GETTING DOWN TO FACTS II

Technical Report

Making California Data More Useful for Educational Improvement

Meredith Phillips
University of California, Los Angeles

Sarah Reber
University of California, Los Angeles

Jesse Rothstein
University of California, Berkeley

September 2018

About: The *Getting Down to Facts* project seeks to create a common evidence base for understanding the current state of California school systems and lay the foundation for substantive conversations about what education policies should be sustained and what might be improved to ensure increased opportunity and success for all students in California in the decades ahead. *Getting Down to Facts II* follows approximately a decade after the first *Getting Down to Facts* effort in 2007. This technical report is one of 36 in the set of *Getting Down to Facts II* studies that cover four main areas related to state education policy: student success, governance, personnel, and funding.

Stanford University

PACE Policy Analysis for California Education

Making California Data More Useful for Educational Improvement

Meredith Phillips
University of California, Los Angeles

Sarah Reber
University of California, Los Angeles

Jesse Rothstein
University of California, Berkeley

Acknowledgments

We thank Elizabeth Dabney, Jon Fullerton, Dan Goldhaber, Chris Kingsley, Heather Hough, Helen Ladd, Jay Pfeiffer, Morgan Polikoff, Evan White, and participants at the Getting Down to Facts II meetings for helpful conversations and comments on earlier drafts. We are grateful to Elsa Augustine and Rachel Young for excellent research assistance and to the Getting Down to Facts II project funders and the Laura and John Arnold Foundation for financial support. The views expressed are those of the authors and do not necessarily reflect those of the funders.

Abstract

Modern computing technology makes it possible for governments at all levels to use the data they already collect to improve service coordination and delivery, and to conduct research and evaluation to inform policymaking. California is well behind other states in taking advantage of this opportunity, in education and in other fields. The state has a patchwork of data systems that are not well integrated and do not provide satisfactory answers to the state's most important policy questions. A particular need is for better linkages across providers – for example, between K-12 and higher education, or among the state's three public higher education systems.

Regional collaborations have developed to fill this need. While these efforts are important and valuable, they are also difficult to set up, and necessarily leave large gaps. Regional efforts are not an adequate substitute for statewide systems.

While there are political, technological, and organizational barriers to the creation of improved statewide data systems, they are not insurmountable. Other states have overcome these barriers and demonstrate the substantial value of better data systems. Enormous IT projects are not required – significant progress can be made at relatively low cost, given political will to overcome bureaucratic and organizational inertia.

A stronger statewide data infrastructure is an essential part of a modern education system today and will help California deliver a world-class education to its students.

Introduction

Improvements in computing technology have made it possible for public agencies to use their records as data that can be analyzed and used to support improved service delivery. In education, administrative data are used to support individualized learning strategies, continuous improvement efforts, accountability systems, and fundamental research that informs a wide range of educational policy, curriculum, and instructional choices. Education data linked to information from other public services, such as child welfare, foster care, social services, health care, and criminal justice, can improve coordination across these functions, supporting better educational and non-educational outcomes for children.

Taking advantage of these opportunities requires building administrative datasets. California has made considerable progress in developing individual-level, administrative datasets in education and other sectors. These have generated real value – to take one example, these data are used to implement the new Local Control Funding Formula (LCFF) funding allocations.

The state has made less progress, however, in linking data across systems to support continuous improvement efforts at the school or district level (see Hough, Byun, and Mulfinger, this volume), or for analyses that can lead to system improvement. California's existing data could be used to enhance the state's education system and the well-being of California residents if a secure infrastructure were developed for linking individual-level data within and across sectors and making such data accessible to school districts and higher education systems, state agencies and policy makers, and researchers inside and outside the government.

While stakeholders have legitimate concerns that connecting data across sectors and expanding data availability pose risks to privacy, or that the data will be used for undesirable political purposes, regional efforts in California—as well as the experiences of other states—show that such hurdles are surmountable, that investing in a statewide, cross-sector data infrastructure could reduce duplication of effort, and that the benefit of such an infrastructure to schools, school districts, service providers, policymakers, and Californians could be substantial. Indeed, investment in a modern, statewide data system would arguably *reduce* the privacy risks currently posed by the multiplicity of ad hoc linking efforts that have arisen in the absence of a coordinated system.

Improved statewide data systems, with appropriate privacy protections, would support many different purposes—from service provision to predictive analytics to evaluation, all of which would help agencies and policymakers better serve Californians. At the most basic level, more information sharing within and between sectors could make it easier for service providers to provide more timely supports to students and better coordinate their services. For example, Allegheny County, Pennsylvania, has connected academic and human-services-related data to help school staff understand students' mental health and child welfare involvement while helping social workers and caseworkers understand children's school performance, attendance, and disciplinary history (Fraser 2015).

Linked individual-level data would also make it possible to generate descriptive evidence that can support continuous improvement efforts (see Hough, Byun, and Mulfinger, this volume), such as targeting programs and interventions to the individuals who need them most. One of the best-known illustrations of the potential value of this type of use comes from Chicago. Beginning in the 1990s, researchers at the University of Chicago Consortium on Schools Research used individual-level, longitudinal data from the Chicago Public Schools to develop a 9th grade measure of course completion and course failure that was a strong predictor of whether students would fail to graduate from high school (Allensworth and Easton 2005). The school district incorporated this "on-track" for high school graduation indicator into its data systems and promoted initiatives that helped schools monitor students' drop out risk and reduce students' course failures (Pitcher, Duncan, Nagaoka, Moeller, Dikerson, and Beechum 2016). Chicago students' 9th grade "on track" and high school graduation rates improved as schools used the new indicator to identify students in need of extra support and developed ways to help students improve their grades (Allensworth 2013; Roderick, Kelly-Kemple, Johnson, and Beechum 2014).

A third potential use of improved data systems is to identify organizations that stand out as particularly effective. Agency staff or researchers can then investigate those organizations' practices and share what they learn more broadly. In Tennessee, the National Center on Scaling Up Effective Schools at Vanderbilt University has been working with two urban school districts to use longitudinal data on high school students to identify more and less effective high schools and learn about the practices in those high schools that may be contributing to their effectiveness (see, e.g., Cannata, Smith, and Haynes 2017). In North Carolina, researchers used linked longitudinal data from the K-12 system and the community college system to examine whether otherwise similar students were more likely to earn a degree or complete the required transfer coursework if they attended particular community colleges (Clotfelter, Ladd, Muschkin, and Vigdor 2013).

Finally, linked data are useful for evaluating the effectiveness of policies and programs. Researchers have used linked birth records and public school records in North Carolina to measure the impact of statewide early childhood policy initiatives, including funding to improve child care quality and expand preschool slots, on children's math and reading achievement by the end of elementary school (Dodge, Bai, Ladd, and Muschkin 2017). In Michigan, researchers have used statewide longitudinal data to evaluate whether a change in the state high school curriculum improved academic achievement (Jacob, Dynarski, Frank, and Schneider 2017). In Florida, researchers have used statewide longitudinal data on teachers and their students to evaluate the effectiveness of state policies designed to reduce teacher shortages in math, science, and special education (Feng and Sass 2015). Also in Florida, researchers used longitudinal data from a large urban district to show that a universal screening program for identifying gifted students increased the representation of low income and minority students in that district's gifted programs (Card and Giuliano 2016).

Although California has not been at the forefront in developing its statewide data systems, the state offers many examples of effective data use. The California Longitudinal Pupil

Achievement Data System (CALPADS) matches individual students throughout California with information about those same students from the California Department of Social Services. This allows the California Department of Education (CDE) to provide information to each school district about which of their students are foster youth and/or are categorically eligible for free meals under the National School Lunch program. Another example comes from higher education: The California Community College system now routinely merges the records of former students with wage records from the Employment Development Department (EDD) to provide publicly available estimates of earnings by discipline and college. These data have been used by researchers to examine the returns to career-technical certificates and degrees (see, e.g., Stevens, Kurlaender, and Grosz 2015).

