A Better Student Data System for California

THE PUBLIC FORUM ON
SCHOOL ACCOUNTABILITY

Members of the Data Systems Panel,
The Public Forum on School Accountability

Phil Daro, Director, The Public Forum on School Accountability

Don Barfield, Chief Development Officer, WestEd
Kathleen Barfield, CDE
George Bohrnstedt, Senior Vice President, AIR
Russ Brawn, Information Systems Administrator, CSIS
Camille Esch, Education Policy Analyst, SRI
Neal Finkelstein, Director, UCOP
Ron Fox, Administrator, Intersegmental Relations, CDE
Bob Friedman, Chief Operations Officer, CSIS
Mike Garet, Chief Research Scientist, AIR
Pete Goldschmidt, Senior Researcher, UCLA
Laura Hamilton, Behavioral Scientist, RAND
Phillip Kaufman, Senior Research Associate, MPR Associates
Don McGlaughlin, Chief Scientist, AIR
Bill Padia, Director, Policy and Analysis Division, CDE
Russ Rumberger, Director, UC Linguistic Minority Research Institute, UCSB
Patrick Shields, Director, SRI
Brian M. Stetcher, Senior Social Scientist, RAND
Brad Strong, Legislative Director, EdVoice

The Public Forum on School Accountability would like to thank the following foundations for funding this project:

The William and Flora Hewlett Foundation
The James Irvine Foundation
The Stuart Foundation
The San Francisco Foundation
The East Bay Community Foundation
The S.H. Cowell Foundation
The Clarence E. Heller Charitable Foundation
The Walter S. Johnson Foundation
The Wallace Gerbode Foundation.

The views expressed herein are those of the authors, not of these organizations.

Summary

California needs legislation to establish a student data system that complies with the federal law *No Child Left Behind* and serves the state's own long-term needs for accurate data on the effectiveness of its public schools. The accountability systems at state and local levels require a high-quality student data system. Research and evaluation efforts also depend on such a system to answer questions about impact on student achievement. A student data system with features described in this report can help the state meet the following critical goals:

- Reduce the chances of erroneous conclusions due to measurement error, sampling error, and non-response error. Volatility in year-to-year scores based on comparing different students leads to unnecessary risk of erroneous conclusions. More accurate conclusions can be drawn from combining same-grade comparisons with longitudinal results of the same students over time.
- Focus schools on improving each student's achievement as well as increasing the number who score above the standard passing score.
- Hold schools accountable for the value they add to students' achievement. Linked longitudinal scores enable scientific control for prior achievement of students.
- Sustain local public and parental support for the accountability system by providing more accurate and stable data on trends from year to year.
- Fulfill new federal requirements for measuring Adequate Yearly Progress with more valid and accurate longitudinal information than cross-sectional data alone can provide.
- Coordinate data between higher education and K-12 in both directions for better evidence-based planning in higher education and K-12.
- Supply data for policy research on the relationships among school, employment, and the workforce.
- Save millions of dollars in future costs of evaluating state programs. Because the state lacks a data system that measures students' longitudinal growth, current studies of state initiatives such as class size reduction or state-sponsored professional development programs require evaluators to construct the needed databases from scratch, often at great expense.

Legislation that facilitates such a system must address a variety of issues, including designating the responsible entity. This report focuses on four issues we think should be addressed in the legislation establishing California's data system:

- Establish encrypted student identifiers based on Social Security Numbers.
- Ensure the highest level of data confidentiality.
- Clearly delimit access and uses.
- Enable ongoing coordination across client systems.

Introduction

Over the last decade, Californians have invested billions in improving their schools. Class size reduction, expanded teacher professional development, increased emphasis on achievement testing at all grades, high school exit exams, and new college outreach programs have all promised improved educational outcomes for all California's youth.

Under Dede Alpert's leadership, the California Legislature's Joint Committee to Develop a Master Plan for Education, Kindergarten Through University, has drafted a new *Master Plan for Education in California.* As legislation develops to implement the recommendations of the Joint Task Force, the opportunity arises to address student data system problems. Indeed, the Task Force specifically proposes that:

> *The State should develop and report yearly on a comprehensive set of educational indicators, constructed form the data provided by an integrated, longitudinal, learner-focused data system and from other school-level data about educational resources, conditions and learning opportunities.*[1]

Without a cohesive and first-rate data system, it will be difficult to track the state's progress toward the larger goal of educational excellence for all students. California's current patchwork of data collection and reporting systems can't inform Californians whether the massive investment they have been making in their educational system is making a difference. This paper provides a framework for creating the data system that our investment in education demands.

In the spring of 2002, the Public Forum on School Accountability assembled a group of educational researchers to discuss California's data needs. The group, which comprised some of the most experienced and esteemed scholars of educational research on policy analysis in the state, was charged with laying the groundwork for a coherent, working data system that will address the needs of educators, researchers, and policymakers alike.

[1] *Master Plan for California Education,* draft 1, p. 46.

I. The Current Data System and Its Limitations

Currently, data on education in California come from several sources. Most information is collected from schools and districts in aggregated form by various distinct programs within the California Department of Education (CDE) and thus cannot be integrated. The CDE annually collects a broad array of information about students and programs in 30 to 40 separate reports from school districts and county offices.

One of the most comprehensive reports on the K-12 systems comes from the California Basic Educational Data System (CBEDS), a database that collects data directly from county offices of education and school districts. CBEDS gathers information on staff and student characteristics as well as enrollment and hiring practices. Three separate forms are used to collect this data: the County/District Information Form, which gathers data on staff and enrollment; the School Information Form, which collects staff and enrollment data specific to schools; and the Professional Assignment Information Form, which collects data on certificated staff from county offices of education and school districts.

The University of California and the California State University systems maintain separate data systems, and each community college district maintains its own data system. An Intersegmental Data Coordination Workgroup tries to span the K-12 and postsecondary segments within the overall K-16 system, but it is an informal group that meets only quarterly and focuses on sharing information about each separate system rather than truly integrating the systems.

While the current data system can be useful for examining basic descriptive information about schools and colleges, it has several major limitations for providing information to various constituencies that can be used to improve the full K-18 education system in California.

