



EDUCATION

- THE ARTS
- CHILD POLICY
- CIVIL JUSTICE
- EDUCATION
- ENERGY AND ENVIRONMENT
- HEALTH AND HEALTH CARE
- INTERNATIONAL AFFAIRS
- NATIONAL SECURITY
- POPULATION AND AGING
- PUBLIC SAFETY
- SCIENCE AND TECHNOLOGY
- SUBSTANCE ABUSE
- TERRORISM AND HOMELAND SECURITY
- TRANSPORTATION AND INFRASTRUCTURE
- WORKFORCE AND WORKPLACE

This PDF document was made available from www.rand.org as a public service of the RAND Corporation.

[Jump down to document](#) ▼

The RAND Corporation is a nonprofit research organization providing objective analysis and effective solutions that address the challenges facing the public and private sectors around the world.

Support RAND

[Purchase this document](#)

[Browse Books & Publications](#)

[Make a charitable contribution](#)

For More Information

Visit RAND at www.rand.org

Explore [RAND Education](#)

View [document details](#)

Limited Electronic Distribution Rights

This document and trademark(s) contained herein are protected by law as indicated in a notice appearing later in this work. This electronic representation of RAND intellectual property is provided for non-commercial use only. Unauthorized posting of RAND PDFs to a non-RAND Web site is prohibited. RAND PDFs are protected under copyright law. Permission is required from RAND to reproduce, or reuse in another form, any of our research documents for commercial use. For information on reprint and linking permissions, please see [RAND Permissions](#).

This product is part of the RAND Corporation monograph series. RAND monographs present major research findings that address the challenges facing the public and private sectors. All RAND monographs undergo rigorous peer review to ensure high standards for research quality and objectivity.

Toward a K–20 Student Unit Record Data System for California

Georges Vernez, Cathy Krop, Mirka Vuollo, Janet S. Hansen

Prepared for The William and Flora Hewlett Foundation

Distribution unlimited



RAND

EDUCATION

The research described in this report was conducted within RAND Education and supported by The William and Flora Hewlett Foundation.

Library of Congress Cataloging-in-Publication Data is available for this publication.

ISBN 978-0-8330-4206-4

The RAND Corporation is a nonprofit research organization providing objective analysis and effective solutions that address the challenges facing the public and private sectors around the world. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

RAND® is a registered trademark.

© Copyright 2008 RAND Corporation

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from RAND.

Published 2008 by the RAND Corporation

1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138

1200 South Hayes Street, Arlington, VA 22202-5050

4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665

RAND URL: <http://www.rand.org>

To order RAND documents or to obtain additional information, contact

Distribution Services: Telephone: (310) 451-7002;

Fax: (310) 451-6915; Email: order@rand.org

Preface

Education, more than ever, is key to a person's lifetime economic prospects as well as to making the most of one's talent and interests in a world that is rapidly changing economically and politically. Policy-makers and educators are working to align K–12 and postsecondary standards, assessments, and accountability policies to discourage student dropout, encourage smoother transition from secondary to postsecondary education, and encourage student retention in college. To take steps that will achieve the goal of improving student progression and quality, states need accurate information on student enrollment and retention, the effectiveness of programs, and factors that may affect how students move through the education system. To this end, they are developing robust data systems that are commonly termed “student unit record” (SUR) systems because they contain individual electronic records of each student enrolled in an educational institution. SUR data systems permit the tracking of an individual student's progress over time—from entry in kindergarten to exit from college and eventually into the labor market as well—to answer questions that are at the core of educational effectiveness.

California is lagging behind other states in developing such a student data system. Currently, 18 states can match student individual records from K–12 and postsecondary education systems and several other states are currently developing this capability. California is not one of them.

We undertook this study (1) to document the state of the various student data systems available for the four California education segments—K–12 public schools, community colleges, California State

University, and the University of California; (2) to assess the feasibility of and identify the challenges to and opportunities for developing a California integrated SUR data system; and (3) to identify steps that could be taken toward building and maintaining an integrated SUR system for California.

This project was conducted by RAND Education, a unit of the RAND Corporation, and was sponsored by The William and Flora Hewlett Foundation. The findings of this study should be of interest to legislators, education administrators, teachers, and faculty and all those in the general public who believe in data-based decisionmaking and seek ways to improve the effectiveness of education.