These and similar efforts have generated real value. But they have been costly in time, effort, and expense to set up, because California's data systems are not designed to support them. Data systems are typically isolated, not linked to each other, and mutually inaccessible. Thus, while the community college system's agreement with EDD allows it to merge to earnings and employment data, enabling measurement of students' labor market outcomes, it cannot routinely measure relevant outcomes for students who transfer to California State University or University of California (UC) campuses to complete four-year degrees, because there is no standing link between community college, CSU, and UC data. For the same reason, elementary and secondary education districts typically cannot follow their students into college or the labor market, so are unable to use information on students' college or employment outcomes to guide K-12 improvement efforts (Moore, Grubb, and Esch 2016). Many other states have developed systems to support this kind of use, but California has not. The state's lack of a linked longitudinal data system inhibits its ability to design and run the world-class education system that its people need.

Fortunately, California could draw on the experiences of both other states and existing efforts in the state to develop more useful data systems. This paper makes the case for the value of a more comprehensive educational data system and provides guidance about how California might proceed to develop such a system for the benefit of educators, policymakers, and, most important, California's next generation.

What Could We Do with Better Data Systems and What Do We Need?

Potential Uses and Users

California's education system involves many different actors who use—or could use—data to help them achieve their goals. A comprehensive data system could support a range of educators, administrators, researchers, and policymakers in different ways. Some of the potential users and uses – including some that are already being pursued and others that are impossible given the current fragmented data infrastructure – include:

- Teachers and other school instructional and support staff could use data on individual students' past performance to customize instruction and interventions such as English language instruction or reading assistance.
- School leaders could use similar data to inform classroom groupings or staffing decisions. They could also use data on students' subsequent performance – for example, college going, college completion, and career success rates for a high schools' past students – to help them understand whether their students are graduating well-prepared for the next stages of their lives.
- Districts could use similar data to inform staffing, curricular, and resource-allocation decisions.
- Districts could also use data on social service program participation to coordinate efforts with local social service agencies, to ensure, for example, that youth in foster care receive appropriate educational and non-educational services.
- The California Department of Education could use college and career success measures, along with student achievement scores, to understand how students progress through the state's education system and which programs are more and less effective at preparing students for later success.
- CDE could use – and indeed, under the Local Control Funding Formula, does use – demographic and other student characteristics as an input into funding formulae. CDE could also use similar data, combined with finance, staffing, or other resource allocation data, to understand how state funds are allocated to and used by schools.
- Higher education institutions could use data collected from students' high school transcripts to streamline admissions procedures, course placements, advising, and even financial aid awards.
- California's colleges could also use this information to inform outreach campaigns aimed at identifying high-performing students who might not be contemplating college. Institutions that accept transfer students might use the information in community college records for similar purposes.
- Providers of non-educational public services, such as welfare and nutrition programs, criminal justice authorities, and health providers, could use data on children's educational enrollment and outcomes to coordinate service delivery and to measure outcomes for service recipients. For example, a juvenile justice diversion program might monitor the progress of the young people it serves by measuring their attendance and progress in school.
- Research units within districts, the CDE, and other government agencies could use outcome measures derived from education records to evaluate programs or interventions. For example, a district might assess the impact of a new curriculum, or a juvenile justice agency might compare the effectiveness of alternative diversion programs.
- Academic and non-academic researchers might use education data for basic research aimed at uncovering structural patterns in the education system or the economy. For example, researchers might examine the impact of family economic circumstances on student success in school, or they might develop "early warning indicators" of juvenile

delinquency. Such studies enhance our understanding of educational processes generally and inform the development of the next generation of interventions.

California's current data systems provide limited support for these potential users and uses. This leads to duplication in data collection, as agencies collect data elements for themselves because they do not have access to the data system in which the information already exists. It also leads to ineffective service delivery and holes in the web of education services that the state provides. A more integrated California education data system would facilitate all of the above potential uses, in some cases making things possible that are simply not possible today and in other cases substantially reducing costs, increasing efficiency, and allowing students to be better served.

Moreover, California policymakers are often forced to rely on evidence from elsewhere, often in settings with limited resemblance to California, simply because California's data systems make it difficult to assemble evidence here. For example, the chapter in this volume on teaching English Learners in California relies on evidence from studies conducted in Texas, Florida, and North Carolina (Santibañez, this volume). Similarly, we understand much less about educator mobility—about how teachers' and principals' careers unfold—in California than in other states where data systems better facilitate research on this important topic.

Types of Uses and System Requirements

Different users and uses of education data systems require different data. Unlocking the value of each potential use thus requires different characteristics of the data system, with different operational and technical requirements and different security risks.

One useful distinction concerns the "freshness" of the data. For some uses, it is sufficient for the data user to have access only to "frozen" data, perhaps several months old at the time of use. For example, in an evaluation of a pilot juvenile justice program, the evaluator might need access to attendance and completion records of students who participated in the program last year but these records do not need to be up-to-the-minute. This study could proceed based on periodic extracts from attendance and graduation records. Other uses, however, require real-time access to the district's IT system – for example, a probation officer who needs to check a student's attendance this week, not several months in the past. This user would get no value from access to frozen records that are months out of date.

A related distinction is between uses requiring access to the individual records of identifiable students and those that can use aggregated or de-identified data. A juvenile probation officer knows the name of the student with whom he or she is working, and needs that student's attendance records, not those of some other student; the same goes for the college placing students into courses based on their high school records. This kind of sharing involves personal information that is rightly considered private. Currently, this type of sharing happens in ad hoc ways—for example, students may be asked to obtain their own school records and turn them over to their probation officer or their college placement office. Such

arrangements create disclosure risks; carefully designed automated systems could facilitate necessary sharing while offering stronger privacy protections.

Most uses of administrative data do not require access to personal information, however. Many uses require only data that are aggregated to a level higher than the student – for example, school averages are often sufficient to support district funding and staffing decisions. Other uses require access to individual students' records but do not require access to personal information. For example, evaluators of a particular program may need individual-level records about students who participated in the program, but they can typically work with "de-identified" data, from which personal identifiers such as names, birthdates, addresses, and student ID numbers have been removed, to reduce risk of accidental release of personal information.

It is possible to design data systems to serve each of these potential use cases, while preserving student privacy and information security. But these different uses call for different designs. Rather than trying to build a single, fully integrated, PK-20 data system that serves all potential uses, it seems to us more feasible for California to proceed incrementally, steadily improving statewide data systems and working gradually toward a more comprehensive system – or set of systems – that would meet many of the state's needs. In this paper, we emphasize data systems that can provide "frozen" data from the recent past, either de-identified or with individual identifiers, for use in continuous improvement, analysis, evaluation, and research. In contrast to "live" data systems that maintain continuous connections to each of the constituent databases, "frozen" data systems do not require enormous IT projects or extensive redesign of the existing databases used by districts and other agencies in the state, so can be implemented relatively quickly and at lower cost.

Current State of Data Availability and Access in California

California's education system is famously uncoordinated, with a frequently shifting division of responsibility and authority among the state, the counties, and the school districts. This arrangement has been described as "a remarkably crazy quilt of interacting authorities that are not aligned" (as quoted in Brewer and Smith 2007, p. xv). Although Brewer and Smith (2007) noted a substantial shift of authority to the state level from 1964 onward, more recently the Local Control Funding Formula has reversed this trend to some extent. At the postsecondary level, the three distinct systems of public higher education have struggled throughout their history to coordinate effectively with each other or with the elementary and secondary education system. Since the 2011 closure of the California Postsecondary Education Commission, there is no formal structure within which such coordination can take place, and the systems have operated largely independently (Moore, Grubb, and Esch, 2016).

Existing State Education Data Systems

Without a central authority to facilitate joint work and support collaboration, each agency has developed its own data systems, largely independently. As a result, unlike other states that have databases that can be described as a single statewide education data system,

California has dozens or hundreds of separate systems that are, at best, loosely connected (Moore et al. 2017). Limiting our attention only to data on individual students, the state education data systems include:

- Data on elementary and secondary students and staff, generated by individual school districts and maintained in the districts' own data systems.
- The CALPADS system maintained by CDE, which rolls up some (but not all) of the student- and staff-level information in the individual district data systems.
- The three higher education systems (the community colleges, the California State University, and the University of California) each have registration, transcript and financial aid data on their own students, as well as admissions data for the selective campuses. Each campus has its own data system, but each of the three sectors also maintains its own centralized data system that, like CALPADS, contains a subset of data elements for students from all the campuses in that sector. No data system combines data across the three sectors in a systematic way.
- The California Student Aid Commission (CSAC) maintains data on financial aid eligibility and receipt for students who applied for financial aid.
- The National Student Clearinghouse (NSC) is a non-governmental organization that collects information on enrollment and graduation from institutions of higher education in California and nationally. The data can be purchased by districts, states, colleges, and researchers.