One of its primary limitations is in answering questions about students and student learning within the K-12 system. That is, the current system is adequate for what it was originally designed for: reporting school-level and grade-level statistics. The current school report cards now report on student achievement—what students within a given school know. They do not (and cannot) report on student learning—how much students are gaining in their academic knowledge.

Under the new accountability requirements of the state, which hold schools responsible for student performance, this structure is no longer adequate. With the passage of the *No Child Left Behind Act* of 2001, states are required to set challenging academic standards and measure students' progress against a set of standards. The new law encourages (but does not require) states to create data systems that link individual student test score, enrollment, and graduation records over time. Many states are taking this as an opportunity to upgrade their data systems to be able to track students over time. As Chrys Doherty of Just for the Kids states:

> *States that create these [upgraded] systems will benefit in two ways:*
>
> *1. They will more easily satisfy specific reporting requirements in the NCLB legislation.*
>
> *2. They will give educators, parents, and policymakers the powerful information they need to improve schools.*[2]

Just for the Kids finds California's data system deficient in two ways—in linking data and in access:

> ***Linked Records***
> *California collects information on the performance of students present on the day of the test. However, California does not collect student-level fall enrollment data from the school districts and does not have a consistent student ID to link the spring test score information with the enrollment data from the current and previous years. Therefore, the state data cannot be used to identify how long each tested student has been enrolled in the same school or district. Nor can California match each student's test scores with the same student's scores from prior years.*
>
> *Linking fall student-level enrollment and test score data would make it possible to report separately on students who have been enrolled in the same school for three years. Linking test score data for the same students would make it possible to disaggregate middle or high school students based on their level of academic preparation when they entered the school, and compare each California middle or high school with other schools whose students walked in the door equally or less well prepared.*
>
> ***Access to Data***
> *Replication of the Just for the Kids data picture in California requires that the organization or agency doing the replicating have access to student-level data under conditions that safeguard the privacy of individual students. We have not yet received*

[2] Doherty, Chrys; *Getting Smart about Data: Satisfying Federal Reporting Requirements While Helping Schools Improve.* (Forthcoming.)

> *information from California on the conditions under which student-level data would be available to outside researchers or nonprofit organizations.*[3]

The California School Information Services (CSIS) system is an attempt to overcome some of the management problems with California's education data by coordinating the electronic transfer of student records among Local Education Agencies (LEA) and between the LEA and the State. The mission of CSIS is to:

- Build the local capacity of LEAs to implement and maintain comparable student information systems that will support information exchange through CSIS.
- Promote the timely electronic transfer of student records between LEAs as students move between districts, and to postsecondary institutions as students seek entry to higher education.
- Streamline the reporting of data between LEAs and the CDE to reduce the local reporting burden.

CSIS is a voluntary program and is not implemented within all California districts. Furthermore, CSIS does not include data from the STAR system and thus does not help link test score data for individual students, nor does it help California comply with the requirements of *No Child Left Behind.*

The current system's inability to adequately measure student progress is not its only weakness. The current system cannot link the various components of the educational system. For example, it cannot track students through the K-12 system into the Community College, State University, or University of California systems, hampering the ability of higher education planners to plan and evaluate programs.

At long last, serious scientific research into educational programs has become a focus of federal and academic interest. The most important step in desired research programs is connecting instructional variables to student outcomes. Yet California's system cannot match students with teachers. This effectively weakens research into instructional programs, teacher practices, programs for teachers, and other key input variables that could be associated with student outcomes, and it makes such research unnecessarily expensive.

It is also not currently possible to track students out of the school system into the world of work. While California does have the Performance-Based Accountability (PBA) system to help meet the requirements of the federal Workforce Investment Act, it follows only participants in several key programs, not all students within the state. PBA does use

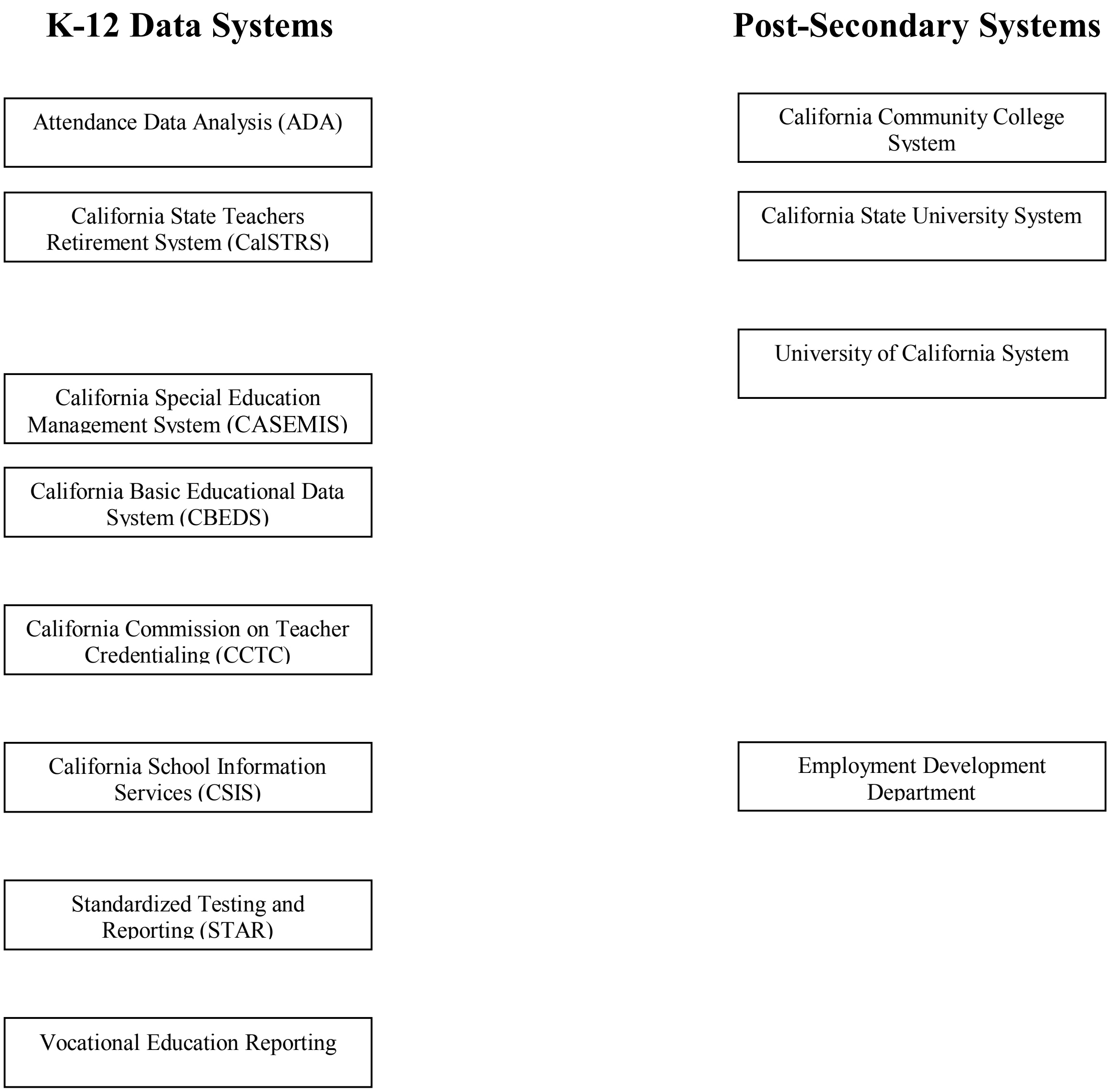
[3] Excerpted from the Just for the Kids Web site, at http://www.just4kids.org/US/California.asp.

