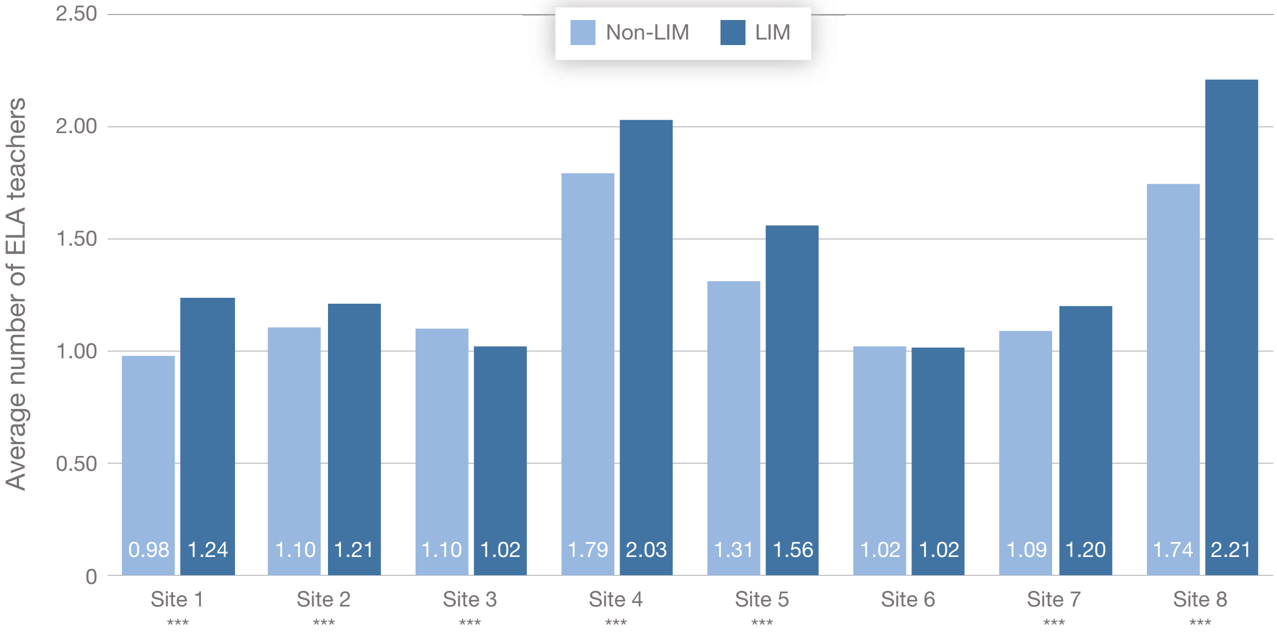
Average Number of Enrollment Periods, by LIM Status



NOTES: * $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.

FIGURE 5

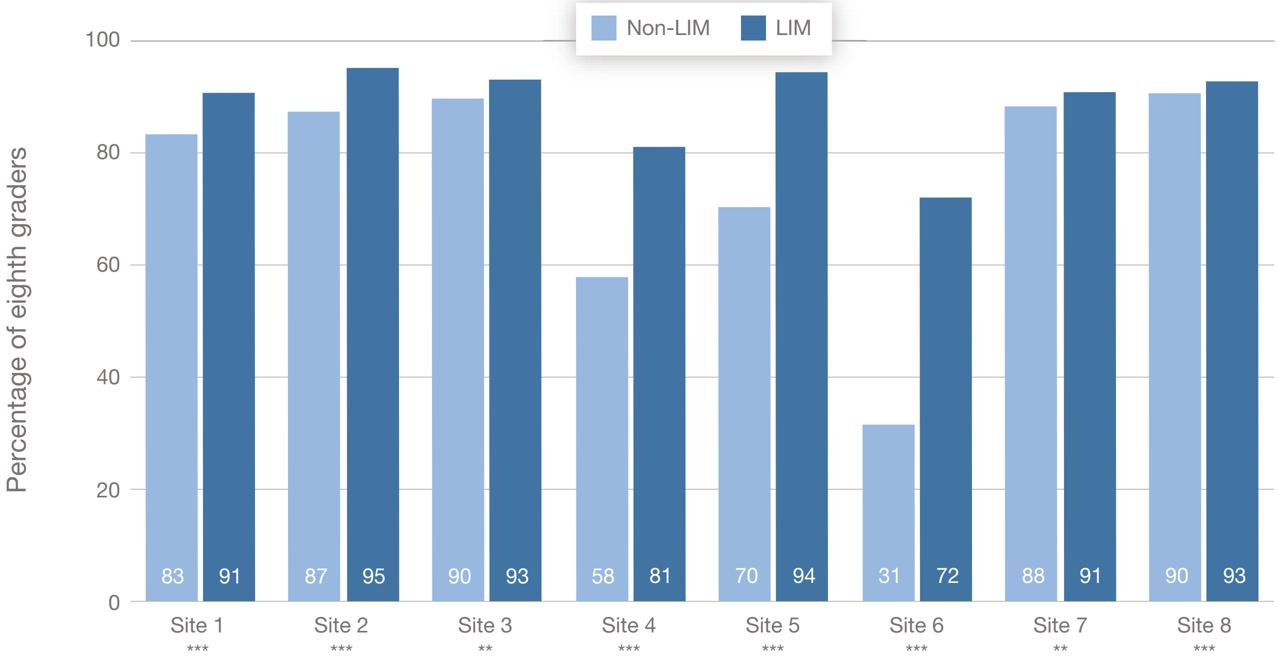
Average Number of ELA Teachers for Seventh Graders, by LIM Status



NOTES: * $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.

FIGURE 6

Percentage of Eighth Graders Enrolled at the Site in the Previous Year, by LIM Status



NOTES: * $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.

In contrast, non-LIM eighth-grade students were more likely to be absent from sites' data from year to year than LIM students. That is, non-LIM eighth graders changed school systems more frequently than LIM eighth graders. The pattern was particularly noticeable in sites 4, 5, and 6. More-complex analytic cases (e.g., cases where there are multiple teachers, multiple transfers, or lapses in year-to-year data) are sometimes dropped from analyses. These patterns suggest that non-LIM students were more likely than LIM students to be excluded from analyses that used data from consecutive years, such as analyses of student achievement growth, while LIM students were more likely to be complicated analytic cases because of their intradistrict mobility and subsequent multiple-teacher scenarios.

The data sets prepared for use in regression analyses often contain missing values. Cases with missing values often are dropped, although there are strategies for handling missing values in statistical analysis. Where longitudinal analyses can account for attrition using weights, our dashboarding process reported only on cases with valid, non-missing data. The conclusions drawn from such dashboards can depend on which students were included in the analytic sample (Zvoch and Stevens, 2005). Student subgroups that are examined separately in such analyses might appear in test or enrollment files at different rates. In our analyses, this was a consistent pattern without a straightforward solution. For researchers who are conducting similar longitudinal analyses, we suggest searching for patterns of longitudinal missingness by key groups (e.g., LIM students, English language learners, students with disabilities). If significant patterns are found, researchers can report details about specific subgroups that have different representation rates based on the available data (e.g., 80 percent of LIM students with assessment test data in consecutive years were included in the analysis, compared with just 60 percent of non-LIM students).

Identifier Issues

Data on students and teachers are kept and processed for different reasons, often by different organizations or departments in a school district or CMO. With so many goals and entities involved, data systems are not always integrated. The breadth of the data collected for this project necessarily involved working with multiple identifiers for each site. In addition to making these data sets linkable across different data systems, an issue identified by West (2012), protecting the identity and privacy of students and staff is a priority of research involving K–12 education for both school districts and researchers, as noted by Matheus, Janssen, and Maheshwari (forthcoming).

Multiple Identifiers

We found that a single student or teacher could have several identification numbers for a variety of reasons. Teachers might have one ID for managing student information, another ID for their personal human resources (HR) records, and yet another for tracking PD. Without existing crosswalks, it was challenging to link data that came from multiple sources and that were used for different purposes. Over this multiyear study, sites reorganized, schools merged, SISs changed, and ID systems that tracked students and/or teachers changed. We constructed ID crosswalks to link data sourced from multiple information systems; however, because the primary purpose of the data might not have been for research, some gaps in the crosswalks existed. For instance, in one site, the SIS issued an ID variable for each teacher who entered student data (e.g., grades, attendance). The teachers were assigned this identifier when they first used the SIS. Therefore, it was likely that only teachers received these IDs. In contrast, all district staff received IDs in the HR data system at the site. If, for some reason, a teacher did not enter information into the SIS or if the analysis included nonteachers, the crosswalk would not resolve the identification issue.

Although sites continued to update and make improvements to existing information systems, the primary purpose and use of the data often predated research goals. For instance, to estimate teacher exposure to students, multiple data sources might need to be used (e.g., teacher HR records and data extracted from a database designed to track student attendance). The database might not have been designed to be integrated with personnel records because its focus was simply to track students.

In other instances, privacy concerns led to the use of scrambled or omitted IDs, which complicated tracking individuals over time and across data systems. Protecting sensitive information is a priority when using administrative education data, especially as it relates to students. Names and identifiable information were removed from data files before analysis began and were replaced with research IDs.

Site Deidentification

Because of the longitudinal nature of the dashboards, the data files that were needed to produce the dashboards were collected annually throughout the project. Each time we received data, we deidentified them to protect the identities and sensitive information of teachers and students. The real identifiers were replaced with research IDs to be used in the analysis. At the same time, sites could take similar steps to protect identities.

Although there might be several ways to deidentify data, our data procedure linked each distinct teacher and student ID to one and only one research ID. As new teachers were hired or new students enrolled, their new IDs were added to the crosswalk and research IDs were assigned to them. Because of

> Although sites continued to update and make improvements to existing information systems, the primary purpose and use of the data often predated research goals.

staff turnover and new students, we expected that some percentage of IDs would be new each year.

Researchers and sites must communicate their expectations about what identifiers are being used in each file to ensure data integrity and reliability across files and years. Also, where possible, research IDs should be visually distinct in some way; for example, they can be of differing lengths (i.e., the research ID might have a length of ten characters and the real ID might have a length of eight characters) or the research IDs could have extra characters (i.e., all research IDs begin with a letter "R" to distinguish them from numeric-only IDs).