Beyond these, publicly provided programs, outside of the K-12 and higher education systems, obtain relevant information that is not routinely collected in standardized databases. For example, the state funds or subsidizes a range of programs in the early education sector, but does not maintain individual-level data on participation that would be useful for measuring students' success as they progress to elementary school. Finally, several databases not at the student level are useful for understanding individual students' educational experiences – for example, the publicly searchable information on teaching credentials and certificates maintained by the Commission on Teacher Credentialing.

This fragmented structure represents a substantial *improvement* from the situation a decade ago. In 2008, Loeb, Beteille, and Perez enumerated over a dozen K-12 datasets maintained by CDE, but not linked at the student level. The integration of many of these datasets into CALPADS is a big step forward.

CALPADS is underutilized, however. The CDE has developed relatively standardized procedures by which agencies and researchers can gain access to CALPADS data for specified research projects. Despite this, a shortage of resources at CDE and a lack of clarity among potential users about permissible uses have prevented the data from being used to nearly the extent that similar data systems in other states are used by agency staff and others to understand and improve educational practices in those states. Moreover, the CALPADS access procedures are the exception, not the rule: Most other California data systems have no regular processes or procedures by which analysts from other agencies, policymakers, or researchers

can obtain access. While there have been efforts to develop linkages among them (discussed below), these can take years of effort to establish, and often produce one-off linkages that can be used only for a single, specific purpose but not for others.

Other State Data Systems

Beyond the education data systems listed above, other important student-level data that reside elsewhere in state and county government would be useful for the effective operation of the education system:

- The Employment Development Department (EDD) and Franchise Tax Board (FTB) maintain data on the employment and earnings of California workers that could be used to measure economic outcomes of the education system. These data might be used, for example, to measure career readiness of high school graduates participating in Career Technical Education (CTE) programs.
- A range of state agencies, including the EDD and the Department of Social Services, administer programs that provide job training, apprenticeships, and other skill-building interventions to adults. Data on participation in these programs would be useful for assessing the effectiveness of the education system at preparing students for careers and for identifying new programmatic needs.
- The state's various social service systems, themselves fragmented across multiple state-level departments and between the state and its counties, contain information on current and former students' involvement with cash and non-cash welfare programs (such as CalWorks and CalFresh) and interactions with the child welfare and foster care systems. These data would be useful for service coordination – it would be helpful, for example, for school social workers to know whether a student's family has interacted with the child welfare system.
- Many students receive state-funded or subsidized health insurance through MediCal or Covered California (the state's health insurance exchange), and receive health care through public health providers in the state. Data linkages with these systems could make it possible for schools that operate health clinics to provide appropriate basic health services to their students and could facilitate the coordination of health care and special education services.
- Many students and their parents have interactions with the criminal justice system, itself fragmented with several distinct data systems at the state and county level. Coordination across systems would be useful for targeting interventions at students at risk of dropout, abuse, criminal justice involvement, or other adverse outcomes. Links to criminal justice data could also be used for evaluating programs that may impact criminal justice involvement.

The state of California or its agents already own each of the above data systems, and each already contains a great deal of information, in a readily usable form, that could usefully support improved administration of the education system, if only it were available to those who could benefit from it. But data access is limited and unsystematic.

Data Integration Efforts

In recent years, there have been a number of efforts to integrate data across the various systems (Moore and Bracco 2018). These have largely come from the bottom up, starting with individual districts or other education providers who saw the value that better data could add to their ability to provide needed services. They have also mostly been localized, limited to a single pair of data systems (e.g., a single district's records merged to the local community college district) or to a particular geographic area. This limits the opportunity to use these efforts to improve systems statewide. Nevertheless, these efforts have demonstrated the value that could be generated through more intentional, comprehensive efforts to integrate data across the state.

Several efforts focus on linkages between K-12 and higher education data:

- Cal-PASS Plus contains linked, longitudinal education data provided by California school districts, community colleges, and universities. Funded by the Community College Chancellor's office, and managed by the Educational Results Partnership and San Joaquin Delta College, Cal-PASS Plus makes it possible for participating agencies, and collaborating researchers, to match students from K-12 education to community college (and some four-year college) records (Cal-PASS Plus, no date). The initiative's flagship study examines the use of multiple measures – not just a single placement exam – to inform course placement for new community college students (Bahr et al. 2017). Other projects include creating dashboards to measure the effectiveness of Career Technical Education (CTE) programs or educational outcomes for foster youth.
- The CORE Districts, a consortium of eight large, primarily urban school districts, have developed a data collaborative that combines student-level data from these participating districts and others. The CORE districts use these data to generate district-designed school accountability measures and to conduct research in partnership with Policy Analysis for California Education (PACE). Data elements include a number of locally-generated measures not available through the CDE's CALPADS system, including measures of socio-emotional skills and high school readiness. The CORE data system has also recently added measures of college enrollment and completion from the National Student Clearinghouse (NSC). See Hough, Byun, and Mulfinger (this volume) for more information about CORE and the use of data for continuous improvement.
- Several individual districts have developed their own bilateral partnerships with university-based researchers or small research centers. These include the Los Angeles Education Research Institute (LAERI) in Los Angeles, San Diego Education Research Alliance (SanDERA) in San Diego, and the Stanford/San Francisco Unified School District Partnership in San Francisco. In some cases, the districts have also purchased college enrollment and completion data from the NSC to support this research (see, e.g., Phillips, Yamashiro, and Jacobson, 2017).
- The California Ed Lab at UC Davis (led by Michal Kurlaender), in conjunction with PACE, has been engaged in a broad research effort to investigate alignment and transitions between K-12 and higher education systems in the state, including college and career

readiness, collegiate remediation, college persistence and degree completion, and college quality. This work, largely funded by grants from the Institute of Education Sciences in the U.S. Department of Education, links K-12 data provided by the CDE to California State University and Community College system-wide data.

Other initiatives link education data and data from other state or local programs:

- Federal regulations holding institutions of higher education accountable for the "gainful employment" of their students require colleges to measure the eventual employment and earnings of their students. As part of the implementation of these regulations, the higher education institutions obtain earnings records for their former students from EDD. However, because each system may only obtain data for its own students, the systems are unable to construct the control groups needed to measure the impacts of programs of study relative to the available alternatives.
- The Education Equals Partnership, an effort to improve educational outcomes for students from foster care, has worked with four California counties to link individual-level child welfare and education data and make these data available in real time to support youth. This effort builds on prior work by the California Child Welfare Indicators Project and the Center for the Future of Teaching and Learning at WestEd that linked statewide child welfare and education data to describe foster care students' educational experiences and outcomes (Wiegmann et al. 2014).
- The Silicon Valley Regional Data Trust (SVRDT) is working with three counties in Silicon Valley to build data systems that integrate data from the 27 local school districts with data from child and family services, juvenile justice, mental health, and other social service programs run by the counties. The goal is to permit the development of personalized learning strategies for students that take into account the importance of non-school influences in students' lives.
- The California Policy Lab (CPL), of which the three present authors are part, is a joint initiative of the Berkeley and Los Angeles campuses of the University of California that aims to support better use of the state's administrative data for research that informs program and policy improvement. CPL fosters partnerships between university researchers and state and local agencies, in education and beyond, focused on problems that the agencies identify as essential to their effective operations. CPL has active projects underway with a number of education agencies, including LAUSD, CSAC, and the Community College Chancellor's Office, as well as with a range of non-education agencies.