Social Security Numbers to link to Unemployment Insurance data, but it does not include those students who do not go on to postsecondary education, nor does it include all levels within the system. An expanded system that included all California students would give policymakers a fuller picture of workforce development in the state.

As Figure 1 below shows, the system is comprehensive in that it covers most areas of the educational enterprise, but it is disconnected in that data from different sources cannot be linked.

Figure 1

Current Data Systems



K-12 Data Systems

- Attendance Data Analysis (ADA)
- California State Teachers Retirement System (CalSTRS)
- California Special Education Management System (CASEMIS)
- California Basic Educational Data System (CBEDS)
- California Commission on Teacher Credentialing (CCTC)
- California School Information Services (CSIS)
- Standardized Testing and Reporting (STAR)
- Vocational Education Reporting

Post-Secondary Systems

- California Community College System
- California State University System
- University of California System
- Employment Development Department

II. Advantages of a New Data System

California needs a student-based data system that can answer three fundamental questions about our educational system: 1) How much does each school add to its students' prior achievement? What are the trends in achievement over time? 2) How well are current reforms working to increase student learning? 3) What are the factors that impede student or accelerate educational progress? The advantages of such a data system are that it will:

- Answer questions about each school's annual contribution to its students' achievement.
- Improve accuracy in determining whether a school is improving, getting worse, or stagnating. Reduce year-to-year wavering back and forth across achievement targets due to statistical error.
- Answer questions about the impact of major state programs on student achievement.
- Provide a database for answering future policy, policy research, and educational research questions about student achievement and its relationships to other variables.
- Provide a shared core of student data accessible by client systems across agencies and institutions.

1. How well are students learning over time?

The most fundamental question for any educational system is: How much are students learning? California's current data system is not fully capable of answering that basic question. The main factor inhibiting the current data system's ability to accurately assess student progress is that it is only cross-sectional in nature. The Standard Assessment and Reporting (STAR) system reports annually on each class of students within a given school—for example, this year's 3rd grade class at Kennedy Elementary School. These types of scores are generally referred to as "level" or "status" indicators (Meyer 1995; Carlson 2002). In the next year, STAR reports how the next cohort of Kennedy 3rd graders performed.

The Academic Performance Index (API) also is based on the difference between the performance of last year's Kennedy 3rd graders and this year's. This "successive cohort" approach (Carlson 2002; Robert L. Linn 2002) has serious weaknesses, the most obvious of which is that differences between the composition of one year's class of 3rd graders and the next year's class of 3rd graders can cause changes in aggregate test score results.

Linn and Haug, in a paper entitled "Stability of School Building Accountability Scores and Gains," found that:

> *A number of states have school-building accountability systems that rely on comparisons of achievement from one year to the next. Improvement of the performance of schools is judged by changes in the achievement of successive groups of students. Year-to-year changes in scores for successive groups of students have a great deal of volatility. The uncertainty in the scores is the result of measurement and sampling error and nonpersistent factors that affect scores in one year but not the next. The level of uncertainty was investigated using fourth grade reading results for schools in Colorado for four years of administration of the Colorado Student Assessment Program. It was found that the year-to-year changes are quite unstable, resulting in a near zero correlation of the school gains from years one to two with those from years three to four.* [4]

The current API provides a misleading picture of academic progress within schools and districts. It does not measure growth in individuals over time; it just compares overall achievement of successive classes over time.

STAR and the API could be modified to follow the same class from year to year. That is, to compare the scores of the 3rd grade class of Kennedy Elementary School in 2000 to the scores of Kennedy 4th graders in 2001. But this approach, sometimes referred to as "quasi-longitudinal," also has serious limitations. For example, given high levels of student mobility, many members of last year's 3rd grade class may have transferred out of Kennedy. Many other students may have transferred in during the year, further affecting the mix of students from grade to grade.

A true longitudinal data system would be able to track students from year to year regardless of whether or not they re-enrolled in Kennedy Elementary School. Unless they transferred out of state or dropped out of school, such a system would be able to measure

[4] Linn, Robert and Haug, Carol; *Educational Evaluation and Policy Analysis*, Spring 2002, Vol. 24, No. 1, pp. 29–36.

their progress from 3rd grade to 4th grade regardless of where they went, thus providing a more accurate portrait of student learning.

Los Angeles Unified School District recently conducted an actual longitudinal study of its students' learning, tracking individual students' progress in 2000 and 2001. The individual gains were then aggregated up to the school level to show school-level gain scores. The following table of reading score gains shows instances in which cross-sectional measures and longitudinal measures go in opposite directions.