Reliability and Validity of Available Data

We encountered several cases in which available data that appeared to be suitable for analysis were determined to be invalid for research purposes for a variety of reasons (consistent with Raftree, 2015). We highlight cases where the data are present but where their meaning is ambiguous. We also show that the information contained in a variable might change from year to year, thus making the interpretation less reliable. Finally, we discuss data that, upon closer review, were not valid measures for the metrics for which they were intended to be used.

Years of Teaching Experience

Teacher experience was used to create categories for the dashboard metrics, teacher sampling, and analyses of outcomes. Most sites provided a measure of teacher experience using HR records. Sometimes, this measure reflected experience in the teacher's current position and sometimes it reflected total time in the school district. As an example, the metric on the performance of *new teachers* (defined as those with three or less years of teaching experience) tried to capture the extent to which sites recruited effective teachers by comparing new teachers' effectiveness ratings with the ratings of teachers with more experience. This led to two questions when comparing this metric within and across sites:

- **Did the total experience variable reflect total teaching experience or total experience teaching in the district?** The distinction might be irrelevant for administrative purposes, but from a research perspective, it is valuable to identify new and experienced teachers accurately. Unfortunately, sites typically only tracked career records within the district in these administrative data sources, making it nearly impossible to measure the experience a teacher had when first hired. In the CMOs, which had only been in existence for a few years, experienced teachers (based on the reported site-level experience) who recently joined the CMOs were grouped with inexperienced teachers for this metric.
- **How are years of experience being counted?** We found that there was variation in how sites reported these data. Consider teachers who have never taught before. These teachers have zero years of experience at the beginning of their first year and technically one year of experience at the end of the school year. Years of experience typically were recorded as integers. A site had either a 1 in the experience field for the teacher (because they were in their first year) or a zero in the experience field until the teacher completed a school year. Furthermore, we were told that some teachers had taken leaves of absence, and it was not known whether such leaves of absence were included or excluded from experience calculations.

Fortunately, these questions were mostly answered in communication with the sites about how their data were structured. Also, because the methods used to record teacher experience were consistent within sites, it was possible to track trends in the outcomes of new teachers. This highlights the importance of being as explicit as possible when defining the metrics and research questions. If we assume that 180 days represents one full year of teaching experience, will we treat the three-years-or-less experience requirement to mean that a teacher has between zero and 540 days of instruction by the end of the school year or that a teacher has no more than 540 days of instruction at the outset of the school year? By acknowledging that the data are stored in different ways (and in some ways that are less suitable for

We encountered several cases in which available data that appeared to be suitable for analysis were determined to be invalid for research purposes . . . .

specific research questions), we can mitigate confusion in understanding the calculated metrics and their meanings.

Course Section Enrollment Information

Excellence in student achievement metrics was measured by comparing state assessment scores from one year to the next. One research team used students' progress to measure the value added of the teacher who taught them. In these cases, we needed to determine how much instruction a student received and how much of that instruction was done by a particular teacher. However, this sort of dosage attribution was challenging to implement.

Several approaches to linking students to teachers have been identified (Hock and Isenberg, 2017; McEachin et al., 2018), and the most suitable approach often depends on the available data and common practices at the site level (e.g., the prevalence of co-teaching). Researchers need to make key decisions in implementing an approach when other abnormal cases complicate the analysis. Consider the following examples, in which the analysis of TVA was complicated by course section enrollment information:

- **partial enrollment in the school district.** A student appeared in 2011 and 2012 assessment data, but the student only spent the last quarter of 2011–2012 in the school district. Under one attribution method, all the student's progress would be associated with a teacher who taught that student only 25 percent of the time. Under another attribution method, only one-quarter of the progress would be associated with the teacher. Alternatively, if the instruction length failed to meet some minimum threshold, the student-teacher link would be ignored altogether in the analysis.
- **multiple concurrent course sections of the same subject.** A student had two classes in the same subject with two teachers over the school year. One course was remedial and the other was the grade-level course. During this period, the student received twice as much classroom instruction as most of that student's peers. Unfortunately, without knowing the content of courses, it is difficult to know whether both courses should be treated equally in terms of their effect on the student's growth.
- **multiple teachers teaching the same course section.** Two teachers taught the same section but at different times throughout the year. It was unknown exactly when one teacher's instruction ended and the other's began. Without this knowledge, we could not accurately determine how much of a student's progress was attributable to one teacher or the other. Hock and Isenberg, 2017, detailed approaches for handling students with multiple teachers, but in their examples, data existed on how long a teacher taught students.
- **enrollment period.** Course records indicated that one ineffective teacher (based on the teacher's rating from the prior year) taught a particular student. In reality, the student was only in the section for ten days. Depending on the way course enrollment and drop dates were stored (if the dates were stored at all), it would not be possible to determine precisely how long the student was in the section. Even if the actual ten-day period was known, how many days would be enough to consider this

The data collection method and quality need to be considered to resolve how exposure measures will be treated.

student as having been taught by an ineffective teacher?

These are just a sample of the cases we encountered. Similar issues stemming from student mobility and co-teaching models also introduced complexity. In addition, the degree to which sites tracked enrollment and teaching dates in course sections varied. Some sites validated enrollment at certain points throughout the semester and did not bother with the actual dates. Other sites used actual dates, but we were told that the validity of the dates was questionable. Still others recorded only the transfers and not the actual dates. Because of these issues, we used imperfect substitutes for course enrollment dates (i.e., school enrollment dates) or had to make assumptions (e.g., all enrolled students were enrolled for at least half the course).

Our solution is that the best course of action depends on the site's available data and the research question. Researchers could ask sites how best to determine how long a student was enrolled in a course; they might have other ways of obtaining this information. Alternatively, the research question or metric could be crafted to suit the available data, although this was not possible in a project such as ours, with so many sites and methods of collecting data. The data collection method and quality need to be considered to resolve how exposure measures will be treated.

Teacher Tenure

Many school systems provided teachers with a pathway to tenure, ensuring that teachers cannot be terminated without just cause. Where available, tenure generally was awarded based on some combination of experience, performance, and PD. Data systems tracking tenure typically included at least a binary indicator of each teacher's tenure status and some included progress toward tenure. Because tenure data systems were intended for internal administrative use and not for outside researchers conducting analysis, it was a challenge for researchers who lacked institutional knowledge about the school district to determine values for teacher tenure. Teacher tenure has been controversial in the past few years, and the process has been reformed or abolished for new teachers in some states and districts (e.g., North Carolina, Kansas, Indiana, and Florida) (Underwood, 2018; Thomsen, 2014; Gardner and Muehler, 2014). In the sites' cases, data systems were not updated, and hiring dates were necessary to determine which set of tenure rules applied to a given teacher. The termination process for tenured teachers depended on state legislation and contracts with local teachers unions, meaning that the job security provided by tenure varied across states and districts. We found future tenure dates in the data, potentially indicating the date the tenure would have been granted or the date the teacher would first become eligible for review.

As a result, we had difficulty evaluating data on progress toward tenure. As noted previously, years of experience was based either on years in the school district or years in the current position, while tenure could have been based on total career years teaching. Ultimately, we stopped calculating metrics based on tenure because the changes in the tenure process were too substantial to measure change while also meeting the initial aims of the metrics. It would have been difficult to anticipate the substantial changes to tenure across several states at the beginning of the project.

In this situation, available data (i.e., tenure dates) that were still being collected became invalid because of a policy change, even though the data were still being recorded automatically. In our communication with the sites, we learned either to request additional data to measure tenure or that the data were not available. At a minimum, tracking teacher tenure requires that site administrative resources provide data on which teachers are covered under the tenure rules. Tracking progress toward tenure or anticipating tenure eligibility might become more complicated and could require the institutional knowledge of the school district and its data systems. As policies change, researchers should verify that corresponding data and data-collection methods have been updated as well.

Professional Development and Coaching

PD encompassed a variety of activities geared toward improving teacher practice. We saw data related to taking online courses and modules and participating in arranged courses, seminars, or district-arranged instructional sessions. Throughout the project, sites worked to improve how they captured these activities, but the resulting data sets did not capture all PD activities systematically. For example, in one site in the earlier years of the IP initiative, online modules were recorded but with a different ID than the primary teacher ID. The teachers self-reported, and the site was not confident that all online courses were captured. Although the data-collection methods improved over time, comparing the effectiveness of PD longitudinally using data of differing quality was too problematic.

We were asked to identify coaches and teachers who had been coached, and to provide estimates of how much a teacher had been coached. Various sites told us that coaching was tracked differently at different schools and that many activities were informal. Although the site leaders were confident that teacher coaching occurred, we rarely received data on detailed coaching activities that were considered to be comprehensive. The activities associated with coaching and mentoring did not seem like activities that lend themselves to a data trail in the same way that explicit courses and sessions did. If these activities are to be analyzed, a deliberate plan of recording and detailing what activities qualify as coaching must be determined beforehand.

We also tried collecting applicant, hiring, and compensation data throughout the project. In some cases, we received data but the data were not in a suitable or sufficiently standardized format to be used in analysis.

As sites update their information systems, these data, including courses, online training, self-study sections, and collegial meetings, might be captured. The ability to integrate these data with HR data (including applicant, hiring, and compensation data) more broadly would facilitate this type of research.

Summary

The examples discussed in this section highlight key challenges related to the standardization and management of data over time and across sites. There are various types of educational data for which standardization is likely to be a problem, including course enrollment information, teacher tenure information, and information about PD. Additionally, there are ID issues that involve changes related to time and databases created for different purposes. Finally, key fields needed for research might be incomplete, missing, or populated with outdated data (e.g., fields that are no longer maintained).

In Table 8, we summarize challenges we encountered in standardizing and managing data and the recommended ways to address them. As noted earlier, the shading reflects recommendations based on things we did during the project. Cells without shading reflect suggestions that we did not try but that could be used by others facing these challenges.

The Importance of Confidentiality, Data Sensitivity, and Partnerships

In addition to data-collection and management issues, proper data reporting presents its own set of challenges. Political concerns at school districts can complicate the reporting of some metrics, and privacy concerns can present ethical issues. In the following sections, we discuss our approach to data security and privacy issues and how we established working relationships with the sites during the project.