Many of these efforts have yielded important policy improvements. One notable "win" relates to course placement of new community college students. Cal-PASS Plus's flagship project, the Multiple Measures Assessment Project (MMAP), is a study of the relationship between success in entry-level college courses and various predictors that might be used for placement decisions. These decisions have historically been made based on standardized placement tests administered by the community colleges. But evidence indicated that these placement tests were not as informative as other information, such as high school GPA and

course grades, which were already available in high school districts' data systems (Bahr 2016; Scott-Clayton et al. 2014). Several community colleges, as well as the CSU system, have moved to "multiple measures" placement systems, which has increased the share of students placed initially into college-level courses rather than remedial courses, while maintaining or increasing course passage rates (Bahr et al. 2017). Because initial placement into remedial courses is a strong predictor of eventual dropout without a degree, this simple change promises to meaningfully expand the number of students successfully completing college, while if anything reducing public costs. MMAP relied on a dataset, built specifically for that project, which linked high school transcripts, test scores, and community college data.

The Promise of a Statewide Comprehensive Data System

These efforts illustrate the value of linked data for informing education policy and practice. But they also have limitations. Most important, each of these efforts involved large investments of time and resources, typically over several years. In some cases, the result of these investments is a one-off data system that is useful for the purpose for which it was built but cannot – due to legal or technological limitations – be used for other purposes that arise later. New questions may require a separate effort to build a relevant data system, starting from square one.

A second important limitation is the spotty coverage of this patchwork of data-linking efforts. Many of these efforts required negotiating MOUs, one by one, with each participating district and postsecondary education provider. (A notable exception is the Community College Chancellor's Office, which created Cal-PASS Plus and provided systemwide data with a single MOU, though even Cal-PASS Plus relies on bilateral MOUs with individual school districts and four-year universities.) This means there are holes in the coverage – some districts or colleges participate while their neighbors do not, and the data from some districts and colleges are more up to date than the data from others. This limits the value of the data, because some students disappear from a longitudinal database simply because they move to a non-participating district or college, and it is difficult to know how the results would generalize to the state as a whole.

The incomplete coverage of the existing linked data systems also means that nonparticipating districts do not have access to the valuable information that is available to participants (Moore and Bracco 2018). Districts vary in their capacity to participate in these types of partnerships and it is not surprising that a system based on voluntary initiatives will tend to exclude smaller and rural districts. A more comprehensive, statewide data system would enable these districts to benefit from the same data access that larger, urban districts have.

Aside from equalizing access to data across different types of districts, a statewide data repository would have a number of additional advantages over the patchwork described above and elsewhere (e.g., Moore et al 2017). Most important, statewide data systems make it easier for agencies to serve mobile students, by making it possible for agencies to access relevant

educational, human services, and health data even when students move across district or county lines. Likewise, statewide data make it possible for K-12 districts to more accurately understand students' outcomes in higher education and the workforce even if their former students have moved to a different part of the state.

Statewide data systems also make it possible for districts and counties to compare their outcomes to other districts and counties that serve similar populations, to gain a better understanding of their relative strengths. This facilitates evaluations of local programs because more detailed data make it possible to identify a good comparison group for program evaluation purposes. Of course, having statewide data also makes it easier to evaluate the impact of state policies and to understand how that impact varies across localities.

There are also strong efficiency and equity arguments for a statewide data system. With statewide systems, agencies or regions can benefit from a centralized resource for linking data and developing protocols for sharing and displaying data, rather than re-creating those systems de novo in each locality. These benefits are likely to be especially important for smaller agencies and localities that typically have fewer resources for data management and analysis than their larger counterparts. And while local and regional partnerships are likely to still be important for ensuring that data and analyses remain relevant to local needs (London and Gurantz 2010), local agencies can augment the statewide linked data with their own local data, as desired.

Models from Other States

Many other states have created data systems linking K-12, higher education, and employment data. Some states have also connected data on human services, criminal justice, and/or health, creating integrated data repositories either at universities or within a state agency. The specific system designs vary -- some data systems are updated regularly enough for research and analysis while others have elements that are updated frequently enough to be used in real-time by service providers. These efforts provide useful examples for California to draw upon.

Washington State's linked longitudinal data system, the Education Research and Data Center (ERDC), was created by the Washington Legislature in 2007 and is housed in the Office of Financial Management. The ERDC includes linked data from the K-12, higher education, workforce, and criminal justice sectors. ERDC data have been used to answer a number of descriptive questions about how K-12 students fare in higher education and the labor market (see, e.g., Patterson and Weeks 2016) and how interaction with the criminal justice system, through juvenile courts or incarceration, is related to young adults' educational and employment outcomes (Hough 2016; McCurley, Kigerl, and Peterson 2017). These data are also being used to evaluate state programs, such as a college scholarship targeted to low-income middle school students and intended to encourage college attendance and make college affordable (Goldhaber, Long, Gratz, and Rooklyn 2017).

Texas has also developed a linked longitudinal data system, the Texas Education Research Center (ERC), created by the Texas Legislature in 2006 to house K-12, teacher certification, higher education, and workforce data at two university research centers (UT Austin and UT Dallas). The ERC data have been used to study the relationship between students' math and science coursework in high school and students' completion of a STEM degree in college (Borman et al. 2017), the effects of community college tuition on students' college enrollment (Denning 2017), and the effects of attending Texas charter schools on college enrollment and earnings (Dobbie and Fryer 2016), among other topics. The linked ERC data can also be used by the sponsoring state agencies, which must review and approve every external request for access to the data.

An important part of the development of a state education data repository is the creation of governance rules that specify the conditions under which the data can be used. Typically, state agencies have preferential access, but systems also provide mechanisms for outside researchers to request access. The box below describes the governance processes and outside researcher data access rules in several states that have developed linked, longitudinal data systems, including Washington and Texas.

Data Access and Governance in Selected States

Florida: Florida's Education Data Warehouse (EDW) contains de-identified, individual-level longitudinal data from the K-12, postsecondary, and workforce systems. The Florida Department of Education has a research agenda that outlines priority research topics. When researchers submit proposals to use the data, the EDW seeks a program sponsor for the research and then forwards the proposal to a methodology committee for review. Researchers do not have access to all data elements and, to prevent confusion, are not allowed to calculate their own versions of measures already calculated by the department. Researchers must submit results to the department for review before dissemination. Florida also has an extensive PK-20 Education Information Portal that provides public access to aggregated data on education and employment.

Kentucky: The Kentucky Center for Education and Workforce Statistics (KCEWS) houses individual-level data from the preschool, K-12, postsecondary, and workforce systems. These data are de-identified and made available to state agency staff, approved researchers and, in aggregate form, to the public. Operational costs are offset by data access fees. Agencies contributing data have an opportunity to review research reports before dissemination.

Maryland: The Maryland Longitudinal Data System (MLDS) contains de-identified, individual-level student and workforce data. A Governing Board, composed of K-12, postsecondary, and workforce agency heads, as well as additional members (including the chair) appointed by the Governor, guides the development of the research agenda. For in-depth research, the MLDS partners with faculty at the University of Maryland.

North Carolina: Located at the Duke Center for Child and Family Policy, the North Carolina Education Research Data Center (NCERDC) stores and manages individual-level data on K-12 students and teachers. The NCERDC shares de-identified data with researchers who have received Institutional Review Board approval, have submitted approved Data Security plans, and have agreed to provide results to the North Carolina Department of Public Instruction prior to publication. The NCERDC charges a data access fee to non-students to cover administrative costs. North Carolina is also establishing a linked longitudinal data system (North Carolina SchoolWorks) that includes individual-level early learning, K-12, higher education, and workforce data and is governed by heads of the agencies contributing data to the system.

Texas: Located at UT Austin and UT Dallas, the Education Research Center (ERC), which holds linked, de-identified K-12, postsecondary, and workforce data, reviews proposals to ensure they are methodologically sound and have the potential to benefit education in the state and then forwards the proposals to the ERC Advisory board, consisting of representatives from the Texas Education Agency, the Texas Higher Education Coordinating Board, and the Texas Workforce Commission, for consideration. Operational costs of the ERC are offset by data access fees for accepted proposals. Authorized users access the data through secured Texas ERC computers on a private network, and results are reviewed for disclosure compliance before release. The ERC requires that researchers create a policy brief about the research for public dissemination.