Figure 2: Reading Score Gains at Several LAUSD Schools, 2000-2001

School	Cross-sectional gain [5]	Longitudinal gain [6]
Clifford Elementary	-15	0.6
Clover Elementary	-10	2.2
Kittridge Elementary	4	-2.2
Lanai Elementary	27	-0.3

Lanai Elementary is apparently getting many more 3rd grade students over the 50th percentile, but last year's 3rd grade students have not gained ground as 4th grade students, and may have even slipped slightly. Lanai needs to look at the matched longitudinal scores to see which students are gaining ground and which are slipping.

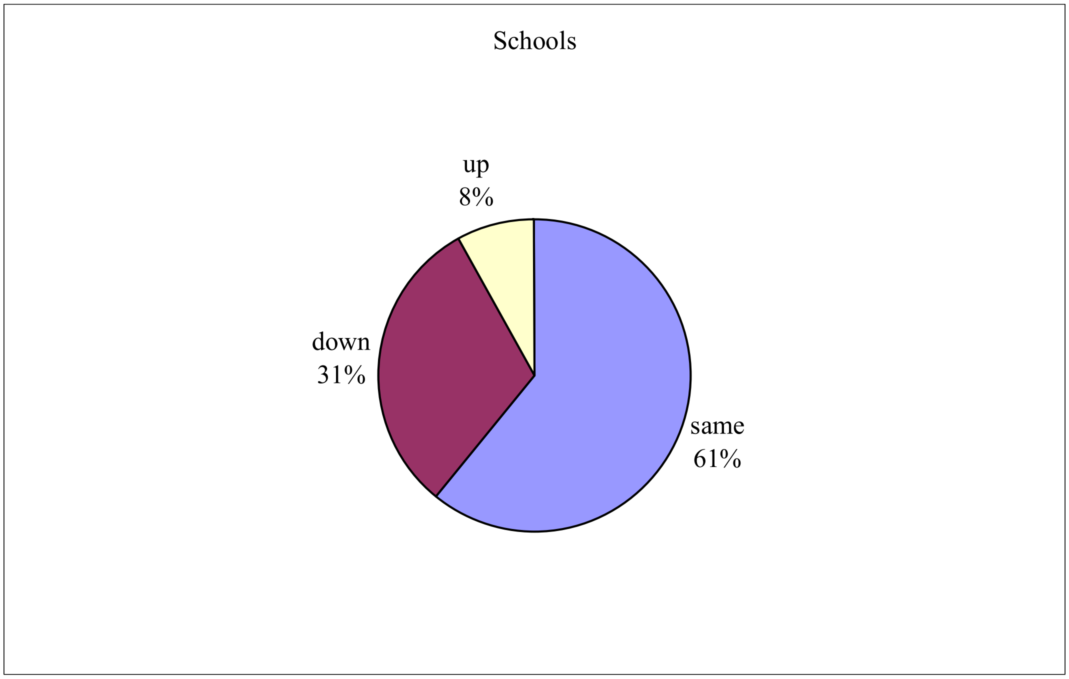
The cross-sectional data are from the state STAR reports, comparing 2000 3rd graders to the 2001 3rd graders. The longitudinal data are the Stanford 9 mean reading NCE gain from 3rd grade students in 2000 to those same students in 4th grade in 2001, from the Los Angeles Unified School District's *2000-2001 Stanford 9 Matched Individual Student NCE Gains* report of October 2001.

Examining STAR reading scores from 2000 to 2001, we performed a simple count of the number of times cross-sectional data and longitudinal analysis yielded contradictory information about the direction of change within a school. In how many schools did the cross-sectional API show an increase in achievement when the longitudinal analysis showed a decrease, and vice versa? Figure 3 below displays the results of this analysis.

[5] Gains in percentage of students at or above the 50th percentile.

[6] NCE gains.

Figure 3: Percent agreement on school progress between the California State API and longitudinal data from LAUSD.



These data show that the API, based on solely quasi-longitudinal analyses, yields a reverse conclusion about school progress about 40 percent of the time compared to an actual longitudinal analyses. Of the 406 schools in LAUSD, about 8 percent showed increases in their achievement when examined longitudinally, while the State API had their scores declining. About 30 percent had longitudinal declines and API increases. (This simple analysis is not intended to explain why there are so many more disagreements in declining scores than in improving scores.)

A combination of the two types of analysis would provide a more stable and accurate measure of school progress. While the longitudinal analysis is more precise, the cross-sectional comparison includes mobile students and compares, however bluntly, achievement in a particular grade-level program from year to year. Local planners could certainly make more informed decisions if they knew how this year's 3rd grade students compared to last year's 3rd graders *and* how last year's 3rd graders are doing this year in 4th grade.

Longitudinal analysis can be approximated by comparing this year's 4th graders to last year's 3rd graders. This quasi-longitudinal method fails to account for the fact that only some of the students are the same from year to year. Researchers have shown a correlation of just .36 when comparing true longitudinal analysis to quasi-longitudinal analysis of school achievement (Dyer, 1969).[7] Dyer also found that the correlation between true longitudinal data and successive cohort data was actually negative: –0.13.

2. How well are state programs and policies working?

California has implemented a number of reforms over the last few years in an effort to improve the overall quality of education in the K-16 system. These include:

- Class size reduction
- Adoption of standards, instructional programs, and tests aligned to one another
- The end of social promotion
- Implementation of high-stakes high school graduation exams
- New investments in teacher professional development
- Modifications to bilingual education
- The end of affirmative action
- A major expansion of outreach activities by the University of California to improve the future UC eligibility of educationally disadvantaged students.

It is essential that the State knows how well these reforms are working—not only how well they are being implemented, but also what impact they are having on student outcomes. Currently, when the State conducts evaluations of these reforms, the evaluators must supplement the information that can be gleaned from CBEDS or the higher education data systems in order to adequately assess them impact of the reforms.

For example, a few years ago a group of research firms formed a consortium to evaluate the class size reduction reform in California (SB 1777). The consortium had to supplement CBEDS data by sending surveys to a representative sample of district superintendents, school principals, and classroom teachers in grades 1 through 4. The surveys contained questions about the program implementation, resource usage, district and school administration, and classroom practices. Answers to the survey provided valuable information on the impacts of class size reduction, but it was relatively costly

[7] Dyer, H.; Linn, R.L.; Patton, M.J.; American Educational Research Journal 6(4): 591-605.

and it was a unique data collection—a "one shot" effort at measuring the impact of the reform. If an electronic student record system were in place that could link all of the systems together, much of the data collected in this one-time survey could be collected on a routine basis. A one-time evaluation of the impact of class size reduction could be transformed into a continuous evaluation of this important reform.