The principal author of this work may be contacted by email at vernez@rand.org or by phone at 310-393-0411, extension 6211. For more information on RAND Education, contact the Director, Susan Bodilly. She can be reached by email at Susan_Bodilly@rand.org; by phone at 703-413-1100, extension 5377; or by mail at RAND Corporation, 1200 South Hayes St., Arlington, VA 22202-5050. More information about RAND is available at www.rand.org.

Contents

Preface	iii
Figure and Tables	ix
Summary	xi
Acknowledgments	xix
Abbreviations	xxi

CHAPTER ONE

Introduction	1
Interest in K–20 Data Systems Is Growing in the United States	2
Many States Have Made Considerable Progress in Building a K–20 Data System	4
California Has Many Similar Reasons for Creating a K–20 Data System	5
California Currently Lags Behind Other States	7
Differences Between California and Other States Might Make Creating a K–20 System Challenging	9
Focus of This Study	10
Study Approach	12
Study Limitations	15
Organization of the Report	16

CHAPTER TWO

California’s Current Student Data Systems	17
Individual Student Databases for K–12 Students	18
California School Information Services	18
California Longitudinal Pupil Achievement Data System (CALPADS)	21

Individual SUR Databases for Postsecondary Students	23
California Community Colleges	23
California State University	26
University of California	28
Partially Integrated SUR Databases	31
California Postsecondary Education Commission	31
Fully Integrated SUR Databases	34
California Partnership for Achieving Student Success	34
Summary	38

CHAPTER THREE

Major Challenges and System Design Issues	41
Major Challenges	41
Overcoming the Culture of Independence and a Protective Mindset Among the Stakeholders	41
Gaining Support and a Consensus from Stakeholders for a K–20 SUR File	42
Assigning Responsibility for Operating the K–20 Data System	45
Assigning Decisionmaking Authority for the K–20 Data System	48
Costs	49
System Design and Technical Issues	51
Technical Feasibility	51
Incompleteness and Uneven Quality of the Student Data	52
Coordination of Data Submission by the Segments	55
A Common Student Identifier	55
Determining Who Should Have Access to the Data	58
Student Privacy	61
Summary	62

CHAPTER FOUR

What Next for California?	63
Complete the Design and Implementation of CALPADS	63
Identify a Champion to Be an Advocate for a K–20 Data System in California	64
Obtain Legislative Authority	64
Begin Building the K–20 Student Data System Incrementally	65

Develop an “Objective” Analytical Capability and Expertise 69
 Is There a Better Alternative? 70

APPENDIX

A. Selected Characteristics of States Selected for Interviews 73
B. Illustrative Interview Protocol 77
C. California Student Record Data Systems 87

References 101

Figure and Tables

Figure

2.1. Common California’s Seven SUR Data Systems in Use or Under Development.....	18
---	----

Tables

A.1. Characteristics of Student Data Systems of States Selected for Interviews.....	73
C.1. Common Data Elements and Other Characteristics.....	88
C.2. Other Characteristics of the Data Systems and Their Use.....	95

Summary

To provide the information needed to help expand educational opportunities and improve the quality of education from kindergarten through college, U.S. states are increasingly building robust student data systems. These systems are commonly termed “student unit record” (SUR) systems because they contain individual electronic records of each student enrolled in an educational institution. SUR data systems permit the tracking of an individual student’s progress over time—from entry into elementary school to exit from college and eventually into the labor market. These data are vital to attempts to formulate policy and answer questions at the core of studies of educational effectiveness.

Currently, 18 states are able to match individual student records between their K–12 and postsecondary education systems. Several other states are developing this capability. California, however, lags behind other states. Despite the California legislature’s clear intent that a longitudinal student data management system able to track students from elementary school through college be developed and maintained, the state is, at this time, neither planning nor implementing such a system. A K–20 SUR system would enable the state, education administrators, faculty, and the public to address questions that currently cannot be addressed with current data systems—questions such as how California’s students are progressing over time; how course articulation between high school and college can be improved; how effective are programs to retain, prepare, and help students succeed in college; what classes of students may need special intervention and attention; and how well are students prepared to meet California’s future labor demands.