Washington: Located in the state Office of Financial Management, the Education Research and Data Center (ERDC) has a Memorandum of Understanding with all the agencies that contribute data and has three committees that govern data use. The Research and Reporting Coordination committee is composed of ERDC staff, representatives of the organizations contributing data, and representatives of other stakeholder organizations. This committee recommends research priorities. The Data Stewards Committee includes ERDC researchers and staff who are familiar with the data contributed by their agencies. This committee maintains consistent data definitions and considers what new data should be collected. The Data Custodians Committee is composed of technical experts from the agencies contributing data and the ERDC. This committee coordinates data sharing and data protection and storage. The ERDC shares data with other state agencies and researchers through a process in which the ERDC and the agencies that contribute data to the ERDC provide feedback at both the proposal and reporting stages of the research.

Barriers to Developing a Comprehensive Statewide Data System

Developing data systems that will serve teachers, school leaders, policy-makers, and, ultimately, California's children will require the state to overcome a variety of technical and political barriers. That so many states have, in fact, developed effective data systems following a range of models suggests these barriers can be surmounted. We discuss technical barriers first, and argue that these are in fact not difficult to overcome. We then discuss the, often thornier, political barriers.

Technical Barriers

An integrated data system requires obtaining data from many sources, linking them together, storing the linked data, ensuring data quality, sharing them with authorized users, and keeping data secure from those who should not have access. These steps have been accomplished in many other states, and the technical challenges are well understood. California could draw on the experience of other states, the guidance of organizations such as the Data Quality Campaign, and expertise residing in agencies, non-profit organizations, and universities to determine the best solution to these issues. In fact, in the K-12 domain, California has already overcome these technical challenges to develop CALPADS.

Linking Data over Time and Across Agencies

Linking data from different sources and time periods is central to creating an integrated education data system. California law established the CALPADS ID as the unique identifier for use in education data, and the state has successfully linked K-12 data in the CALPADS system.

CALPADS currently includes some data about teachers, and some districts have created and used student-teacher linked data. CDE planned, and the U.S. Department of Education agreed to fund, a statewide database covering teachers and other certified employees, the California Teacher Integrated Data Education System (CalTIDES). The proposed CalTIDES could have been used on its own to understand how teachers move through training and into the profession, or it could have been linked with student-level data in CALPADS. However, Governor Brown vetoed $2.1 million the Legislature appropriated for CalTIDES in 2011, and the project has been tabled since then (EdSource 2016, Senate Bill 87 2011).

Linking K-12 education data to information from other agencies—such as higher education systems, social services agencies, health programs, and Employment Development Department (EDD) data on employment and wages—presents additional challenges since those agencies do not use the CALPADS ID.[1] Without common, unique ID variables, data must be merged using so-called “fuzzy matches” that recognize the possibility of duplicates. For example, sometimes the only common elements between data sets are names and dates of birth. These do not uniquely identify individuals (though they may come close). Methods for conducting fuzzy matches and for recording potential mismatches are well established and used in a wide variety of settings. For example, the North Carolina Education Research Data Center (NCEDRC) has a core database of K-12 data, but the data have been matched to a wide range of data from other agencies, allowing the examination of critical questions for education.

Data linked using fuzzy matches are not appropriate for some uses. Depending on the quality of the original data and the number of data elements available, accuracy rates well above 90 percent are common. But this still leaves some students who will be mis-matched.

---

[1] Many of those agencies, including EDD, identify individuals by Social Security Number (SSN), but California law requires the use of the CALPADS ID in education agencies rather than SSN. The experience of other states shows that data can be reliably linked across sectors even without SSNs.

One would not want to make course placement decisions using these data, for example, without confirming that each student's information is accurate. (Confirming the accuracy of data is, of course, important even when unique IDs are available, because of the possibility of error in any large-scale real-world database.) However, for many other uses, fuzzy matched data are indispensable, and perfectly appropriate. Occasional mis-matches pose little problem for most research, including descriptive studies and program evaluations, or for low-stakes performance measurement or improvement efforts.

A particularly important type of match is between K-12 and higher education data. State law requires the state's three higher education systems to collect the CALPADS ID for incoming students. Although coverage has improved in recent years, this provision has still not been fully implemented (Warren and Hough 2013; Moore, Bracco, and Nodine 2017). Connecting the K-12 CALPADS data to data from the higher education system is an important next step toward a more comprehensive data system that would allow practitioners and researchers to answer important questions about both sectors; for example, we could learn about how students transition from K-12 to higher education and how a range of K-12 programs impact college-readiness and college performance. Warren and Hough (2013) identify several options for collecting CALPADS IDs for college and university students. A recent MOU between CDE and the community college system now allows for automatic transmission of the CALPADS ID to the community college application, and the CSU system is now seeking a similar MOU (Moore, Bracco, and Nodine 2017). Fuzzy matches are also an option for many purposes. Colleges typically record from which high school a student graduated, and often confirm this via official transcripts. In combination with basic biographical information like name and date of birth, these data would support a very accurate match.

Ensuring Data Quality

A key contribution of a centralized education data system is improving data quality for the range of uses described in Section II. Agencies collect data for particular uses, and those data are not always appropriate for other uses or organized in a user-friendly way. In other states with well-developed data systems, the organization housing the data receives various data elements from agencies, converts those data into more usable formats, checks the data for consistency and accuracy, and provides documentation to potential users.[2] These processes improve the quality of both the data and the analysis, and reduce duplication of effort. In California's decentralized system, there is no practical way for data cleaning efforts to diffuse across all of the different systems, so many datasets are plagued with errors and poorly documented, and are generally not harmonized over time.

Political Barriers

While the technical barriers to developing a more comprehensive education data system in California are, for the most part, easily solved, political concerns also loom large.

[2] See Ladd and Munschkin (2008) for a helpful discussion of procedures in North Carolina.

Many states have overcome political barriers and developed comprehensive and accessible data systems through effective leadership and extensive consultation with a wide range of stakeholders, including parents, teachers, school districts, state agencies, and researchers.

Efforts to develop education data systems in California need to acknowledge that data construction and use are not purely technocratic endeavors. They entail political risks—for agencies, advocates, and others—that need to be acknowledged. When measurement is possible, it has the potential of revealing something surprising – for example, that a project or program is not working as well as might have been assumed. Stakeholders may be concerned about this; they may be even more concerned that low-quality research will yield misleading results about program effectiveness.

Agencies will be hesitant to participate in a statewide system if they fear that their data will be misused. (Parents have similar, reasonable concerns – we discuss privacy considerations below.) Agencies have legitimate fears that sharing their data could create work, interfere with their operations, or yield misleading conclusions about their programs' effectiveness. Some agencies may also have an interest in preventing research – even high-quality research – that might reveal their own shortcomings. The value of data will be limited if agencies have veto power over any use of their data that might be inconvenient, without regard to other perspectives. These issues should be addressed transparently in consultation with agencies as data systems are built, and the agreements governing how different data can be used, and by whom, should balance these concerns. In many states, committees composed of agency staff and outside stakeholders develop mutually acceptable rules about data access and the process by which findings are shared (e.g., internally first and publicly only after a review period).

Broadly, three factors are important for maintaining political support for comprehensive education data systems. First, strong leadership is critical; successful efforts in other states have typically required both political leadership and individuals inside the Department of Education and other agencies who understand the value of the work and are committed to developing better data systems and to improving the effectiveness of the education system. Second, state-level actors need to be consistently engaged with stakeholders, particularly county and district officials, throughout the process. Third, the agency in charge of developing data systems should clearly articulate the value of the data system, how it will be used, and how privacy will be protected. In developing support for a comprehensive data system in California, it seems particularly important to emphasize that data can be used for a wide range of purposes other than accountability (see box), and that more comprehensive data systems will make it easier for agencies to serve Californians well.

Data and Accountability

Much of the impetus in other states for the creation of improved data systems in education has come from efforts to construct performance measures to be used in accountability regimes. The No Child Left Behind Act of 2001, for example, mandated that states administer standardized tests in elementary and secondary schools and use these to construct school-level measures that formed the basis of school accountability scores. Many states created integrated data systems as part of this process. Later, the Obama Administration provided funding for states to construct data systems that linked teachers to their students' achievement, and encouraged the use of these systems in teacher evaluations. These policies have been highly controversial, opposed by many who thought that they were unfair to schools and teachers serving high-need populations, wasted time and money that would be better devoted to instruction, and encouraged educators to focus excessively on test scores.