Another example of the need for comprehensive education data is shown in the current evaluation of the California Professional Development Institutes. Assembly Bill 2881 in 2000 expanded the state's current system of teacher professional development centers. In so doing, the legislation required the University of California's Office of the President (UCOP) to submit annual evaluation reports to the legislature beginning January 1, 2002. The evaluations ask three questions:

1. Did teachers increase their content knowledge in their field?
2. Did students in the classes or schools served by teachers participating in the institutes demonstrate increased achievement in the relevant field?
3. For whom and under what conditions are the answers to questions 1 and 2 related?

Answering these questions necessitates student achievement data linked over time and linked to their teachers. Since such data did not exist, the evaluators had to go to the districts themselves for the data—a costly and cumbersome procedure. Going to 156 different districts meant dealing with 156 different data systems and data procedures, with each district providing unique obstacles to collecting and reporting the appropriate data.[8]

No system would be able to anticipate all of the data needs for all evaluation. But a system such as the one we propose would provide the basis for supplying a great deal of common data to evaluators, cutting down on costs to the state and burden to the schools and districts.

What factors promote or impede student learning?

An effective educational data system would also be able to support research into the reasons why some students succeed and others do not. For example, what enables some students to succeed in school despite impediments to their progress? What enables some schools to promote student achievement above all expectations?

[8] Dorf, Rena, UCOP, personal communication.

Currently, researchers and policymakers are forced to collect their own data or rely on national data to make judgments about the relationships among student characteristics, school characteristics, and student achievement. Conducting a separate data collection for each study is costly, and the resulting data samples are not always representative of the population under study. Relying on national data (such as the data produced by the National Center for Education Statistics) does not yield reliable conclusions about California's students since the data does not usually support separate state analyses. Furthermore, the national data does not always use samples large enough to support studies of important subgroups such as language-minority students or students with learning disabilities.

Furthermore, the existing data cannot be linked together to provide a research base. While the STRS system provides teacher data and the STAR system provides student data, the two cannot be linked to show whether there are characteristics of teachers that are particularly effective with particular types of students. The state has amassed a great deal of data in various systems, but it cannot be accessed in ways that will yield answers to the most important research questions.

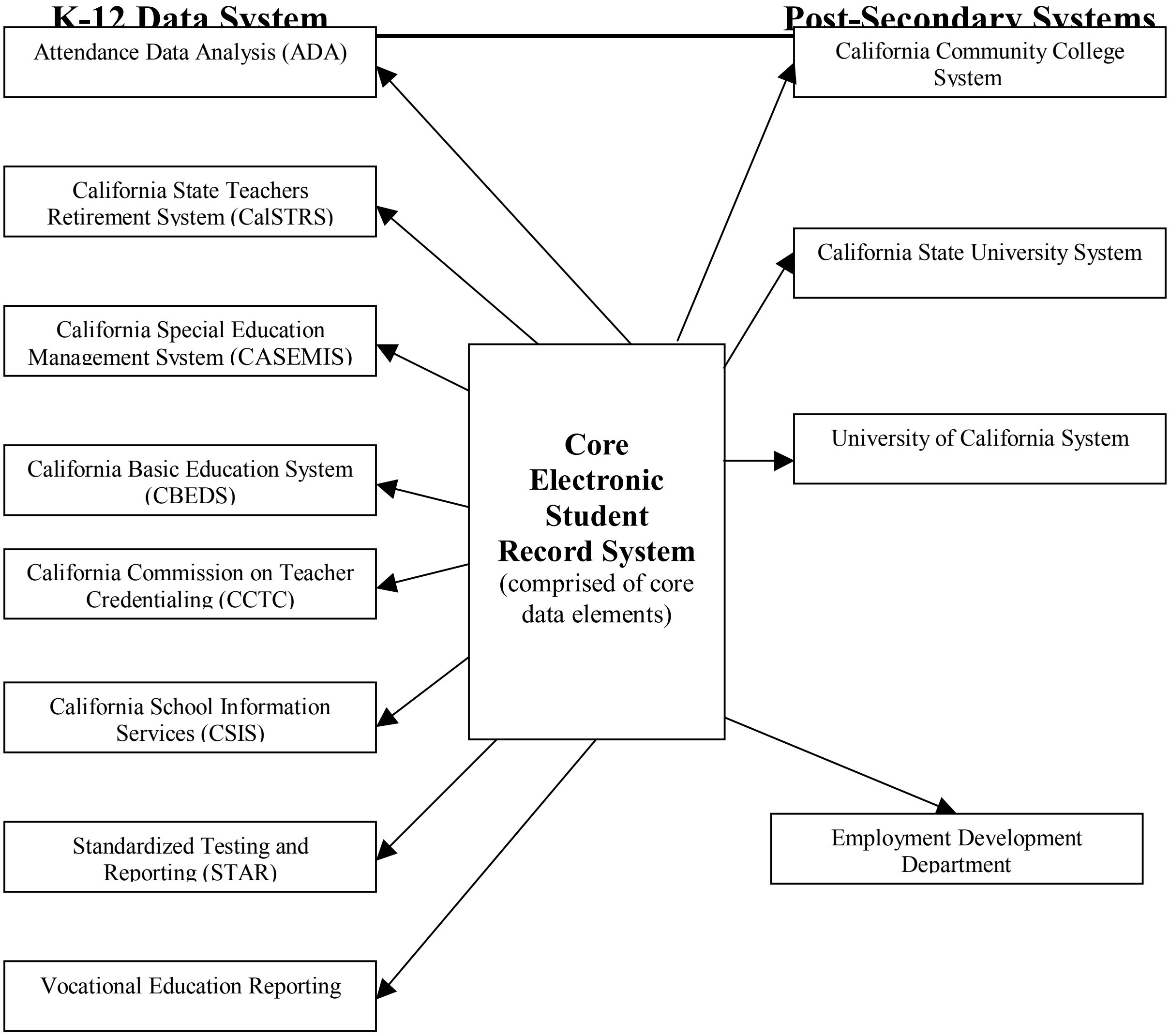
A data system connected by a common student identifier could be used to conduct studies on educational effectiveness. Samples drawn from the full universe file would become the basis for these studies. For example, samples of students, teachers, schools, and districts could be constructed so that the data could match the needs of the individual study. Since much of the data would already be part of the ongoing data collection system, once the samples were drawn, very little new data would have to be collected.