At the request of The William and Flora Hewlett Foundation, the RAND Corporation undertook this study to identify the issues and challenges involved in developing an integrated K–20 SUR data system for California. To this end, we

- reviewed and documented the characteristics of the SUR data systems currently in use or under development in all four segments of California education: K–12 public schools, community colleges, the California State University, and the University of California
- interviewed the administrators of SUR systems in California and other states
- reviewed the literature on longitudinal SUR systems and the experiences of various states with these systems
- interviewed key state-level stakeholders, leaders of education segments in California, and researchers.

California’s Current Student Data Systems

California currently has seven student data systems either in use or under development:

- The California School Information System (CSIS), a mandated public K–12 system that contains limited enrollment and demographic data on all students and additional data on 60 percent of those students in districts participating voluntarily—currently in use
- The California Longitudinal Pupil Achievement Data System (CALPADS), a mandated public K–12 system—under development
- One system for each of California’s three public higher-education segments, the community colleges, California State University, and the University of California—all currently in use
- The California Post-Secondary Education Commission student database, a partially integrated system that brings together

student-level data from all three higher education segments—currently in use

- The California Partnership for Achieving Student Success (CalPASS), a voluntary and regional system that fully integrates student data from individual institutions in all four of California’s educational segments—currently in use and expanding.

Detailed information on each of the above student data systems is contained in Chapter Two.

Major Challenges and System Design Issues

These seven systems give California the embryonic architecture that would enable it to develop and maintain a comprehensive K–20 longitudinal SUR data system. The state also has the necessary technical expertise. Furthermore, our respondents indicated that there are no technological barriers to developing such a data system.

Still, a number of major challenges will have to be addressed before an integrated K–20 SUR system can become a reality in California. Each of the state’s four education segments has developed its own policies and administrative practices and all have developed strong separate cultures and identities as well as a protective mindset. A consensus among all stakeholders that building such a system is desirable in the first place will have to be developed. Beyond that point, it is also likely to be difficult to develop a consensus on the purpose of the system—who should hold decisionmaking authority over its development and maintenance, who should have access to it, and who should operate it—because this may require that the current overseers of the four education segments give up some decisionmaking authority. Given California’s budget woes and its experience with cost overruns when developing some previous state data systems, providing the necessary funding may well prove to be a roadblock, even though the funds would be minimal compared with the billions spent on education in the state.

In addition, the state will need to gain a consensus among the four education segments and other stakeholders on various system design issues:

- determining how comprehensive the system should be
- standardizing data elements across segments
- improving the quality of the data collected
- selecting a common student identifier across segments
- protecting student privacy.

But as experience in other states bears witness, all of these design issues are solvable.

What Next for California?

Overcoming the interrelated challenges involved in developing an integrated K–20 SUR data system in California, and eventually one even more comprehensive, may be a daunting task that could take several years. In our view, the state will need to take several steps to resolve these challenges:

1. Complete the Design and Implementation of CALPADS

The first step is to complete the design and implementation of CALPADS. Much remains to be accomplished before the system is operational in 2009–10 as anticipated—including selecting a vendor, which is not expected until fall 2007, and providing the funding to build the capability of Local Education Agencies (LEAs) and schools to collect and transmit student data electronically. Funding for this purpose was deleted from the state’s 2007–08 budget.

2. Identify a Champion to Be an Advocate for a K–20 Data SUR

According to our California respondents, getting the process started will require an influential individual or group of individuals to perform several key functions:

- champion the development of an integrated K–20 SUR data system
- gather the necessary support from stakeholders and the public
- develop a consensus on the system’s governance and design
- sponsor the legislative action that would mandate and define its purpose and capabilities and identify its limitations.

Potential candidates include the governor, a well-connected business leader, or a legislator. Alternatively, a K–20 SUR could be championed by a commission that might include a representation of committed stakeholders and potential users.

3. Obtain Legislative Authority

Even though new legislative action is not needed to develop and maintain a K–20 SUR data system, our California respondents agreed unanimously that such legislation would be required to enable the state to develop an *integrated* K–20 SUR system. Many of the issues involved in developing such a data system are so sensitive and important that they need to be aired, debated, and eventually resolved in an open public debate. Nor is it likely that the disparate views on governance of the data system (i.e., who holds decisionmaking authority and who owns the system and access to the data) could be resolved without legislative authority.

The California legislature would need to vet and address the following questions:

- What purposes would a K–20 SUR data system primarily serve?
- Who should have decisionmaking authority over the design of the system?
- Who should operate the system and where should it be located?
- Who should have access to the data it contains?
- What common student identifier should be used?
- What is the minimal content to include in the data file?
- What level of funding should be allocated to develop, maintain, and use the system?