Because the development of data systems was so closely tied to the accountability movement, those who are skeptical of accountability policies have sometimes opposed the data systems themselves (Henig 2012). But it is important to realize that many uses of linked administrative data, including those discussed in Section II, are unrelated to accountability (particularly the high-stakes form of teacher evaluation that is the focus of most controversy). When data systems become pawns in the accountability wars, this impedes all other uses of improved data, with negative consequences for students and educators.

Governance and Access

The state will need a plan to determine where data will be stored and, critically, who has access under what conditions. This requires secure IT infrastructure that can prevent private information from being inadvertently released. Other states offer models for how this can be accomplished (see box on states' data access and governance): Data can be consolidated and stored within an existing agency, within a newly created single-purpose agency, in a university or non-profit organization, or in some combination or consortium of these. The relationships among the entities holding the data, the original agencies, and different users of the data are typically governed by Memoranda of Understanding (MOUs) and Data Use Agreements (DUAs) that specify who can have access to which types of data under what conditions. In consultation with stakeholders, California could develop procedures for de-identifying data for analysis, access guidelines and review protocols for assessing requests to use the data, and IT systems for sharing the data securely.

As the state develops IT infrastructure and protocols, California should not give data access short shrift. If the state builds the data without making them available to practitioners, agency staff, policymakers, and researchers, the effort will be wasted. As discussed above, the state successfully built the K-12 CALPADS database, but the data are underutilized for research and practice. The state should work with stakeholders to develop policies allowing those whose work could improve the education and well-being of California's children to access the statewide data repository. Those policies should pay particular attention to privacy and security, ensuring compliance with relevant privacy laws and minimizing the risk of unauthorized disclosure (see below for further discussion of privacy). For example, for many purposes, de-identified or aggregate data are sufficient and reduce the risk of unauthorized

disclosure. Moreover, a well-designed central repository has the potential to significantly enhance privacy relative to the current patchwork of systems.

Another essential part of the system design is ensuring that resources are available to conduct the analyses that state agencies, districts, schools, and educators most need to serve students well (Connaway, Keesler, Schwartz 2015). This can be accomplished by giving agencies priority access to data, by prioritizing projects identified by agencies and educators as important in the review process, by explicitly funding this work, and by developing partnerships with outside researchers that can be funded externally. Tennessee has developed an internal Research and Strategy Division that offers a potential model for how California might organize its own research capacity and develop partnerships with other organizations (see the discussion by Moffitt et al. in this volume).

Privacy

Increased collection and use of student data has raised privacy concerns among parents, particularly when those data are turned over to for-profit education technology firms or other outside groups for purposes parents may not perceive as beneficial to their children (EdWeek 2014). Ensuring the protection of student privacy is required by the Federal Education Rights and Privacy Act (FERPA) and by state laws. Additional privacy laws, such as HIPAA, will also apply when other types of data such as health records are linked to education data. Protecting privacy is critical to building trust with agencies, students, parents, stakeholders, and the public.

Discussions and policy around data privacy should therefore be central to the planning for education data systems. At the same time, these concerns need not prevent the use of education data for analytical purposes, with appropriate safeguards, and procedures should be put in place to guard against efforts to use privacy concerns as an excuse to restrict data access. Moreover, it is important to recognize that a well-designed, centralized state data system offers the opportunity to *enhance* privacy protection relative to the status quo, by developing state-of-the-art information security systems and reducing the number of users who obtain access to personally identifiable information (PII).

What is FERPA?

The Family Educational Rights and Privacy Act (FERPA) was first passed in 1974 and is the main federal law governing education data privacy. Education agencies that receive federal funding—including CDE, school districts, and schools—must comply with the provisions of FERPA specifying the conditions under which education data can be disclosed to third parties. In general, students and/or parents must consent to having their education data disclosed to third parties for a particular purpose. However, the law permits disclosure without consent under some circumstances, including for research (the "studies" exception) and to evaluate and audit education programs (the "audit or evaluation" exception).

FERPA is often thought to prevent the use of identified data for research and other purposes, but this is a misconception. The U.S. Department of Education has issued guidelines clearly stating that a wide range of research with individual-level data is permitted by FERPA, so long as the potential benefits of the research outweigh the risks and proper procedures are followed to limit the risk of unauthorized disclosure (EducationCounsel, LLC, 2011).

The most recent regulations clarified how the law affects state education agencies and State Longitudinal Data Systems in particular. These regulations clarify several points relevant to the use of state-level education databases under the studies or audit and evaluation exceptions. The regulations (1) allow education data to be stored and analyzed at a non-education agency; (2) clarify that data from one part of the education system can be disclosed to another part of the education system; for example, the higher education system can provide data to the state or local districts for purposes of evaluating programs and practices in the K12 system; (3) broaden the definition of education programs that can be studied under the evaluation and audit exceptions; and (4) clarify that state and local agencies can enter into agreements with organizations conducting research under the relevant exceptions and makes the research studies provisions of the law applicable to data held by state education agencies, even if those agencies did not initially collect the data. (EducationCounsel, LLC, 2011).

Altogether, the updated regulations make it clear that FERPA need not be a barrier to developing and using state-level education data systems. Hundreds of FERPA-compliant studies have been conducted using administrative data received directly from agencies, from SLDS's, or from city- or district-specific data consortia. California can turn to the U.S. Department of Education and California guidelines (see, e.g., Harris et al., n.d.), as well as the experiences of other education data consortia, for models of how to protect privacy without unnecessarily hampering analysis.

In addition to FERPA, a number of state laws regulate the use of California administrative data. The state constitution recognizes an "inalienable right" to "pursuing and obtaining" privacy. The primary law governing the application of this to administrative data is the Information Practices Act (IPA), which protects the right to privacy by regulating the "maintenance and dissemination of personal information" by the government. The act prohibits state and local agencies from disclosing personal information in ways that link the information to the individual to whom it pertains without the individual's consent. However, IPA contains a number of exceptions to this prohibition to permit uses that do not violate individual privacy. One allows disclosure in non-identifiable form for "statistical research or reporting purposes" (section 1798.24(h); another allows disclosure to the University of California, a nonprofit educational institution, or another nonprofit entity for scientific research, provided that there is a plan in place to protect personal information from improper use and disclosures and that the study has been reviewed by the state Committee for the Protection of Human Subjects (section 1798.24(t). Nearly all of the uses of "frozen" data for research, analysis, or performance improvement discussed above are permitted under these or other clauses.

The Data Quality Campaign identifies broad principles for safeguarding student data (Data Quality Campaign 2016)[3]:

- Information about how data are collected, stored, and used should be clearly communicated to agencies and the public.
- Appropriate processes and procedures to physically, technically, and legally safeguard the data should be developed and implemented.
- Appropriate governance structures should be put in place to ensure privacy procedures are implemented consistently and with fidelity. This should include the creation of a chief privacy officer or similar high-level position.

The Data Quality Campaign (2016) provides more detail and best practices related to each of these principles.

One part of a strategy for protecting privacy is to limit access to the most sensitive data to the small share of users who truly need it. As discussed in Section II, different use cases require different forms of data. Many states have set up tiered access systems, where users obtain access only to the data they need for their approved research or measurement purposes. Users who do not require personal identifiers can be provided access only to de-identified data. This can greatly reduce the risk of accidental disclosure of personal information, particularly if identifiers are stored completely separately.

An advantage of a centralized California repository is that it could also centralize data linking work, when personal information must be used. If data linking happened more centrally, it would be much easier to manage the process and ensure that the highest security standards were upheld. Then, once data were linked, identifiers could be removed, and made available only to projects that required them, and only under appropriately secure conditions. In the current patchwork, a much wider group of analysts accesses sensitive information, and standards for protection and segregation of this information are disparate.

A careful strategy for the segregation of identifiable information could address one of the most serious concerns that has been raised about improved data systems in California. In recent years, California has clashed with the federal government about immigration enforcement, and some are concerned that data collected and stored by California agencies could be used by federal officials for immigration enforcement. Care should be taken to make sure that any new data systems do not increase the risk to immigrant children and their families. At the same time, understanding the educational needs and outcomes of immigrants is critical so that California schools can better meet the needs of these students. These risks and benefits will have to be carefully considered and balanced. The state might consider, for example, storing any information about immigration status separately from other data, without

[3] https://2pido73em67o3eytaq1cp8au-wpengine.netdna-ssl.com/wp-content/uploads/2016/03/DQC-roadmap-safeguarding-data-June24.pdf

direct links to personally identifiable information, and restricting the set of uses for which such data can be made available.