III. A proposal for a new integrated data system

California needs a student-level data system that tracks student achievement over time. Such a system would use a common identification number for students that would follow them through their educational careers. This system would be focused on a limited core of data elements collected at the individual student level. Among the core elements might be enrollment status, achievement on state tests, gender, race/ethnicity, grade, school attended, and other transcript elements of priority importance in state policy development. A common identification number could then link these core elements to other elements within related databases of the current system. The various data systems currently in place would be linked via this new student-based record system, providing a rich and powerful tool for school accountability and program improvement. Figure 4 below depicts the structure of such a system.

Figure 4

Proposed State Longitudinal Data System


K-12 Data System
Post-Secondary Systems
Attendance Data Analysis (ADA)
California State Teachers Retirement System (CalSTRS)
California Special Education Management System (CASEMIS)
California Basic Education System (CBEDS)
California Commission on Teacher Credentialing (CCTC)
California School Information Services (CSIS)
Standardized Testing and Reporting (STAR)
Vocational Education Reporting
Core Electronic Student Record System (comprised of core data elements)
California Community College System
California State University System
University of California System
Employment Development Department

The core of the system would be a very limited set of data elements such as race, gender, age, and the like. These common elements would be linkable to all of the other data systems by a common identification number. As a student moved from school to school—or into college or job training—their records would follow them. A crucial component of this system would be that it would be capable of reporting gains in achievement for each student, which would then be aggregated up to school mean gains. For example, the 4th grade students who were in school A last year would be tracked through the year. Their achievement gains from last year to this year would then be aggregated to the school level to represent a school growth indicator of achievement gains from the 4th to the 5th grades.[9]

Such a system has other advantages. Cohorts of students could be followed throughout their educational careers to show the impact of current reform efforts. True gains in learning could be used to access impacts rather than just status indicators. For the first time, the state could follow the progress of the numbers of highly mobile students. The University of California and the State University system would be able to anticipate the future needs of their student populations as they track the progress of young cohorts of students who will eventually become eligible for admission. The community colleges could evaluate the work-related outcomes that are so important to their mission—the benefits that vocational programs provide for young adults. The impact of various programs on "nontraditional" students could be assessed. In brief, a vast array of tools would be available to the state to provide continuous feedback to the whole K-16 system.

The advantage of linkability to student achievement in the proposed system compared to the current system is summarized in Figure 5 below.

[9] Technical details that would have to be considered include how long a student would need to be enrolled in a school to have his or her score assigned to that school.

Figure 5: Comparison of the Current Data System to the Proposed Data System

DATA SYSTEM	CONTENTS	CURRENT SYSTEM LINKABLE TO STUDENT OUTCOMES?	PROPOSED SYSTEM LINKABLE TO STUDENT OUTCOMES?
Core Electronic Student Record System	Multi-level reporting that integrates core data elements for educational analysis.	**NO**	**YES**
Attendance Data Analysis (ADA)[10]	Attendance data for all the schools in a local education authority gathered together at one point and analyzed in various ways according to local requirements.	**NO**	**YES**
California State Teachers Retirement System (CalSTRS)[11]	Teacher data on retirement and benefits of teachers within the K-12 education system.	**NO**	**NO**
California Special Education Management System (CASEMIS)[12]	Student-level data on over 550,000 students and individual student information regarding participation in special education programs.	**NO**	**YES**
California Basic Educational Data System (CBEDS)[13]	Collects school level data school course from California public schools (K-12) on issues such as staffing, enrollment, retention rates, and vocational education.	**NO**	**YES**
California Commission on Teacher Credentialing (CCTC)[14]	Teacher data on issues of credentialing, certification and retention of teachers in K-12 system.	**NO**	**NO**
California School Information Services (CSIS)[15]	Exchange of student transcripts between Local Education Agencies and postsecondary institutions.	**NO**	**YES**
Standardized Testing and Reporting (STAR)[16]	School aggregations of student characteristics and outcomes.	**PARTIAL[17]**	**YES**
Vocational Education Reporting	Vocational Education Reporting	**NO**	**YES**
California Community College System[18]	Postsecondary student-level reporting for an institution that consists of 72 locally governed districts, 109 colleges, and numerous off-campus centers.	**NO**	**YES**
California State University System[19]	Post-secondary student-level reporting for an institution of 23 campuses, 388,700 students, and 42,000 faculty and staff.	**NO**	**YES**
University of California System[20]	Postsecondary student-level reporting for a system that consists of 10 campuses and more than 187,000 students.	**NO**	**YES**
Employment Development Department[21]	Population data reporting on rates of employment, employment opportunities, and training services in California.	**NO**	**YES**

Implementing an integrated data system raises several key questions, including:

[10] Attendance Data Analysis (ADA), http://www.becta.org.uk/slict/software/index.

[11] California State Teachers Retirement System (CalSTRS), http://www.strs.ca.gov/.

[12] California Special Education Management System (CASEMIS), http:///www2.sac-co.k12.ca.us/speced/general.

[13] California Basic Educational Data System (CBEDS), http://www.cde.ca.gov/demographics/coord/index.html.

[14] California Commission on Teacher Credentialing (CCTC), http://www.ctc.ca.gov/.

[15] California School Information Services (CSIS), http://www.csis.k12.ca.us.

[16] Standardized Testing and Reporting (STAR), http://star.cde.ca.gov.

[17] Link to student outcomes is partial because STAR cannot be linked to prior achievement data.

[18] California Community College System, http://www.cccco.edu.

[19] California State University, http://www.calstate.edu/.

[20] University of California System, http://www.universityofcalifornia.edu.