TABLE 8

Challenges Related to the Standardization and Management of Data and Recommended Ways to Address Them

Challenges	Recommended Ways to Address Them
Course names and subject-matter categories are not consistent across sites	Work with sites to develop crosswalks between course titles, obtain detailed course descriptions and use them to classify courses into subjects, and work with SMEs to understand course content and make proper classification
It is difficult to track student enrollment across years, particularly to understand the reasons for student withdrawal	Work with sites to understand withdrawal codes and encourage them to standardize codes across schools and over time
It is difficult to develop representative longitudinal measures because different groups of students were missing over time to different degrees	Calculate the extent to which key subgroups of students are retained in longitudinal data sets to estimate their representativeness; when there is uneven missingness across groups, report on cohort retention as part of reports on progress
Because internal systems are developed at different times and for different purposes, teachers and students might have multiple identification numbers and sites might not have built crosswalks to be able to link them	Develop a single research ID and make it distinctive so that it is not easily confused with existing IDs (e.g., make it one digit longer or add a distinctive character) and work with sites to establish a consistent system for converting internal IDs to research IDs before transferring data to the research team
Site data on teaching experience are not collected consistently and might reflect total years of teaching, years of teaching at the site, or years of teaching in the current position; partial years might be rounded to whole years in different ways	Understand what information sites collect about teaching experience and request the information needed; it might be necessary to adjust analyses and reports to correspond to available data on years of experience
Metrics that link students and teachers and require multiple years of data, such as value-added measures, can be difficult to compute because students can enroll in multiple courses of the same type, multiple teachers can be responsible for a given course, students can transfer in and out of the course at different times, and some students do not persist in the district for all years	Understand exactly what data are available at the outset and try to develop metrics and analytic methods that reflect the available options, but recognize that there are no simple ways to address these challenges
Data on teachers' tenure statuses are not collected and reported consistently across sites; some states do not offer tenure, while others have changed tenure rules recently	Understand fully what the rules are about tenure and how records are kept before adopting metrics based on tenure
Most sites did not collect or retain data on teachers' participation in PD, mentoring, or coaching that would allow researchers to use the data for monitoring; although formal PD courses could easily be monitored, other forms of support, such as coaching, are more difficult to track	Collect these data more systematically to enable analysis focused on PD. Find out what systems exist for keeping track of PD received by teachers and whether additional mechanisms can be put in place at reasonable cost before deciding to include PD as a study variable

Protecting the Identities of Teachers and Students

FERPA protects student data, and researchers must assure school districts that they will put procedures in place to protect student identities during transfer, storage, linking, and reporting of findings based on student data. Compliance with FERPA requires a secure system to protect student identities while managing and analyzing data and reporting results. Although no student names are included in data sets or reports, it might be possible to identify a student by inference if, for instance, the reporting involves cross-tabulating by a demographic characteristic (e.g., ethnicity) where very few members of a group occur in the data. Appropriate precautions must be taken to guard against this.

However, privacy rules governing teacher data often are less strict and less clear. School districts routinely use teacher data to make management decisions, and many types of teacher data are made public for transparency purposes. For example, the public interest organization Transparent California publishes individual salaries for all public school district employees. Because many schools and districts list faculty names on their websites, the problem of identification by inference is greater for staff than for students. TE ratings are key parts of teachers' performance appraisals and are not made public. However, we needed access to these data because they were used extensively in constructing the dashboards. To guard against any inadvertent release of this sensitive information, we used extra precautions associated with personal information classified as sensitive in the RAND environment and we made sure that these data were reported only at a high level of aggregation. Similarly, we took extra precautions when we linked effectiveness ratings to survey responses and shared them with the evaluation research teams; we created a separate ID in place of the standard research ID so that no information from other files (e.g., staff, job, course teacher) could be used to link a response to an individual teacher.

Most research studies enter into DUAs with data providers to ensure proper data safeguarding. The DUA used in this project was a contractual agreement among the RAND Corporation, the foundation, AIR, and the sites that outlined the responsibility of all parties; it included a detailed list of the data elements that were to be shared. In addition, a data safeguarding plan was developed that required removing teacher IDs before data were given to researchers. The deidentification was done by the RAND team because the process might have resulted in errors if it were done inconsistently by different departments (e.g., the personnel office has teacher hiring data, the PD office has TE data, and the finance office has teacher salary data) or if the same office prepared the data differently across multiple school years. Also, when personally identifiable information (PII) was needed to contact subjects for the surveys, the research team provided the information needed to contact respondents using a separate unique ID to prevent the linking of other sensitive data.

Relationships with Data Providers

The RAND team worked with the IP sites for seven years. During that period, the sites inevitably experienced turnover. This created occasional complications because new staff were sometimes unfamiliar with the study, unaware of the procedures that had been established, or wanted to implement different processes. The RAND data team was fortunate to have continuity of staff throughout the project.

As noted by Raftree (2015), we recognized from the outset that it would be important to establish personal relationships with the site staff because we needed a close collaborative working relationship. We organized the RAND data team by assigning a primary point of contact for each site. Prior to submitting the first data request, we made in-person visits to each site. During these meetings, we interacted with the administrators and data managers and with each of the individuals responsible for primary data extracts. In these initial meetings with the site technical and research staff, we learned a great deal about their underlying data structure and the limitations of the data maintained by each site. Although the DUA contained a detailed description of the requested data, the additional knowledge of the content and structure of data systems enabled us to customize data requests to make it easier for the sites to provide us with the data extracts in a timely fashion. The result was a customized data request for each site that minimized the burden of data transformations imposed on the site. This helped establish a collaborative working relationship with each of the data providers. These meetings also provided the site personnel with an opportunity to ask questions and obtain a better understanding of the overall research effort.

> Because many schools and districts list faculty names on their websites, the problem of identification by inference is greater for staff than for students.

Developing personal relationships with site personnel and establishing a team approach for the data-collection effort provided direct project benefits. It improved the timeliness of data delivery and increased site engagement in working through data issues. Staff turnover at the sites did present a problem; for this reason, it would be beneficial to have multiple site visits. These visits would have continued to reinforce the collaborative working relationships with site personnel.

Although the foundation introduced the RAND data team to the key site personnel via email, our interaction with the sites was independent of the foundation. Our initial site visit included only the RAND data team and site personnel, and subsequent communications and interactions were primarily between the RAND data team and site personnel.

To avoid errors in reporting, we developed a prerelease feedback loop with the sites, allowing them to review all dashboard information prior to finalizing the dashboards. This feedback loop allowed the sites to update or correct some of the source data and provide useful context for some of the data points. This process improved the overall accuracy of the dashboard and helped build trust between the RAND data team and site personnel.

In 2017, the project was extended for an additional two years with continued data collection and dashboard production. There was a hope that restricted-use files would be produced and distributed for use by the broader research community through the Inter-University Consortium for Political and Social Research (ICPSR). Although all but one of the sites agreed to participate in the additional two years of data collection, several sites did not agree to the release of any public-use files. Despite our strong working relationship with each of the sites, the sites had little incentive to agree to the creation of the public-use files and expressed concerns about the potential risks in releasing data collected for the initiative. One obstacle to the creation of public-use files was the site leaders' concerns that the data would be misinterpreted and that the site stakeholders would not have the opportunity to comment or provide feedback on analyses done by other researchers. If creating public-use files had been a stated goal at the outset of the project, sites might have been more likely to agree. However, it would have been difficult at the beginning to foresee public-use files as a potential deliverable.

Summary

We encountered several challenges associated with data safeguarding and privacy, most of which we were able to address through thoughtful communication and careful data-handling procedures. In Table 9, we summarize those challenges and recommended ways to address them. Once again, shading reflects solutions addressed in the project, while unshaded cells reflect solutions that were not tried in our project but that could be used by others.

Conclusion

In 2009, the foundation launched the Intensive Partnerships for Effective Teaching initiative in three school districts and four CMOs to improve student achievement, graduation rates, and postsecondary participation. The RAND Corporation worked with the foundation to collect and warehouse data from participating sites and to produce annual data dashboards that presented quantitative information about key indicators of the progress of the reform.

Challenges and Recommended Ways to Address Them

The RAND data team encountered several challenges in the process of designing metrics, collecting the data from individual IP sites, managing and harmonizing the data across sites, and monitoring and comparing trends over time. These challenges and ways to address them are described in previous sections. Table 10 brings together all of the challenges and recommended ways to address them. As we did in each of the tables in the section summaries, we shade the recommendations we used to address challenges and leave unshaded those we did not use but that could be used by other researchers, data scientists, and practitioners. All of the recommendations can help other researchers, data scientists, and practitioners who are planning to use administrative data to monitor

TABLE 9

Challenges Related to Data Safeguarding and Privacy and Recommended Ways to Address Them

Challenges	Recommended Ways to Address Them
Confidentiality requirements, such as FERPA, necessitate DUAs and data safeguarding procedures that will add time and complexity to the project	Make personal visits to each site to establish working relationships with data providers, develop mutually agreeable procedures, and build sites' confidence in researchers' ability to handle sensitive data
Site personnel are likely to change over a long study, and replacement staff might be unfamiliar with established procedures or might want to change existing processes	Organize the data team with a single point of contact for each site and, if possible, visit the sites every few years to keep relationships current
Site leaders do not want to be surprised by the public release of findings they have not seen	Preview results with site stakeholders so that they can be prepared to answer questions that might come from news media or the community
School districts likely will be cautious about releasing student and/or teacher data for secondary analyses; they have legal responsibility for protecting the confidentiality of their students and employees	Begin discussing the potential for restricted-use or public-use files early in a project

large-scale evaluation efforts. Our experience provides suggestions at a much finer level of detail than those currently available in the literature.

Overarching Lessons Learned

In addition to the specific challenges and solutions summarized earlier, we identified some general lessons learned that can be shared with other researchers who work with school district data systems for evaluating improvement efforts.

Long-Term Projects Need the Flexibility to Accommodate Change

This project spanned nine years; as noted, several changes altered both the emphasis and practices. Over such a long time, the project scope can change, and the topics the metrics were initially intended to measure might no longer be priorities. Teaching effectiveness—measuring it, increasing it, and linking it to student outcomes—was a focal point at the outset. Although it remained a key part of the analysis, by the end of the project, we were examining other questions that could be answered using site administrative data through public-use files. Also, sites' SISs and data-collection methods naturally evolve. The changes associated with underlying data would affect the consistency of the computation of metrics. This effort began with several possible aims and research goals; this inherently made it easier to shift as data availability or priorities changed.