Communicate how data will be used to improve practice. To build political support for developing a comprehensive educational data system, state leaders need to consistently communicate the many ways that the system can be used to improve the practice of education in California (see the discussion by Hough, Byun, and Mulfinger in this volume). The state can draw on the work of local consortia—such as CORE, LAERI, and the Silicon Valley Regional Data Trust, as well as the examples of informative research done in other states that cannot be done in California—to show how school districts can benefit. Not all parts of the data system can or should be built at once, and the state should prioritize the areas identified by districts and state policy-makers as most pressing. For example, the Legislative Analyst's Office (LAO) called for the state to develop a teacher database to help understand and address teacher shortages (LAO 2016). Similarly, districts want to know more about their students' needs and experiences outside of school, something linkages to social services and other state data could accomplish. Currently, some districts' need for this information is met by more local data consortia, but smaller districts may not have the funding or capacity to participate. Drawing on the experience of the state's data-use leaders to bring the most useful aspects of those data systems to all districts in the state should be a priority.

Seek funding and partnerships to defray costs and develop expertise. Building, maintaining, and monitoring access to a comprehensive state education data system will require dedicated funding. Such funding would be a tiny fraction of total education spending and would allow policy-makers at all levels to allocate the more than $50 billion that the State of California spends on K-12 education more effectively. There is substantial philanthropic support available, along with federal funding, for this type of project.

To get the most out of its data systems, the state will also need to build the human resources, in local and state agencies and in the state's universities and independent research organizations, to analyze, display, and interpret education data. The state could begin by building on the financial and human capital investments already made by private organizations and local agencies, for the benefit of the whole state (Hough, Byun, and Mulfinger, this volume).

Many other states have found partnerships with researchers an important part of their strategy for expanding their capacity for data linking and analysis. The California Policy Lab exists to support this type of research partnership, and its affiliates and other researchers throughout the state are eager to help build systems that can support the types of data use and analysis that will contribute to better outcomes for California's children.

Conclusion

Improvements in computing and database design have dramatically lowered the cost of storing and analyzing large administrative datasets. California has a great opportunity to make use of the data it already collects to better understand and coordinate the activities of the

state's education system. California has made considerable progress in recent years, by developing the CALPADS database at CDE and through a number of more localized partnerships. But despite this progress, the state lags behind its peers, which have done much more to create infrastructure that allows them to use their data systems to support improved instruction and educational practice.

California can catch up. The state already collects a great deal of useful data, and a lot of it is already included in CALPADS and other databases. The state's research community has extensive expertise that could support public agencies in building and using a more integrated data system. Political leadership is critical to overcoming inertia to develop more comprehensive education data systems and, critically, to make them available to educators, state agencies, researchers, and policy makers.

California's next governor has a great opportunity to move California forward on this issue. Clear instruction to the state's education agencies, and an insistence that the data that each holds is a public trust to be used for the good of the state's children, could set in motion the creation, in relatively short order, of a statewide integrated data system. This would allow the state's schools and districts, institutions of higher education, and other public service providers to improve their service coordination, to better understand how well programs and policies are working, and to generate the knowledge that will support the next generation of improvements in California education.

References

Allensworth, E. (2013, January). The use of ninth-grade early warning indicators to improve Chicago schools. *Journal of Education for Students Placed at Risk (JESPAR)*, 18(1), 68–83. doi:10.1080/10824669.2013.745181.

Allensworth, E., and Easton, J. (2005, June). *The on-track indicator as a predictor of high school graduation.* Chicago, IL: The University of Chicago Consortium on School Research. Retrieved from https://consortium.uchicago.edu/publications/track-indicator-predictor-high-school-graduation.

Bahr, P. R., Fagioli, L. P., Hetts, J., Hayward, C., Willett, T., Lamoree, D., Newell, M. A., Sorey, K., and Baker, R.B. (2017, October 11). *Improving placement accuracy in California's community colleges using multiple measures of high school achievement*. San Rafael, CA: the RP Group. Retrieved from https://rpgroup.org/Portals/0/Documents/Projects/MultipleMeasures/Publications/Bahr_et_al-2017-Improving_Placement_Accuracy_in_California.pdf.

Bahr, P. R. (2016, April 13). *Replacing placement tests in Michigan's community colleges*. Ann Arbor: Center for the Study of Higher and Postsecondary Education, University of Michigan. Retrieved from https://umich.app.box.com/v/Bahr-2016-PlacementTests.

Borman, T., Margolin, J., Garland, M., Rapaport, A., Park, S. J., and LiCalsi, C. (2017). *Associations between predictive indicators and postsecondary science, technology, engineer-ing, and math success among Hispanic students in Texas* (REL 2018–279). Washington, D.C.: U.S. Department of Education, Institute of Education Sciences, National Center for Edu-cation Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved from https://files.eric.ed.gov/fulltext/ED577564.pdf.

Brewer, D., and Smith, J. (2007). Evaluating the "crazy quilt": Educational governance in California. *Getting Down to Facts*. Stanford, CA: Stanford Center for Education Policy Analysis. Retrieved from http://cepa.stanford.edu/content/evaluating-crazy-quilt-educational-governance-california.

California Budget Act of 2011, SB 87 (2011-2012), Chapter 33, (Cal. Stat. 2011). Retrieved from California Legislative Information http://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201120120SB87.

Cal-PASS Plus. (n.d.). *About Cal-PASS Plus*. Sacramento, CA: Cal-PASS Plus. Retrieved from http://www.calpassplus.org/MediaLibrary/calpassplus/publicweb/Documents/help/About_Cal_PASS_Plus.pdf

Cannata, M. A., Smith, T. M., and Haynes, K. T. (2017, July). Integrating academic press and support by increasing student ownership and responsibility. *AERA Open*, 3(3). Retrieved from doi:10.1177/2332858417713181.

Clotfelter, C. T., Ladd, H. F., Muschkin, C. G., and Vigdor, J.L. (2013, November). Success in community college: Do institutions differ? *Research in Higher Education*, 54(7), 805–24. Retrieved from doi:10.1007/s11162-013-9295-6.

Conaway, C., Keesler, V., and Schwartz, N. (2015, May). What research do state education agencies really need? The promise and limitations of state longitudinal data systems. *Educational Evaluation and Policy Analysis*, 37(1), 16S–28S. Retrieved from doi:10.3102/0162373715576073.

Data Quality Campaign. (2015, January 12). *Roadmap to safeguarding student data*. Washington, D.C.: Data Quality Campaign. Retrieved from https://dataqualitycampaign.org/resource/roadmap-safeguarding-student-data/.

Denning, J. T. (2017, May). College on the cheap: Consequences of community college tuition reductions. *American Economic Journal: Economic Policy*, 9(2), 155–88. Retrieved from doi:10.1257/pol.20150374.

Dobbie, W. S., and Fryer, Jr. R. G. (2016, August). Charter schools and labor market outcomes. Working Paper, 22502. *National Bureau of Economic Research*. Retrieved from doi:10.3386/w22502.

Dodge, K. A., Bai, Y., Ladd, H. F., and Muschkin, C. G. (2017, May). Impact of North Carolina's early childhood programs and policies on educational outcomes in elementary school. *Child Development*, 88(3), 996–1014. Retrieved from doi:10.1111/cdev.12645.

Ebudget. (2017, June 27). *State budget: Enacted budget detail*. Retrieved from Ebudget http://www.ebudget.ca.gov/budget/publication/#/e/2017-18/BudgetDetail.

EducationCounsel, LLC. (2011, December 2). *U.S. Department of Education final FERPA regulations: Advisory and overview*. Washington, D.C.: Data Quality Campaign. Retrieved from https://dataqualitycampaign.org/resource/u-s-department-education-final-ferpa-regulations-advisory-overview/.

Feng, L., and Sass, T. (2015, September). *The impact of incentives to recruit and retain teachers in "hard-to-staff " subjects: An analysis of the Florida critical teacher shortage program* (Working Paper, 141). Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research. Retrieved from https://caldercenter.org/sites/default/files/WP%20141.pdf.