[21] California Employment Development Department, http:// www.edd.ca.gov.

- What should be used as a common identifier for tracking students?
- How can data confidentiality be ensured?
- Who should get access to the data?
- How should the interchange of data between the different components of the system be coordinated?

What should be used as a common identifier?

A single common identifier is critical to the success of a linked system. Without one, linking the various datasets would have to use some sort of matching technique based on combinations of name and demographic characteristics. There are various ways in which a common identifier can be constructed. Barbara Clements of Evaluation Software Publishing, Inc. (and formerly with the Council of Chief State School Officers) is the nation's foremost authority on this issue. She has found that generally there are four options to designing and implementing a student identifier:[22]

1. **Local Approach.** Locally assigned numbers can be used in combination with preexisting state identifiers for the school. The main disadvantage of this approach is that students are likely to be assigned different numbers as they move from district to district, thus making longitudinal analysis difficult.
2. **Algorithm-Assigned Identifiers.** This is essentially how CSIS currently assigns identification numbers; it uses the Soundex encoding system for names and birthdays. The main disadvantages of this option are that a person might give different names or birth dates when entering different parts of the education system or the workforce, and that the algorithm changes when a person's name changes. These weaknesses would be particularly critical when attempting to track students into college and/or the workforce.
3. **Social Security Numbers.** SSNs are unique and assigned nationwide; almost all students enter school with a SSN already assigned, and the University of California already uses students' SSNs for identification. The main disadvantage of using SSNs relate to the privacy concerns that many Californians have expressed. However, several states currently use SSNs as their systemwide identifier and have been able to address the privacy concerns successfully.
4. **Hybrid of 2 and 3.** Each student is assigned a unique identifier based on the student's SSN. The identifier, which stays with the student's data from year to year, cannot be

traced back to the SSN or to the student's identity, except deep inside the most secure vaults of the data system. Florida uses this approach, employing an encryption formula to scramble the SSN and then using this scrambled identifier as the common link within each system. At the most secure core of the system is a link between the common identifier and the original SSN.

Such a hybrid appears to be the best way in which data can be shared among systems. If different random identification numbers were used within each system, trying to link student data across systems—for example, between the community college system and the employment insurance data (which already use SSN)—would be extremely difficult and prone to error. Only options 3 and 4 allow linking across systems.

Legislative Implication: Establish Common Identifier

Establish a longitudinally sustainable student identifier unique to each student. Stipulate that the identifier be based on Social Security Number encrypted deep within the most secure area of the data system.

How can data confidentiality be ensured?

Students and their families value their privacy. States that keep education data on individual students owe it to the public to keep that data confidential. The essence of this responsibility is to keep the link between data about a student and the identity of the student confidential. It is the identity, not the data, that requires the utmost protection—because Social Security Numbers identify individuals, they must be strictly protected.

The Family Educational Rights and Privacy Act (FERPA), passed by the United States Congress in 1974, provides for the privacy and confidentiality of student education data. Fortunately, there has been a great deal of work done in this area in other states and with the federal data collections.

Barbara Clements, one of the nation's leading experts in this area, has worked for years on this issue and has helped several states develop confidentiality safeguards for

[22] In her paper *Designing and Implementing a System for Assigning Student Identifiers in New York*, Dr. Clements and her colleague Glynn Ligon go into great detail on the advantages and disadvantages of each

their education data records. She has developed a list of specific recommendations to ensure that education data within the state data system remains secure.[23] Clements argues that individual student records should be accessed by state staff only on a "need to know" basis. Furthermore, if the data is to be used as an analytical database and accessed at the individual level, personal identifiers should be stripped from the file before it is released.

The federal government also has guidelines for keeping data secure. For example, in collecting data on individuals, the National Center for Education Statistics (NCES) has a variety of procedures in place to ensure the confidentiality of those records. (These standards are detailed in *NCES Statistical Standards,* available at http://www.nces.ed.gov/pubsearch/pubsinfo.asp?pubid=92021.) The standards include the following:

- All contractors (both data collectors and analysts) with NCES must submit a list of staff members who will have access to confidential data. These staff members must take an oath of confidentiality confirmed by a notarized affidavit of nondisclosure.
- Researchers outside of the federal government who want access to the data also must take the confidentiality oath and must demonstrate that they have specific procedures in place to protect the privacy of individual data.
- Violations of confidentiality are punishable by a fine of $250,000 or five years in jail.
- In reports, care is taken to ensure that individuals cannot be identified in tables. For example, cells within a table must contain at least three cases.
- Before releasing the dataset, the data must be subjected to a disclosure analysis to ensure that individuals cannot be identified by outside analysis using any statistical routine. A Disclosure Review Board reviews each dataset before it is released.

Legislative Implication: Protect Data Confidentiality

Charge the entity responsible for the student data system with protecting students data. In carrying out its responsibilities to provide analyzable data files to legitimate analysts, including student data linked longitudinally or linked to other files by means of student identifiers, the entity should strip or encrypt identifiers traceable to the student's identity.

of these options.

Who should get access to the data?

In a system that allows only a small group of state employees to access the core data elements and links to the other components in the system, a few standard reports, much like the API and STAR reports that are now produced, could be produced with greater accuracy. But such a system wouldn't increase the public's ability to understand how well our schools are doing. An expanded range of analysts and researchers would be able to produce answers to a broader ranges of questions.

The user set should be expanded to a larger set of state employees .For example, the UCOP may want access to the data to do a special study on the progress of students through high school into the community college system and then transferring into the UC system. The workforce development program might want to do special analyses of the outcomes of training programs on job placement and earnings. Or high school vocational coordinators may want to explore the pathways of vocational students through the educational system and out into the world of work.

Finally, the circle of users can be expanded more widely to include outside researchers. This would allow the data to be used for independent research purposes that would inform policymakers on the root causes of educational inequity in the state. Allowing access to the data to outside researchers would also serve as an independent audit system for the whole state data system. This has been done in other states, including Texas and North Carolina, where independent researchers are able to use state collected data to affirm or contradict the published results of state reports. In those states, this has facilitated a healthy debate about the educational system and how to improve it. It has helped inform both policymakers and the public about the complexities of the issues of educational reform and has increased the faith the public has in the data.