Success Might Depend on Establishing Good Relationships with Counterparts in District Offices

Data teams tend to focus on the details of data structures, transmittal procedures, and information processing. All these concerns about accurate data handling are essential to a successful data-based analytic project, but it might be equally important to invest in building positive working relationships with client data administrators and IT personnel. There were many times during this study where progress occurred because we had developed strong professional connections with our counterparts in the sites.

Think Creatively About Redefining and Rescaling Existing Measures to Create Comparable Metrics Across Sites

Often, it is simpler to consider indicators conceptually than it is to implement them. For example, it makes good sense to think about creating an indicator of effectiveness—everyone would benefit from tracking program effectiveness over time. However, it can be surprisingly difficult to do this meaningfully when sites use different measures of student outcomes. Is a two-point gain on one state's test equivalent to a two-point gain on another state's test?

TABLE 10

Summary of Challenges and Recommended Ways to Address Them

Challenges	Recommended Ways to Address Them
Challenges related to developing metrics	
There are differences in subjective views about how to operationalize a general construct, such as effectiveness	Conduct conversations with stakeholders to come to a shared understanding about key constructs
There are differences in existing locally adopted definitions (e.g., the definition of college ready)	Create and report multiple designations that correspond to familiar policies
Sites measure a similar construct (e.g., effective teaching) in different ways, making comparisons between sites challenging	Balance the metrics between those that allow for comparing across sites and those that are more valid for within-site trends
Different tests and scales perform the same function (e.g., SAT, ACT)	Look for research or a practical basis for equating across scales or for research-based correspondences
There are differences in the distribution of performance levels based on differences in the strictness of cut points (e.g., percentage of highly effective teachers)	Use percentiles or other relative designations rather than cut points that cannot be compared
There are different standards for course length, length of the school day, number of courses per student, and other underlying educational elements	Work carefully with local administrators to understand conditions in the schools; just because you can compute the same indicator in two settings does not mean that the two versions will be comparable
There are insufficient levels of detail in files (e.g., teacher job classifications)	Realize that not all constructs might be present and that it might be better to obtain descriptive information from site leaders about the implemented program than to try to extract indicators from extensive data files
Challenges related to data collection	
State data release schedules vary, affecting the timeliness of data processing and reporting	Make proposed reporting schedules contingent on the receipt of external data and plan for multiple reporting cycles with increasing levels of completeness so that stakeholders have results as soon as possible for decisionmaking
The database structures and procedures used by national organizations might change	Inform stakeholders that there are uncertainties about the consistency of external data initially and consult with them if important changes occur
Sites might change their internal data recording policies or procedures (such as demographic categories)	Communicate often with site administrators to be aware of potential changes that might influence analysis or reporting; inform stakeholders about how such changes will affect reporting
Challenges related to standardization and management of data	
Course names and subject-matter categories are not consistent across sites	Work with sites to develop crosswalks between course titles, obtain detailed course descriptions and use them to classify courses into subjects, work with SMEs to understand course content and make proper classifications
It is difficult to track student enrollment across years, particularly to understand the reasons for student withdrawal	Work with sites to understand withdrawal codes and encourage them to standardize codes across schools and over time
It is difficult to develop representative longitudinal measures because different groups of students were missing over time to different degrees	Calculate the extent to which key subgroups of students are retained in longitudinal data sets to estimate their representativeness; when there is uneven missingness across groups, include cohort retention in reports on progress
Because internal systems are developed at different times and for different purposes, teachers and students might have multiple identification numbers, and sites might not have built crosswalks to be able to link them	Develop a single research ID and make it distinctive so it is not easily confused with existing IDs (e.g., make it one digit longer or add a distinctive character); work with the sites to establish a consistent system for converting internal IDs to research IDs before transferring data to the research team

Table 10—Continued

Challenges	Recommended Ways to Address Them
Site data on teaching experience are not collected consistently and might reflect total years of teaching, years of teaching at the site, or years of teaching in the current position; partial years might be rounded to whole years in different ways	Understand the types of information sites collect about teaching experience and request the information needed; it might be necessary to adjust analyses and reports to correspond to available data on years of experience
Metrics that link students and teachers and require multiple years of data, such as value-added measures, can be difficult to compute because students can enroll in multiple courses of the same type, multiple teachers can be responsible for a given course, students can transfer in and out of the course at different times, and some students do not persist in the district for all years	Understand exactly what types of data are available at the outset and try to develop metrics and analytic methods that reflect the available options, but recognize that there are no simple ways to address these challenges
Data on teachers' tenure statuses are not collected and reported consistently across sites; some states do not offer tenure, while others have changed tenure rules recently	Understand the rules about tenure and how records are kept before adopting metrics based on tenure
Most sites did not collect or retain data on teachers' participation in PD, mentoring, or coaching that would allow researchers to use the data for monitoring; although formal PD courses could easily be monitored, other forms of support, such as coaching, are more difficult to track	Collect these data more systematically to enable analysis focused on PD; find out what systems exist for keeping track of PD received by teachers and whether additional mechanisms can be put in place at reasonable cost before deciding to include PD as a study variable
Challenges related to data safeguarding and privacy	
Confidentiality requirements (such as FERPA) necessitate DUAs and data safeguarding procedures that will add time and complexity to the project	Make personal visits to each site to establish working relationships with data providers, develop mutually agreeable procedures, and build sites' confidence in researchers' ability to handle sensitive data
Site personnel likely will change over a long study, and replacement staff might be unfamiliar with established procedures or might want to change existing processes	Organize the data team with a single point of contact for each site, and, if possible, visit the sites every few years to keep relationships current
Site leaders do not want to be surprised by the public release of findings they have not seen	Preview results with site stakeholders so that they can be prepared to answer questions that might come from news media or the community
School districts likely will be cautious about releasing student and/or teacher data for secondary analyses; they have legal responsibility for protecting the confidentiality of their students and employees	Begin discussing the potential for restricted-use or public-use files early in a project

Is an expert teacher in one school district equivalent to a highly effective teacher in another? We encountered these kinds of scaling and categorizing questions repeatedly in different contexts. We used several strategies to try to standardize measures, including equating, converting to relative scales (e.g., percentiles), finding research-based crosswalks, and asking site experts for systematic judgments about equivalencies. Different kinds of solutions worked in different contexts, and a successful data-based analytic project will benefit from creative thinking about ways to rescale or reclassify based on existing measures.

Be Prepared to Collaborate with District Staff in Deidentification and Substitution of New Student and Teacher IDs

Most educators understand their responsibility to protect PII, particularly those who work with sensitive student and teacher information. However, their methods for ensuring confidentiality usually focus on preventing unauthorized personnel from gaining access to the data. Researchers who work with secondary data tend to address confidentiality by avoiding obtaining sensitive information in the first place by scrambling or substituting ID numbers. This process might need to be explained to people who are sending data, and algorithms will need to be developed that can be replicated over time so the de-identified data can be linked across years and across files.

As a final note, although we encountered several challenges, we were able to consistently produce dashboards for the sites and the foundation in a timely fashion and provide reliable data to research teams. The flexibility, thoughtfulness, patience, and creativity on the part of the sites, the foundation, and the research teams in handling these issues ultimately led to a sustained, successful effort. We think that the issues, suggestions, and solutions presented in this report extend the literature on developing metrics for dashboards by providing more-detailed and specific examples, and we hope that this report will inform future efforts related to collecting, processing, and analyzing these types of data.

Appendix. Metrics

In this appendix, we describe the methods used to generate each of the metrics included in the dashboard. These metrics were developed in conjunction with the foundation and the sites, taking into consideration the limitations of available data. Other studies might face different data limitations that would require alternative methodology. Note that, in many cases, alternative definitions were tried that did not work out for a variety of reasons. This report provides the definitions that were used in the dashboards; its purpose is not to justify the specific computations or outline the objections that were raised and discussed that eventually led to these definitions. Some of the computations are unique to these sites and are not likely to have general application. The metrics were selected and designed specifically to assist the foundation in monitoring the progress of the initiative.

Effectiveness

Differentiation of Performance

Percentage of teachers in each effectiveness (rating) category

This metric is the percentage of teachers in each of the effectiveness categories based on the TE ratings provided by the sites. In Table 11, we show the categories of TE as of 2014 as provided by the sites. Several of the metrics use categories labeled highly effective and ineffective, which combine multiple site categories as follows: highly effective combines categories in *italics* in the table and ineffective combines categories in **bold** in the table.

Method:

1. Obtain a count of the number of teachers in each effectiveness category.
2. Divide the count for each category by the total number of teachers with an effectiveness rating:

$$\text{Percentage} = 100 \times (\text{number in category}) / (\text{total number rated})$$

Strength of Tenure Decisions

Percentage of newly tenured teachers who were rated effective or higher in the previous year

This metric is intended to show whether sites are granting tenure to the right teachers. To calculate this metric, for SY 2013–2014, for example, the targeted teachers include those who received tenure beginning in July 2013 and ending in June 2014. The RAND team also calculates but does not include in the dashboard the count and percentage of newly tenured teachers who received ineffective ratings immediately before receiving tenure.

Method:

1. Obtain a count of the number of newly tenured teachers within each effectiveness category. Tenure status in one given school year is compared with the effectiveness category in the prior school year. For example, tenure dates from July 2013 to June 2014 are compared with SY 2012–2013 effectiveness scores, which reflect performance during the 2012–2013 school year.
2. Obtain the count of all newly tenured teachers in the effective or higher categories and ineffective categories (see Table 12).
3. Calculate the percentage of total:

$$\text{Percentage of newly tenured teachers rated effective or higher in the previous year} = \frac{\left(\begin{array}{c}\text{number of newly tenured teachers with}\\ \text{effective or higher rating}\end{array}\right)}{\left(\text{number of total newly tenured teachers}\right)}$$

$$\text{Percentage of newly tenured teachers rated ineffective in the previous year} = \frac{\left(\begin{array}{c}\text{number of newly tenured teachers}\\ \text{with ineffective rating}\end{array}\right)}{\left(\text{number of total newly tenured teachers}\right)}$$

The sites characterize TE differently. In Table 12, we indicate how TE categories from Table 11 are sorted into effective or higher and ineffective categories.