Fensterwald, J. (2016, February 29). Renewed call to create statewide teacher database. *EdSource*. Retrieved from https://edsource.org/2016/renewed-call-to-create-statewide-database-on-teachers/95339.

Fraser, J. (2015, August). *Improving educational and well-being outcomes: School-DHS data sharing in Allegheny County*. Pittsburgh, PA: Allegheny County Department of Human Services. Retrieved from https://www.alleghenycountyanalytics.us/wp-content/uploads/2016/06/Improving-Educational-and-Well-Being-Outcomes-8-19-15.pdf.

Goldhaber, D., Long, M. C., Gratz, T., and Rooklyn, J. (2017). *The effects of Washington's College Bound Scholarship Program on high school grades, high school completion, and incarceration* (Working Paper, 05302017-2–1). Seattle: Center for Education Data and Research, University of Washington. Retrieved from http://www.cedr.us/papers/working/CEDR%20Working%20Paper%20No.%2005302017-2-1.pdf.

Harris, K., Torlakson, T., and Lightbourne, W. (n.d.). Dear colleague letter: Foster youth information sharing. California Bureau of Children's Justice, California Department of Education, and California Department of Social Services. Retrieved from https://oag.ca.gov/sites/all/files/agweb/pdfs/bcj/fy-info.pdf.

Henig, J. R. (2012). The politics of data use. *Teachers College Record*, 114(11), 1-32.

Hough, G. C. (2016). *Education and employment characteristics of incarcerated young adults*. Olympia: Washington State Statistical Analysis Center. Retrieved from

erdc.wa.gov/publications/justice-program-outcomes/education-and-employment-characteristics-incarcerated-young.

Hough, H., Byun, R, and Mulfinger, L. (2018). Using Data for Improvement: Learning from the CORE Data Collaborative. *Getting Down to Facts 2*: Stanford, CA.

Jacob, B. Dynarski, S., Frank, K., Schneider, B. (2016, February). *Are expectations alone enough? Estimating the effect of a mandatory college-prep curriculum in Michigan* (Working paper 22013). Washington, D.C.: National Bureau of Economic Research. Retrieved from doi:10.3386/w22013.

Kamisar, B. (2014, January 7). InBloom sputters amid concerns about privacy of student data. *Education Week*. Retrieved from https://www.edweek.org/ew/articles/2014/01/08/15inbloom_ep.h33.html?r=182194026.

Ladd, H. F., and Muschkin, C. G. (2008, April 15). *Research access to state education administrative data: The North Carolina Education Research Data Center.* Durham: The North Carolina Education Research Data Center.

Loeb, S., Beteille, T., and Perez, M. (2008, February). Building an information system to support continuous improvement in California public schools. *Policy Analysis for California Education*. Stanford, CA: Stanford Center for Education Policy Analysis. Retrieved from https://cepa.stanford.edu/content/building-information-system-support-continuous-improvement-california-public-schools.

London, R. A., and Gurantz, O. (2010, April). Data infrastructure and secondary to postsecondary tracking." *Journal of Education for Students Placed at Risk (JESPAR)*, 15(1–2), 186–99. Retrieved from doi:10.1080/10824661003635259.

McCurley, C., Kigerl, A., and Peterson, A. (2017). *Students before and after juvenile court dispositions: Student characteristics, education progress, juvenile court dispositions, and education outcomes in Washington State*. Olympia: Washington Center for Court Research, Administrative Office of the Courts.

Moffitt, S.L., Cohen, D.K., Lyddon, M.J., O'Neill, M.K., Smith, K.B., Willse, C. (2018). Structures for Instructional Support. *Getting Down to Facts 2*: Stanford, CA.

Moore, C., Bracco, K., and Nodine, T. (2017, August). *California's maze of student information: Education data systems leave critical questions unanswered*. Sacramento, CA: Education Insights Center. Retrieved from http://edinsightscenter.org/Publications/Research-Reports-and-Briefs/ctl/ArticleView/mid/421/articleId/2189/Californias-Maze-of-Student-Information-Education-Data-Systems-Leave-Critical-Questions-Unanswered.

Moore, C., and Bracco, K. R. (2018, January). *Scaling goodwill: The challenges of implementing robust education data sharing through regional partnerships.* Sacramento, CA: Education Insights Center. Retrieved from http://edinsightscenter.org/Publications/Research-Reports-and-Briefs/ctl/ArticleView/mid/421/articleId/2191/Scaling-Goodwill-The-Challenges-of-Implementing-Robust-Education-Data-Sharing-Through-Regional-Partnerships.

Moore, C., Grubb, B, and Esch, C. (2016, November). *Gaps in perspective: Who should be responsible for tracking student progress across education institutions?* Sacramento, CA: Education Insights Center. Retrieved from http://edinsightscenter.org/Publications/Research-Reports-and-

Briefs/ctl/ArticleView/mid/421/articleId/2182/Gaps-in-Perspective-Who-Should-Be-Responsible-for-Tracking-Student-Progress-Across-Education-Institutions.

National Cooperative Education Statistics System. (2016, July). *Forum guide to education data privacy*. Washington, D.C.: National Forum on Education Statistics. Retrieved from https://nces.ed.gov/pubs2016/NFES2016096.pdf.

Paterson, T., and Weeks, G. (2017). *STEM bachelor's degrees in Washington State: The gender deficit by major and race category*. Olympia, WA: Education Research and Data Center. Retrieved from https://erdc.wa.gov/publications/economic-returns/stem-bachelors-degrees-washington-state-gender-deficit-major-and-race.

Phillips, M., Yamashiro, K., and Jacobson, T. (2017, August). *College going in LAUSD: An analysis of college enrollment, persistence, and completion patterns*. Los Angeles, CA: Los Angeles Education Research Institute. Retrieved from http://laeri.org/ed/wp-content/uploads/2017/08/laericollegegoing082017.pdf.

Pitcher, M. A., Duncan, S. J., Nagaoka, J., Moeller, E., Dickerson, L. and Beechum, N.O. (2016, November). *The network for college success: A capacity-building model for school improvement*. Chicago, IL: The University of Chicago Consortium on School Research. Retrieved from https://consortium.uchicago.edu/publications/network-college-success-capacity-building-model-school-improvement.

Roderick, M., Kelley-Kemple, T., Johnson, D. W., and Beechum, N. O. (2014, April). *Preventable failure: Improvements in long-term outcomes when high schools focused on the ninth grade year: Research Summary*. Chicago, IL: The University of Chicago Consortium on School Research. Retrieved from https://consortium.uchicago.edu/publications/preventable-failure-improvements-long-term-outcomes-when-high-schools-focused-ninth.

Santibañez, L. and Snyder, C. (2018). Teaching English Learners in California: How Teacher Credential Requirements in California Address the Needs of ELs. *Getting Down to Facts 2*: Stanford, CA.

Scott-Clayton, J., Crosta, P. M., and Belfield, C. R. (2014, September). Improving the targeting of treatment: Evidence from college remediation. *Educational Evaluation and Policy Analysis*, 36(3), 371–93. Retrieved from doi:10.3102/0162373713517935.

Stevens, A. H., Kurlaender, M. and Grosz, M. (2015, April). *Career technical education and labor market outcomes: Evidence from California community colleges* (Working Paper, 21137). National Bureau of Economic Research. Retrieved from doi:10.3386/w21137.

Taylor, M. (2016, February). *The 2016-17 Budget: Proposition 98 education analysis*. Sacramento, CA: Legislative Analyst's Office. Retrieved from http://www.lao.ca.gov/Reports/2016/3355/prop-98-analysis-021816.pdf.

Warren, P., and Hough, H. (2013, August). *Increasing the usefulness of California's education data*. San Francisco, CA: Public Policy Institute of California. Retrieved from http://www.ppic.org/publication/increasing-the-usefulness-of-californias-education-data/.

Wiegmann, W., Putnam-Hornsten, E., Barrat, V. X., Magruder, J., and Needell, B. (2014). *The invisible achievement gap part 2: How the foster care experiences of California public school students are associated with their education outcomes.* San Francisco, CA: Stuart

Foundation. Retrieved from http://stuartfoundation.org/wp-content/uploads/2016/04/IAGpart2.pdf.