In this report, we have presented a number of possible answers to these questions—various options that the legislature may consider, some that would lead to a more comprehensive integrated K–20 system and some to a narrower set of linkable individual student databases.

4. Build the K–20 Student Data System Incrementally

It will take several years to fully develop an integrated K–20 data system that has maximal utility for policy and administrative decisionmakers at the state as well as at the segmental and individual institution levels. Consequently, we recommend taking an incremental approach. A possible development sequence over a period of four to five years would be:

- Integrate the four existing segmental data systems “as is.” Even this simple merger of the three current postsecondary student data files with CALPADS would provide useful and more accurate information than is currently available on dropout and transfer rates, student mobility, and educational progress.
- Add data elements already being collected at the individual school and campus levels but not currently forwarded to the central student database of each segment—such as individual courses and grades.
- Link the K–20 student data file to other state and federal data systems—such as preschool, employment and wage, welfare, foster care, corrections, military, and private university.
- Add data elements not currently being collected by one or more the segments—but only if deemed desirable and cost-effective after several years of experience using the data system and after the three steps just outlined have been implemented.

Overall, California should take a long-term perspective toward improving the quality and standardization of the data that schools and campuses collect. In the short term, those responsible for collecting and entering data should receive training to ensure that they use appropriate quality control and privacy protection practices. But experience suggests that such training will go only so far and that the quality

of the data will improve primarily through use and user feedback to schools and campuses.

5. Develop an “Objective” Analytical Capability and Expertise

To make maximum use of the integrated K–20 student data and effectively respond to the needs of individual schools, colleges, universities, and state officials, California will need to take one last step: An analytical capability, adequately funded, should be developed either within or independent of the organization operating the K–20 data system. Not only will this capability enable those assigned this responsibility to develop the information required for policymaking, it would also address one of the principal concerns of our respondents in the state’s higher education segments—that their data might be misinterpreted either because users are unfamiliar with how a segment operates or because the segment has made administrative or policy changes. With an experienced analytical team, the frequency of such problems would be significantly reduced.

Experience with the use of SUR data systems in California and other states suggests that ten to 20 analysts might be required to provide the level of capability and expertise required.

Acknowledgments

We want to thank Pamela Burdman of The William and Flora Hewlett Foundation for providing both financial support and encouragement for this effort. She helped share the research questions and saw the potential long-term benefits of this kind of inquiry.

We also want to thank all the respondents who took time from their busy schedules to meet with us and answer candidly the many questions we asked. We hope that we have faithfully represented their views.

Professor Michael Kirst, Stanford University, and our colleague Catherine Augustine reviewed a draft of this report. We appreciate their comments and substantive suggestions; they helped make the report a better product.

Finally, we want to thank Susan Bohandy for her excellent editorial assistance and Louis Ramirez, who cleaned up this final version of the report with professional efficiency.

Abbreviations

AICCU	Association of Independent California Colleges and Universities
AP	advanced placement
API	academic performance index
CAHSEE	California High School Exit Exam
CALPADS	California Longitudinal Pupil Achievement Data System
Cal-PASS	California Partnership for Achieving Student Success
CALTIDE	California Longitudinal Teacher Integrated Data Education System
CBEDS	California Basic Educational Data System
CCC	California Community Colleges
CCCCO	California Community Colleges Chancellor's Office
CELDT	California English Language Development Test
CPEC	California Postsecondary Education Commission
CDE	California Department of Education
COMIS	California Community Colleges Management Information System
CSIS	California School Information System

CSS	Corporate Student System
CSU	California State University
DGS	Department of General Services
DQC	Data Quality Campaign
EDW	Education Data Warehouse
ELC	Eligibility in the Local Context
EPT	English Placement test
ERS	Enrollment Recording System
FASTER	Florida Automated System for Transferring Educational Records
FCMAT	Fiscal Crisis Management Assistance Team
FERPA	Family Educational Rights and Privacy Act
FPCO	Family Policy Compliance Office
GATE	Gifted and Talented program
GPA	grade point average
ID	identification
IEP	Individualized Education Program
IES	Institute for Education Science
IPED	Integrated Postsecondary Education Data System
IR&C	Information Resources and Communication Office
IT	information technology
K–12	kindergarten, elementary, middle, and high school education system
KIDS	Kentucky Instructional Data System
LAO	Legislative Analyst’s Office