Performance of New Teachers

Effectiveness percentile rank of new teachers

This metric is intended to reflect improvement in recruiting based on the average percentile rank of the effectiveness scores of new teachers (three or less years of experience).

Method:

1. Calculate percentile points of the continuous TE score for all teachers that appear in the course teacher file (i.e., teachers who actually teach a course).
2. Select teachers with three years of experience or less. In sites that use lagged TE data, this metric will capture only teachers with two or three years of experience. Teachers with one year of experience will not yet have an available effectiveness score. Years of experience is determined from the jobs file. Note that sites capture and calculate years of experience differently. Some do not capture experience gained prior to joining the district. The RAND team attempted to use the most

TABLE 11

Dashboard Teacher Effectiveness Categories, 2014

Site/TE Categories	Alliance	Aspire	Green Dot	PUC	PPS	HCPS	SCS	DPS
Highest	*Master*	*Master*	*Highly effective II*	*Exemplary*	*Distinguished*	*Highly effective (5)*	*Significantly above expectations*	*Distinguished*
	Highly effective	*Highly effective*	*Highly effective*	*Highly effective*	Proficient	*Highly effective (4)*	Above expectations	Effective
	Effective	Effective	Effective	Progressing	Needs improvement	Effective	Meeting expectations	Approaching
	Achieving	**Emerging**	**Emerging**			**Needs improvement**	**Below expectations**	
Lowest	**Entering**	**Entering**	**Entry**	**Emerging**	**Failing**	**Ineffective**	**Significantly below expectations**	**Not meeting**

NOTE: PUC, PPS, and DPS each have only four TE categories.

TABLE 12

Effective or Higher and Ineffective Categories, by Site

Site	Effective or Higher	Ineffective
HCPS	Highly effective (5) Highly effective (4) Effective	Needs improvement Ineffective
SCS	Significantly above expectations Above expectations Meeting expectations	Below expectations Significantly below expectations
PPS	Distinguished Proficient	Needs improvement Failing
Alliance[a]	2012 Highly effective Effective Achieving	2012 Emerging Entry
	2013–2016 Master Highly effective Effective	2013–2016 Achieving Entering
Aspire[a]	2012 Master Highly effective Effective	2012 Emerging
	2013–2017 Master Highly effective Effective	2013–2017 Emerging Entering
Green Dot	Highly effective 2 Highly effective Effective	Emerging Entry
PUC	Exemplary Highly effective Progressing	Emerging
DPS	Distinguished Effective	Approaching Not meeting

[a] Alliance and Aspire changed the categorical names of their effectiveness categories starting with the 2012–2013 school year. As of 2017, Alliance is no longer participating in the IP project and, thus, no further updates from Alliance are expected. Aspire did not submit TE scores for 2018.

consistent and comprehensive measure of years of experience available in each site.

3. Calculate the mean percentile for this subset.

Selective Retention

Percentage of teachers in the highest- and lowest-rated categories and the top and bottom deciles retained for the following year

This metric is designed to show whether the strongest and weakest teachers remain in the classroom after receiving very high or very low ratings. Teacher retention is determined by a teacher being present in the course teacher file in consecutive years. (We cross-check this method by comparing with teachers who have a TE rating in the two relevant years.) Deciles were calculated using SAS PROC UNIVARIATE.

The RAND team also calculated the metric using the top and bottom effectiveness categories in place of top and bottom deciles. The top grouping corresponds to the highly effective grouped categories in Table 11. The bottom grouping corresponds to the ineffective grouped categories in Table 12. Because some teachers receive categorical ratings without continuous scores, the total number of teachers used in the calculation might vary by method.

Method:

1. Identify all teachers in year 1 who are in either the top or bottom grouping.
   a. For the decile method, the top grouping includes teachers with scores in the top 10 percent of TE scores; the bottom grouping includes teachers with scores in the bottom 10 percent.
   b. For the categorical method, the top grouping includes teachers with effectiveness ratings in the highly effective categories according to Table 11; the bottom grouping includes teachers with ratings from the ineffective categories in Table 12.
2. Limit the teachers to those who appear in the top or bottom grouping.
3. Limit the teachers to those who appear in the course teacher file in year 1. Assign all teachers in year 1 a retention variable that is 100 if the teacher is in the course teacher file in year 2, and zero if the teacher is not in the course teacher file.
4. For both the top and bottom groupings, calculate the mean of the retention variable:

$$\text{Percentage} = 100\times \left(\text{present in year 2} \big/ \text{present in year 1} \right)$$

Overall Teacher Improvement

Percentage of teachers who improved minus the percentage of teachers who declined, continuous score

This metric uses the TE rating provided by the sites and is the percentage of teachers who improved minus the percentage who regressed. The metric is calculated in two ways: by using effectiveness categories and by using effectiveness scores. Specifically, the metric calculates the percentage of teachers who (1) are present in two consecutive years, (2) do not have the highest TE rating in year 1, and (3) move from a lower rating category to a higher category in year 2, minus the percentage of teachers who (1) are present in two consecutive years, (2) do not have the lowest TE rating in year 1, and (3) move from a higher rating category to a lower category in year 2. For sites without access to the latest year of TE data, this metric reports the changes using the most-recent data available. For example, if 2014 TE data are unavailable, changes from 2012 to 2013 are reported in the 2014 column. By contrast, all other sites' 2014 calculations for this metric would reflect a change from 2013 to 2014.

Method:

Teacher improvement is calculated in two ways—using changes in effectiveness scores and changes in effectiveness categories.

Using effectiveness scores

1. Select all teachers who have a TE score that is below the top quartile of scores in year 1 and who also have a rating in year 2. Compute the standard deviation of scores in year 1.
2. Compare the effectiveness score from year 1 to year 2 and assign an improvement variable value of 100 if the effectiveness score is 0.5 standard deviations higher in year 2 than in year 1; otherwise, assign a zero.
3. Calculate the mean of the improvement variable:

$$\text{Percentage improved} = 100\times \frac{\left(\text{number of teachers improved in year 2}\right)}{\left(\begin{array}{c}\text{number of teachers below top rank in year 1}\\ \text{with a rating in year 2}\end{array}\right)}$$

1. Select all teachers who have a TE score that is above the bottom quartile of scores in year 1 and who also have a rating in year 2. Compute the standard deviation of scores in year 1.
2. Compare the effectiveness score from year 1 to year 2 and assign a decline variable value of 100 if the effectiveness score is 0.5 standard deviations lower in year 2 than in year 1; otherwise, assign a zero.
3. Calculate the mean of the decline variable:

$$\text{Percentage declined} = 100 \times \frac{(\text{number declined in year 2})}{\left(\begin{array}{c}\text{number above bottom rank in year 1}\\ \text{with a rating in year 2}\end{array}\right)}$$

1. Calculate the difference between the improvement percentage and the decline percentage:

$$\begin{array}{c}\text{Overall teacher improvement} =\\ \text{percentage improved} - \text{percentage declined}\end{array}$$

Using effectiveness categories

1. Select all teachers who have a TE category that is below the top category in year 1 and who also have a rating in year 2.
2. Compare the category in year 1 with that in year 2 and assign an improvement variable value of 100 if the year 2 category is higher than in year 1; otherwise, assign a zero.
3. Calculate the mean of the improvement variable:

$$\text{Percentage improved} = 100 \times \frac{(\text{number improved in year 2})}{\left(\begin{array}{c}\text{number below top rank in year 1}\\ \text{with a rating in year 2}\end{array}\right)}$$

1. Select all teachers who have a TE rating that is above the bottom category in year 1 and who also have a rating in year 2.
2. Compare the category in year 1 with that in year 2 and assign a decline variable value of 100 if the year 2 category is lower than in year 1; otherwise, assign a zero.
3. Calculate the mean of the decline variable:

$$\text{Percentage declined} = 100 \times \frac{(\text{number declined in year 2})}{\left(\begin{array}{c}\text{number above bottom rank in year 1}\\ \text{with a rating in year 2}\end{array}\right)}$$

1. Calculate the difference between the improvement percentage and the decline percentage:

$$\begin{array}{c}\text{Overall teacher improvement} =\\ \text{percentage improved} - \text{percentage declined}\end{array}$$

Addressing Ineffective Teaching

Percentage of ineffective teachers who exited or improved

This metric is designed to show whether sites are improving ineffective teachers or exiting them from the classroom. Teachers are classified as ineffective based on the TE categories they were in at the end of the previous school year (see Table 11). An *improved teacher* is defined as a teacher who receives a TE rating in a category that is at least effective in the following year. An *exited teacher* is defined as a teacher who does not appear in the course teacher file in year 2. The denominator is the number of ineffective teachers in year 1.

Equity

Access to Effective Teachers

Percentage of LIM[E] and non-LIM[F] students taught by highly effective and ineffective teachers

This metric is designed to show whether teachers are distributed equitably among students according to the students' LIM status. The percentage of LIM students who are taught by highly effective and ineffective teachers is compared with the percentage of non-LIM students who are taught by highly effective and ineffective teachers in both reading and math courses. Reading and math courses are identified using the course category column in the course file provided by each site. The calculation is based on student-teacher links in one year and TE results from the previous year. For example, the 2014 calculation reflects a student-teacher link for the 2013–2014 school year and the teacher's effectiveness rating for 2013.

[E] LIM students are those who are both eligible for FRL and identified as African American, Hispanic, or Native American.

[F] Non-LIM students are those who are not both eligible for FRL and identified as African American, Hispanic, or Native American.

The calculation for access to effective teachers was changed for the fall 2014 dashboard. Deciles are now computed within each subject. Because this change is applied to all years, dashboards created from fall 2014 onward might report different numbers for this metric for school years prior to 2014 than previous editions reported. Prior to the 2014 edition of the dashboards, deciles were computed using all teachers with effectiveness ratings, regardless of the subject the teachers taught.