All of this could be handled with the proper confidentiality controls in place. Other states such as Texas and North Carolina have established procedures that protect the privacy of individuals while increasing the states' ability to analyze important aspects of the educational system.

[23] Clements, Barbara (1997); *Protecting the Confidentiality of Education Records in State Databases*, available at http://www.educationadvisor.com/ocio2001/Confidpaper.doc.

Legislative Implication: Designate User Access

Legislation should explicitly list categories of users and uses for the data system, assign responsibility for approving and servicing users and charge the responsible entity to establish and post protocols, criteria, and procedures for use. Approved users should include state agencies with responsibility for education, legislative policy analysts, evaluators of public school programs, and education researchers from established research organizations.

How would the interchange of data between the different components of the system be coordinated?

Coordinating data interchange will be an enormous undertaking under the new system. Clearly, the current system of ad hoc and informal data coordination committees is woefully inadequate to provide authoritative data definitions and protocols for the new system. A greatly expanded CSIS might be the first option to consider. However, CSIS is currently just a voluntary system and does not track students by Social Security Number, nor does it contain student achievement data. Also, if the state decides to give access to data to other than a small group of government employees, substantial resources would be required to perform the necessary data matching and confidentiality checks, and to disseminate the data to users.

A single agency should have the authority to negotiate with client agencies that maintain their own data systems (for example, CBEDS or the community colleges) to come up with common definitions of data elements, protect privacy, maintain accuracy, and to help coordinate data merges. Such a complex system could not be designed, let alone maintained, through cooperative agreements alone. Someone has to be in charge.

Legislative Implication: Designate Responsible Entity

A responsible entity should be designated to establish the core student data system. This entity would be responsible for establishing the student identifier, ensuring confidentiality, producing analyzable files for approved users, and performing regular data audits, in addition to other matters necessary for accomplishing the goals of the legislation. The entity would coordinate the linking to student data with other agencies and users.

IV. Other States' Data Systems

Our proposed approach resembles what many states are already pursuing (some of which pursued it even before the passage of the *No Child Left Behind* legislation). Brian Stecher and his colleagues at Rand recently surveyed those states to determine what they were doing with their student record systems. The following table summarizes the results.

Figure 6: States that maintain or plan to maintain individually identifiable student test data at the state level.

State	Maintains unique student IDs?	If yes, what are they?	Used consistently across years?	Linkable to student background info?
Alaska^	Developing a system	NA	NA	NA
Indiana	Developing a system	NA	NA	NA
Massachusetts^*	Developing a system	NA	NA	NA
Nebraska	Developing a system	NA	NA	NA
Nevada^*	Developing a system	NA	NA	NA
New York^*	Developing a system	NA	NA	NA
North Carolina^*	Developing a system	NA	NA	NA
Alabama^*	Yes	SSN	Yes	Yes
Arkansas*	Yes	SSN or state-assigned ID	Yes	Yes
Delaware*	Yes	State-assigned ID	Yes	Yes
Florida^*	Yes	SSN or state-assigned ID	Yes	Yes
Georgia^	Yes	SSN	Future	Future
Louisiana^*	Yes	State-assigned id	Yes	Yes
Minnesota^	Yes	State-assigned id	Yes	Yes
Mississippi^	Yes	State-assigned id	Future	Future
Oregon	Yes	State-assigned id	Yes	Future
Texas^*	Yes	SSN	Yes	Yes
Vermont*	Yes	State-assigned id	Yes	Yes
Arizona	Yes, partial	Composite	Yes	Yes
Connecticut*	Yes, partial	Composite	Yes	Yes
Maryland^*	Yes, partial	Composite	Yes	Yes
Tennessee^*	Yes, partial	Composite	Yes	Yes

^ States that have an exit exam in place

*States that reward and/or sanction schools based on test scores

Clearly, many states have either already developed or are working on developing a system of student data collection similar to the one we are recommending in this paper. While this option is not without cost, California should be in the forefront of this effort.

The largest state system of education in the nation should be leading the nation both in educational reform in the classroom and in the assessment of the impact of those reforms. Without such longitudinal data, the progress of students in California's schools will remain difficult to assess, the true impact of reforms will remain unknown, and impediments to educational progress will remain hidden.

Conclusion

Virtually every major education program in California ultimately has the same clear, fundamental goal: to improve student learning. The fragmented data system currently in place, however, cannot provide equally clear answers to questions about how well that goal is being met. Californians are unable to accurately track the return on their large investments in education. Researchers and policymakers are forced, at great expense, to construct procedures that can provide sufficient data about how well students are learning. Schools and districts are forced to make "best guess" decisions about the success or failure of various programs.

The solution is to implement a centralized data system that provides data to track students' real progress, rather than approximating their achievement using inaccurate averages. State education agencies, legislative policy analysts, evaluators of public school programs, and education researchers would all be able to access data that truly reflects student progress. Other states have proven that a data system can accurately track student progress without compromising their right to privacy. Every dollar spent on an improved data system would maximize the effectiveness of existing investments and help California make smarter investments in the future.

References

Blackford, L. (2002). "Failing" Schools Report Criticized. *Herald Leader*. Lexington.

Carlson, D. (2002). Four Methods of Judging School Quality and Improvement: Relationships and Implications, Center for the Study of Evaluation, National Center for Research on Evaluation, Standards and Student Testing, UCLA.

Clements, Barbara, and Ligon, Glenn. *Designing and Implementing a System for Assigning Student Identifiers in New York.* [information TK]

Dyer, H.; Linn, R.L.; Patton, M.J. (1969). "A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests." *American Educational Research Journal* **6**(4): 591-605.

Kaufman, P. (2002). "The Feasibility of Developing a California Education Longitudinal Study." Linguistic Minority Research Institute, University of California.

Linn, Robert L., and Haug, Carolyn (2002). "Stability of School Building Accountability Scores and Gains." *Educational Evaluation and Policy Analysis* **24**(1): 29-36.

Meyer, R. H. (1995). "Educational Performance Indicators: a Critique." Institute for Research on Poverty.

Stephens, S. (2002). "New Law May Leave School Behind." *The Plain Dealer,* Cleveland.