The metric is calculated using both effectiveness categories and effectiveness scores. Teachers are identified as highly effective or ineffective per the categories shown in Table 11. Under the effectiveness scores method, teachers in the top and bottom deciles of the distribution within subject are identified. That is, if there are 200 teachers in a site who taught math courses during the school year, 20 of these teachers will be in the top decile and 20 will be in the bottom decile, even if there are 500 teachers in the entire site.

On average, one would expect about 10 percent of students to have access to either a top- or bottom-decile teacher using the effectiveness score method. Under the categorical method, one would expect the percentage of students with access to highly effective teachers to mirror the percentage of teachers who were rated as highly effective in the previous year.

Method:

For math and reading separately,

1. Identify math and reading courses.
2. Identify each teacher separately who taught at least one course during the school year within the subject.
3. Identify teachers in the top- and bottom-decile groups.
   a. Under the effectiveness score method,
      i. calculate TE deciles for all teachers who taught a course in that subject. Deciles will be determined from SAS PROC UNIVARIATE using the TE variables. Keep records only for teachers in the course teacher file who taught a course in the subject.
      ii. identify each teacher in the top 10 percent as a top-decile teacher. Identify each teacher in the bottom 10 percent as a bottom-decile teacher.
   b. Under the effectiveness category method,
      i. keep records only for teachers in the course teacher file who taught a course in the subject.
      ii. identify each teacher who received a highly effective rating according to Table 11 as a top-category teacher. Identify each teacher who received an ineffective rating according to Table 11 as a bottom-category teacher.
4. Join course student and course teacher files matching on school year, school ID, course ID, and section ID.
5. Add variables to each teacher indicating whether they were in the bottom or top group in TE scores from the previous year.
6. Summarize to student level, keeping an indicator if a student was taught by a top- or bottom-group teacher. It is possible for one student to be taught by both a top- and bottom-group teacher. Students are excluded if they are not enrolled in a school where the student-teacher link exists for at least four consecutive months. This is to ensure that the student-teacher links reflect substantial exposure to the teacher.
7. Join student demographics to the student-level file from step 6 to determine each student's LIM status.
8. Calculate the percentages of LIM students and non-LIM students who had top-group teachers and bottom-group teachers:

$$\text{Percentage LIM} = 100 \times \frac{\left(\begin{array}{c}\text{number of LIM students with}\\ \text{bottom- or top-group teacher}\end{array}\right)}{\left(\text{total number of LIM students}\right)}$$

$$\text{Percentage non-LIM} = 100 \times \frac{\left(\begin{array}{c}\text{number of non-LIM students with}\\ \text{bottom- or top-group teacher}\end{array}\right)}{\left(\text{total number of non-LIM students}\right)}$$

Strategic Staffing of Math Teachers

Difference in average prior performance of students assigned to novice teachers compared with experienced teachers

This metric is the average math achievement score of students in the year prior to their placement with the teacher of interest; for example, the grade 5 achievement scores for students currently in grade 6.

This metric is a measure of the disparity in prior math achievement between students placed with novice teachers and students placed with experienced teachers. For this metric, novice teachers are identified as those with one year of teaching experience or less; experienced teachers are those with three years of teaching experience or more. The metric has three parts, and the calculation is based on the methodology described in Strategic Data Project, 2012.

Three metrics are reported: the average score on assessment tests of students taught by experienced teachers reported in standard deviation units (z-score); the average z-score of students taught by novice teachers; and the gap in proficiency between students taught by experienced and novice teachers. When positive, the gap in proficiency indicates that more-proficient students tend to be placed with more-experienced teachers.

Method:

Using average z-scores

1. *Novice teachers* are teachers with one year or less of teaching experience.
2. *Experienced teachers* are teachers with three years or more of teaching experience.
3. Link each student to their respective math teacher and retrieve the student's state assessment scores from the previous year.
4. Identify each student's course section record as being taught by either an experienced or novice teacher.
5. Convert each student's assessment score to a z-score based on the districtwide mean and standard deviation of the state assessment tests by subject, year, and grade.
6. Find the average z-score for students of experienced teachers and the average z-score for students of novice teachers.

Using proficiency rates

1. *Novice teachers* are teachers with one year or less of teaching experience.
2. *Experienced teachers* are teachers with three years or more of teaching experience.
3. Link each student to their respective math state assessment scores from the previous year and their math proficiency level.
4. Identify each student's course section record as being taught by either an experienced or novice teacher.
5. Calculate the percentage of students who are proficient or above proficiency among all students taught by experienced teachers and the percentage of students who are proficient or above among all students taught by novice teachers.
6. Subtract the proficiency rate of students taught by novice teachers from the proficiency rate of students taught by experienced teachers:

$$\text{Proficiency gap percentage} = 100 \times \left(\begin{array}{c} \text{prior-year proficiency rate for students} \\ \text{of experienced teachers} - \\ \text{prior-year proficiency rate for students} \\ \text{of novice teachers} \end{array} \right)$$

Discipline Data

Percentage of LIM and non-LIM students expelled or suspended

This metric is the percentage of students who received disciplinary action (either a suspension of at least one day or an expulsion) at any time during the school year by LIM status. Suspension lengths are provided in the discipline file; expulsions, where applicable, are provided in the enrollment file.

Method:

1. Summarize the enrollment file to a student level, keeping an indicator for student expulsions.
2. Assign a 100 for students that were expelled.
3. Specify the LIM status of each student.

4. Summarize the discipline file to the student level, keeping an indicator for student suspensions.
5. Assign a 100 for students who received a disciplinary action that resulted in DAYS_SUSPENDED greater than zero.
6. Join the student-level enrollment and discipline files. Assign a 100 if a student received either a suspension or an expulsion; assign a zero if a student received neither.
7. Calculate the LIM/non-LIM students who received either suspensions or expulsions:

$$\text{Percentage LIM} = 100 \times \frac{\left(\text{number of LIM students with suspensions or expulsions}\right)}{\left(\text{number of LIM students}\right)}$$

$$\text{Percentage non-LIM} = 100 \times \frac{\left(\text{number of non-LIM students with suspensions or expulsions}\right)}{\left(\text{number of non-LIM students}\right)}$$

Excellence

Student Achievement

Percentage proficient or above on state standardized exam in reading and mathematics for all students and by LIM status

The student achievement metric reflects progress on state standardized testing in math and reading. This metric is based on the site-reported student proficiency indicator for the grades and subjects that were assessed. Percentages are calculated separately for math and reading and for LIM and non-LIM students.

Method:

1. Numerator: Count of students who are reported by the site to be proficient.
2. Denominator: All students enrolled during the school year in tested grades.

$$\text{Percentage} = \frac{\left(\text{number proficient}\right)}{\left(\text{total number of students}\right)}$$

Growth Relative to State

Percentage of students whose scores on state achievement tests increase from grade to grade more than the overall increase in state scores for white students in the same grades

This metric measures year-over-year progress of the sites' LIM students and non-LIM students compared with average year-over-year changes of white students[G] at the state level. Compared with student achievement, this metric focuses more on students' movement within the distribution. For example, it is possible for a student to lack proficiency in two consecutive years while having a higher score in the second year that suggests greater growth than the white student state average.

When states change the format of the state assessment tests between school years, this metric will not be reported on the dashboard.

Method:

1. From the IP data reported by the sites, calculate the difference for each student in their scaled scores between grades for two consecutive school years. For example, calculate the difference between grades 3 and 4 (grade 4 score minus grade 3 score).
2. Compare each student's yearly change with the change of white students in the corresponding grade and school year at the state level. For example, for a student in grade 3 in SY 2011 and grade 4 in 2012, use the difference in the white student statewide mean score for grade 3 in 2011 and the white student statewide mean score for grade 4 in 2012.
3. For students whose yearly change exceeds the change in white student state scores, assign an improvement variable a value of 100; otherwise, assign the student a value of zero.
4. Calculate the mean of the improvement variable overall grades for each school year.

[G] This average includes white students of all income levels.

Closure of the Achievement Gap

Closure of the achievement gap for LIM students for reading and math

The metric compares the achievement of LIM students for each site with state-level white and Asian students' achievement scores. The *LIM achievement gap* is defined as the difference between the percentage of LIM students in the district who are proficient and the percentage of white and Asian students in the state who are proficient.

Method:

Compute separately for reading and math.

1. Identify each student's LIM status using the enrollment file.
2. Join the LIM status to the state assessment scores.
3. Assign students a proficiency variable of 100 if the student is proficient and zero if the student is not proficient.
4. Take the mean of the proficiency variable for LIM students.
5. Subtract the LIM proficiency mean from the state proficiency weighted mean for white and Asian students.

Four-Year Graduation Rate

Percentage of students who graduate within four years of entering grade 9 for all students and by LIM status

This metric tracks the graduation of a cohort of students entering grade 9. Students who transfer out of the district for reasons out of the sites' control are not to be included in the denominator. Students expelled from school will be included in the denominator. Students who transfer into the district after grade 9 are not included in the denominator.

The RAND team provided each site with a list of the withdrawal codes from the enrollment file. The RAND team used the responses from the sites as the basis for defining graduates, and suspensions and expulsions.

Method:

1. Denominator: Identify first-time ninth graders for a given year. Subtract those students who eventually transferred out of the district in the subsequent four years.
2. Numerator: Identify the students who graduated by the end of the fourth year. For example, students entering grade 9 in the fall of the 2009–2010 school year must graduate by the summer of 2013.
3. Identify each student's LIM status using the enrollment file.

$$\text{Percentage} = 100 \times \frac{(\text{LIM graduates})}{(\text{grade 9 LIM cohort})}$$

$$\text{Percentage non-LIM} = 100 \times \frac{(\text{non-LIM graduates})}{(\text{grade 9 non-LIM cohort})}$$

College Readiness Rates

Percentage of students at college level on standardized tests for all students and by LIM status

This metric measures the percentage of students who are considered college ready using standardized tests. Two methods are used to calculate this percentage. The first method uses the highest ACT or SAT composite score a student achieved. The second method, for California students, uses only the EAP ELA and mathematics scores.

For each method, two college readiness percentages are calculated: one for substantial readiness and one for basic readiness (see Table 13). A student is considered college ready if they receive an ACT or SAT composite score or EAP scores above the thresholds shown in the table.

The ACT combined score consists of the sum of the science, math, English, and reading tests. The SAT composite is the sum of the SAT math and verbal tests. The EAP tests are augmented California Standards Tests (CSTs) in eleventh-grade ELA and math; these tests are part of California's public school testing and accountability system and are required of all students.

Students who do not take any exams are counted as not college ready. Students who do not complete all

TABLE 13

Basic and Substantial Readiness Scores for the ACT, SAT, and EAP

	ACT Combined	SAT Composite	EAP per Subject
Substantial readiness	84	1,000	3
Basic readiness	78	930	2

NOTES: As of 2014, there were three scores for the EAP: Level 3 = ready for California State University (CSU) or participating California Community College (CCC) college-level mathematics/English courses; Level 2 = ready for CSU or participating CCC college-level mathematics/English courses – conditional; and Level 1 = not yet demonstrating readiness for CSU or participating CCC college-level mathematics/English courses. From 2015 to 2017, there were four scores for the EAP: Level 1 = standard exceeded: ready for mathematics/English college-level courses; Level 2 = standard met: conditionally ready for mathematics/English college-level courses; Level 3 = standard nearly met: not yet demonstrating readiness for mathematics/English college-level courses; Level 4 = standard not met: not demonstrating readiness for mathematics/English college-level courses.

the sections for a given exam (i.e., all four ACT sections, both math and reading SAT sections, or ELA and math EAP tests) are counted as not college ready for that given exam.

From 2005 to 2016, there were three sections (critical reading, math, and writing), which were each scored from 200 to 800 points. Total SAT scores were based on a maximum of 2,400. However, only the math and critical reading scores are used for the priority metrics. In March 2016, the scoring system reverted to a total score of 1,600 (as it had been prior to 2005). Tests after that date have different assessment codes and must be adjusted to be compared with scores under the previous system. Using concordance tables provided by College Board, the RAND team converted the SAT scores for tests taken after March 2016 with the original scale for the sake of longitudinal comparison.

Method:

1. Denominator: Identify a ninth-grade cohort for each year. Subtract out those students who eventually transferred out of the district.
2. For each exam—SAT, ACT, or EAP—identify students who have basic readiness and students who have substantial readiness. College readiness is calculated separately for each eligible exam. A student must complete all sections in a given exam. If a student is college ready at one or the other level on at least one exam, that student is considered college ready at that level. Students must meet the following criteria for each exam:
   a. ACT: math, science, English, reading
      i. Combined English, reading, math, and science is above four for substantial readiness and 78 for basic readiness
   b. SAT: reading, math
      i. Composite reading and math score is above 1,000 for substantial readiness and 930 for basic readiness
   c. EAP: math and ELA
      i. Through 2014, both ELA and math scores must be greater than or equal to three for substantial readiness and two for basic readiness
      ii. As of 2015, California switched from the CST to the eleventh-grade California Assessment of Student Performance and Progress test. Under the new testing, both ELA and math scores must be Standard exceeded (1) for substantial readiness and Standard exceeded (1) or Standard met (2) for basic readiness.
3. Basic or substantial college readiness for the ACT and SAT requires a composite score at or above the cut point (ACT and SAT component scores do not have to be on the same test date); for EAP tests, basic or substantial college readiness requires a score at or above the cut point on all components of the test.

$$\text{Percentage} = 100 \times \frac{\left(\text{number of college ready}\right)}{\left(\begin{array}{c}\text{ninth-grade first-time enrollees}\\ \text{remaining in district}\end{array}\right)}$$

Participation in College Readiness Assessments

Percentage of each cohort participating in standardized tests to measure college readiness

This metric measures the degree to which students participate in college readiness testing. The metric considers the SAT, ACT, and EAP and uses any recorded score as evidence of participation. The tests must be taken within four years of entering grade 9.

Method:

1. Denominator: Identify grade 9 cohort for each year. Subtract the students who eventually transferred out of the district.
2. Numerator: Number of students who had a reported score for all sections of the SAT, ACT, or EAP.

$$\text{Percentage} = 100 \times \frac{\left(\begin{array}{c}\text{number of students with an SAT,}\\ \text{ACT, or EAP score}\end{array}\right)}{\left(\begin{array}{c}\text{ninth-grade first-time enrollees remaining}\\ \text{in the district}\end{array}\right)}$$

On-Time College Enrollment

Percentage enrolling in college within five years of entering grade 9 for all students and by LIM status

This metric reports on the percentage of students who enroll in college after graduation. *Enrollment in college* is defined as a student having an enrollment record for any college in the National Student Clearinghouse file (which is provided by the site) with an enrollment date on or before September 30 of the fifth year after their first enrollment in grade 9. We construct the denominator for this metric starting with all first-time ninth graders, subtracting out students with a valid transfer out of the site at any time before graduation. The files provided by the National Student Clearinghouse are cumulative. The 2018 file, for example, has new data for students who graduated between July 2016 and June 2017 and more-complete data for students who graduated between July 2015 and June 2016. We found that, because of lags in reporting to the National Student Clearinghouse, the most recent year of on-time college enrollment rates often is underreported and subsequently revised upward as more-complete historical data become available.

Method:

1. Denominator: Identify grade 9 cohort for each year. Subtract students who eventually transferred out of the district.
2. Numerator: Identify students who were enrolled before September 30 of the fifth year after their first grade 9 enrollment. For example, a student entering grade 9 in August 2008 must have an initial college entry date no later than September 30, 2012, to be considered on time.

$$\text{Percentage} = 100 \times \frac{\left(\begin{array}{c}\text{number of students enrolled before the fifth year}\\ \text{after their first grade 9 enrollment}\end{array}\right)}{\left(\text{ninth-grade first-time enrollees remaining in district}\right)}$$

References

Achieve, *Closing the Expectations Gap: 2013 Annual Report on the Alignment of State K–12 Policies and Practice with the Demands of College and Careers*, Washington, D.C., November 2013. As of October 25, 2019:
https://www.achieve.org/files/2013ClosingtheExpectationsGapReport.pdf

Baird, Matthew D., John Engberg, Gerald P. Hunter, and Benjamin K. Master, *Improving Teaching Effectiveness: Access to Effective Teaching—The Intensive Partnerships for Effective Teaching Through 2013–2014*, Santa Monica, Calif.: RAND Corporation, RR-1295/4-BMGF, 2016. As of October 27, 2019:
https://www.rand.org/pubs/research_reports/RR1295z4.html

CCSS—*See* Common Core State Standards.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, Vol. 104, No. 9, September 2014, pp. 2593–2632.

Colorado Department of Education, *2014–15 State Policy Report: Dropout Prevention and Student Engagement*, Denver, Colo.: Division of Innovation, Choice and Engagement, Office of Dropout Prevention and Student Re-Engagement, February 2016. As of January 20, 2020:
http://www.cde.state.co.us/dropoutprevention/2015dropoutpreventionpolicyreport

Common Core State Standards Initiative, "Standards in Your State," webpage, undated. As of October 25, 2019:
http://www.corestandards.org/standards-in-your-state/

Donhost, Michael J., and Vincent A. Anfara, Jr., "Data-Driven Decision Making," *Middle School Journal*, Vol. 42, No. 2, 2010, pp. 56–63.

Dundar, Afet, and Doug Shapiro, *The National Student Clearinghouse as an Integral Part of the National Postsecondary Data Infrastructure*, Herndon, Va.: National Student Clearinghouse Research Center, Institute for Higher Education Policy Research, May 2016.

Dynarski, Susan M., Steven W. Hemelt, and Joshua M. Hyman, "The Missing Manual: Using National Student Clearinghouse Data to Track Postsecondary Outcomes," *Educational Evaluation and Policy Analysis*, Vol. 37, No. 1, May 2015, pp. 53S–79S.

Fraser, Jeffery, *Improving Educational and Well-Being Outcomes: School-DHS Data Sharing in Allegheny County*, Pittsburgh, Pa.: Allegheny County Department of Human Services, August 2015.

Gardner, Kathryn A., and Sarah Muehler, eds., "Law Wise," No. 2, October 2014. As of October 25, 2019:
https://cdn.ymaws.com/www.ksbar.org/resource/collection/F729F565-5794-47F9-BD72-10AEC1A908DA/LW1410.pdf

Herman, Malia, "Data Dashboards a High Priority in National Ed-Tech Plan," *Education Week*, Vol. 35, No. 17, January 11, 2016, p. S14. As of December 16, 2019:
https://www.edweek.org/ew/articles/2016/01/13/data-dashboards-a-high-priority-in-national.html

Hock, Heinrich, and Eric Isenberg, "Methods for Accounting for Co-Teaching in Value-Added Models," *Statistics and Public Policy*, Vol. 4, No. 1, 2017, pp. 1–11.

Humes, Karen R., Nicholas A. Jones, and Roberto R. Ramirez, *Overview of Race and Hispanic Origin: 2010*, Washington, D.C.: U.S. Department of Commerce, Economics and Statistics Administration, U.S. Census Bureau, March 2011.

Kane, Thomas J., and Douglas O. Staiger, *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation*, Cambridge, Mass.: National Bureau of Economic Research, Working Paper No. 14607, December 2008.

Marsh, Julie A., John F. Pane, and Laura S. Hamilton, *Making Sense of Data-Driven Decision Making in Education: Evidence from Recent RAND Research*, Santa Monica, Calif.: RAND Corporation, OP-170-EDU, 2006. As of October 30, 2019:
https://www.rand.org/pubs/occasional_papers/OP170.html

Matheus, Ricardo, Marijn Janssen, and Devender Maheshwari, "Data Science Empowering the Public: Data-Driven Dashboards for Transparent and Accountable Decision-Making in Smart Cities," *Government Information Quarterly*, forthcoming.

McEachin, Andrew, Jonathan Schweig, Rachel Perera, and Isaac M. Opper, *Validation Study of the TNTP Core Teaching Rubric*, Santa Monica, Calif.: RAND Corporation, RR-2623-NTP, 2018. As of October 30, 2019:
https://www.rand.org/pubs/research_reports/RR2623.html

Means, Barbara, Christine Padilla, and Larry Gallagher, *Use of Education Data at the Local Level: From Accountability to Instructional Improvement*, Washington, D.C.: U.S. Department of Education, Office of Planning, Evaluation, and Policy Development, 2010. As of November 25, 2019:
https://files.eric.ed.gov/fulltext/ED511656.pdf

Mihaly, Kata, Daniel F. McCaffrey, Douglas O. Staiger, and J. R. Lockwood, *A Composite Estimator of Effective Teaching*, Seattle, Wash.: Bill & Melinda Gates Foundation, January 8, 2013. As of November 25, 2019:
https://k12education.gatesfoundation.org/resource/a-composite-estimator-of-effective-teaching/

Office of Management and Budget, "Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity," *Federal Register* notice, Washington, D.C., October 30, 1997. As of October 25, 2019:
https://www.whitehouse.gov/wp-content/uploads/2017/11/Revisions-to-the-Standards-for-the-Classification-of-Federal-Data-on-Race-and-Ethnicity-October30-1997.pdf

National Governors Association Center for Best Practices and the Council of Chief State School Officers, *Reaching Higher: The Common Core State Standards Validation Committee*, Washington, D.C., June 2010.

Phillips, Meredith, Sarah Reber, and Jesse Rothstein, *Getting Down to Facts II: Making California Data More Useful for Educational Improvement*, Stanford, Calif.: Stanford University and Policy Analysis for California Education, September 2018.

Porter, Andrew C., and Morgan S. Polikoff, "Measuring Academic Readiness for College," *Educational Policy*, Vol. 26, No. 3, 2012, pp. 394–417.

Raftree, Linda, "13 Tips for Creating Data Dashboards for Decision-Making," *Technology Salon*, June 9, 2015. As of December 16, 2019:
http://technologysalon.org/13-tips-for-creating-data-dashboards-for-decision-making/

RAND Corporation, "About the Intensive Partnerships for Effective Teaching (IP) Initiative," webpage, undated. As of January 20, 2020:
https://www.rand.org/education-and-labor/projects/evaluating-teaching-effectiveness/about.html

Rothman, Robert, *Data Dashboards: Accounting for What Matters*, Washington, D.C.: Alliance for Excellent Education, January 2015.

Saget, Bedel, “Where the SAT and ACT Dominate,” *New York Times*, August 4, 2013. As of October 25, 2019:
https://archive.nytimes.com/www.nytimes.com/interactive/2013/08/04/education/edlife/where-the-sat-and-act-dominate.html

Smith, Veronica S., “Data Dashboard as Evaluation and Research Communication Tool,” *New Directions for Evaluation*, Vol. 2013, No. 140, 2013, pp. 21–45.

Stecher, Brian M., Deborah J. Holtzman, Michael S. Garet, Laura S. Hamilton, John Engberg, Elizabeth D. Steiner, Abby Robyn, Matthew D. Baird, Italo A. Gutierrez, Evan D. Peet, Iliana Brodziak de los Reyes, Kaitlin Fronberg, Gabriel Weinberger, Gerald P. Hunter, and Jay Chambers, *Improving Teaching Effectiveness: Final Report—The Intensive Partnerships for Effective Teaching Through 2015–2016*, Santa Monica, Calif.: RAND Corporation, RR-2242-BMGF, 2018. As of October 25, 2019:
https://www.rand.org/pubs/research_reports/RR2242.html

Stecher, Brian M., Deborah J. Holtzman, Michael S. Garet, Laura S. Hamilton, John Engberg, Elizabeth D. Steiner, Abby Robyn, Matthew D. Baird, Italo A. Gutierrez, Evan D. Peet, Iliana Brodziak de los Reyes, Kaitlin Fronberg, Gabriel Weinberger, Gerald P. Hunter, and Jay Chambers, *Intensive Partnerships for Effective Teaching Enhanced How Teachers Are Evaluated But Had Little Effect on Student Outcomes*, Santa Monica, Calif.: RAND Corporation, RB-10009-1-BMGF, 2019. As of October 25, 2019:
https://www.rand.org/pubs/research_briefs/RB10009-1.html

Strategic Data Project, “Do Low-Performing Students Get Placed with Novice Teachers?” Cambridge, Mass.: Harvard University Center for Education Policy Research, 2012. As of October 25, 2019:
http://sdp.cepr.harvard.edu/files/cepr-sdp/files/sdp-spi-placement-memo.pdf?m=1431356806

The World Bank, “Global Education Policy Dashboard: The Challenge: Understanding What Drives the Learning Crisis,” webpage, January 17, 2019. As of December 16, 2019:
https://www.worldbank.org/en/topic/education/brief/global-education-policy-dashboard

Thomsen, Jennifer, “50-State Comparison: Teacher Tenure/Continuing Contract Policies,” Denver, Colo.: Education Commission of the States, 2014. As of October 25, 2019:
www.ecs.org/teacher-tenure-continuing-contract-policies

Underwood, Julie, “The State of Teacher Tenure,” *Phi Delta Kappan*, Vol. 99, No. 7, 2018, pp. 76–77.

U.S. Department of Agriculture, Food and Nutrition Service, “National School Lunch Program: Income Eligibility Guidelines,” webpage, undated. As of October 25, 2019:
https://www.fns.usda.gov/school-meals/income-eligibility-guidelines

U.S. Department of Education, “United States Education Dashboard,” webpage, undated. As of December 16, 2019:
https://dashboard.ed.gov

Welsh, Richard O., “School Hopscotch: A Comprehensive Review of K-12 Student Mobility in the United States,” *Review of Educational Research*, Vol. 87, No. 3, June 2017, pp. 475–511.

West, Darrell M., *Big Data for Education: Data Mining, Data Analytics, and Web Dashboards*, Washington, D.C.: Brookings Institution, September 2012.

Zvoch, Keith, and Joseph J. Stevens, “Sample Exclusion and Student Attrition Effects in the Longitudinal Study of Middle School Mathematics Performance,” *Educational Assessment*, Vol. 10, No. 2, 2005, pp. 105–123.

Acknowledgments

We are grateful to the partnership sites for responding to our data requests and providing detailed and thoughtful information during the course of the project. Specifically, we thank Anna Brown, Ted Dwyer, Justin R. Rogers, and David Russell of Hillsborough County Public Schools (HCPS); Sam Franklin, Lorraine Marnet, Tara Tucci-Exilus, and Ashley Varrato of Pittsburgh Public Schools (PPS); Tangie Ray and Jeffery Shive of Shelby County Schools (SCS); Heather George, Kendra Wilhelm, Yu-lu Hsiung, and Wesley Brunko of Denver Public Schools (DPS); Vireak Chheng, David Medina, and Liam Mennis of Alliance College-Ready Public Schools; Veronica Chew, Cathleen Li, and David Roth of Aspire Public Schools; Kevin Keelen, Samantha Mita, Quang Nguyen, and Sam Regnier of Green Dot Public Schools; and Matt Goodlaw, George Lewis, Johana Olaya, and Jaime Serrano of Partnerships to Uplift Communities Schools.

We also appreciate the willingness of staff at the Bill & Melinda Gates Foundation to engage with us; we are especially grateful to our program officers, David Silver, Eli Pristoop, Andrew Sokatch, and Alex Resch for their advice and responsiveness throughout the project.

We also would like to thank our reviewers, Fatih Unlu, Tara Tucci-Exilus, and Kata Mihaly for their insightful comments and useful suggestions for improving the report.

At RAND, we are grateful to Chandra Garber and John Engberg for their assistance with earlier versions of this report. We thank Paul Steinberg for his recommendations on how to improve the readability and clarity of the report. We also acknowledge other members of the RAND data team, including Melissa Felician, Rodger Madison, Deborah Wesley, Judith Mele, Marian Oshiro, Chris Chan, Kim Hui, and Andrew Madler.

About the Authors

Gerald P. Hunter is a research programmer at RAND. His research interests span transportation, economic development, education, energy, and urban policy. His work in education has combined administrative data with publicly available demographic data to help develop quantitative measures of the effectiveness of school teachers in multiple school districts across the country. He has worked on multiple projects involving dashboard construction and automation, education-related metrics, and teacher value added for school districts and charter management organizations. Hunter received his master's degree in city planning.

Stephanie Williamson is a research programmer at RAND. Her research includes education and public health. She holds a B.A. in economics.

Asa Wilks is a research programmer at RAND. His projects have included predicting medication adherence, analyzing impacts of a home visiting program on health outcomes and new disease diagnoses, and producing dashboards of student outcomes in several large U.S. cities. He holds a master's degree in public administration.

Janet M. Hanley is the director of the RAND Research Programming Group and a senior statistical programmer. She has extensive experience working with large-scale administrative data files to create individual behavioral and utilization histories for health care and military manpower projects. She has managed the data collection and analytic efforts of several health care and education evaluation projects. She holds a master's degree in statistics.

Brian M. Stecher is an adjunct social scientist at the RAND Corporation. His research focuses on measuring education quality and evaluating education reforms, with a particular emphasis on assessment and accountability systems. He holds a Ph.D. in education.

About This Report

In 2009, the Bill & Melinda Gates Foundation launched the Intensive Partnerships for Effective Teaching (IP) initiative. The initiative's aim was to support school districts and charter management organizations in implementing reforms focused on (1) improving the recruitment and retention of effective teachers; (2) boosting the access of low-income minority students to more-effective teachers; and (3) increasing college readiness and attendance, particularly among underrepresented groups. The ultimate goal of these reforms was to improve student achievement, graduation rates, and postsecondary participation.

In this report, the RAND data team describes the process of creating the dashboards and discusses some of the hurdles encountered and lessons learned in designing indicators, collecting the data from individual IP sites, managing and harmonizing the data across sites, and monitoring and comparing trends over time. This report should be of interest to researchers, data scientists, and practitioners who want to understand the challenges of using administrative data to monitor large-scale evaluation efforts.

This research was conducted by RAND Education and Labor, a division of the RAND Corporation that conducts research on early childhood through postsecondary education programs, workforce development, and programs and policies affecting workers, entrepreneurship, and financial literacy and decisionmaking. This report is based on research funded by the Bill & Melinda Gates Foundation. The findings and conclusions we present are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation. For more information, please visit www.gatesfoundation.org.

More information about RAND can be found at www.rand.org. Questions about this report should be directed to ghunter@rand.org, and questions about RAND Education and Labor should be directed to educationandlabor@rand.org.

For more information on this publication, visit www.rand.org/t/RR2917.

Published (2020) by the RAND Corporation

www.rand.org