LEA	Local Education Agency
LSDS	Longitudinal Student Data System
MOU	Memorandum of Understanding
NCEA	National Center for Educational Accountability
NCES	National Center for Education Statistics
NCLB	No Child Left Behind Act of 2001
NCHEMS	National Center for Higher Education Management Systems
ODE	Oregon Department of Education
OUS	Oregon University System
SAT	Scholastic Aptitude Test
SES	socioeconomic status
SNOR	Student National Origin Report
SSID	Statewide Student Identification
SSN	Social Security number
STAR	Standardized Testing and Reporting
SUR	student unit record
TEA	Texas Education Agency
THECB	Texas Higher Education Coordinating Board
TOEFL	Test of English as a Foreign Language
TOP	Taxonomy of Programs
TPEIR	Texas Public Education Information Resource
UC	University of California
UCOP	University of California's Office of the President

Introduction

There is growing interest in tracking student progression from elementary to secondary school, into postsecondary institutions, and eventually into the labor force. This interest stems from various concerns about the state of the U.S. education system. The United States has fallen behind other nations in the proportion of young adults with a postsecondary credential. The country also continues to see significant gaps in attainment between white students and students of color. At the same time, questions about student dropout and retention rates and effective ways to improve student progression and achievement remain unanswered.

Many states are working to align K–12 and postsecondary standards, assessments, and accountability policies to encourage smoother student transitions from secondary to postsecondary education at the college levels and across different types of postsecondary institutions (L'Orange and Ewell, 2006).

To take steps that will achieve the goal of improving student progression and quality, states need information on, for example, rates of student enrollment and retention, the effectiveness of programs, and factors that may affect how students move through the education system. Data play a central part in providing the information states require. Over the past decade, state policymakers and educators have become increasingly aware that robust state educational data systems are critical to their ability to understand student attainment and achievement, as well as how to improve them. These data systems are commonly termed “student unit record (SUR)” systems because they contain an individual electronic record of each student enrolled in an

educational institution. SUR data systems permit individual student records to be matched with each other over time and across education segments to answer questions at the core of educational effectiveness.

In recent years, the federal No Child Left Behind (NCLB) Act of 2001, which mandates reporting on student progress on a school-by-school basis, has obliged states to develop capable K–12 SUR data systems to meet this requirement. At the same time, higher education institutions that had developed such data systems decades ago have made them increasingly comprehensive.

These systems alone, however, are not as comprehensive as needed to make data-driven decisions that span an entire education system. A system that links or integrates K–12 data with those of higher education—an integrated longitudinal K–20 system—would be the most effective way to provide the information that state policymakers and education administrators require (L’Orange and Ewell, 2006).

Interest in K–20 Data Systems Is Growing in the United States

Interest in creating longitudinal SUR data systems has evolved considerably in recent years, as it has become ever more feasible in technical terms to develop them. In 2005, a national collaborative effort—the Data Quality Campaign (DQC)—was initiated by a group of foundations, national educational administrators, and research associations with the goal of developing a high-quality longitudinal data system in every state by 2009.¹ One campaign goal is to support the creation of data systems that will match students’ P–12 (pre-kindergarten through high school) records with their postsecondary records.

¹ The DQC is supported by the Bill & Melinda Gates Foundation, is managed by the National Center for Educational Accountability, and includes as founding partners Achieve, Inc., the Alliance for Excellent Education, the Council of Chief State School Officers, the Education Trust, the National Center for Educational Accountability, the National Center for Higher Education Management Systems, the National Governors Association Center for Best Practices, the Schools Interoperability Framework Association, Standard & Poor’s School Evaluation Services, and State Higher Education Executive Officers.

The Data Quality Campaign identifies ten essential elements of a state longitudinal data system:²

- a unique statewide student identifier
- student-level enrollment, demographic, and program participation information
- the ability to match an individual student's test records from year to year to measure academic growth
- information on untested students
- a teacher identifier system with the ability to match teachers to students by course taken
- student-level transcript information, including information on courses completed and grades earned
- student-level college admission test scores (e.g., SAT, ACT scores)
- student-level graduation and dropout data
- the ability to match student records from pre-kindergarten through postsecondary systems
- a state data audit system assessing data quality, validity, and reliability.

The campaign suggests that these ten elements are necessary but not sufficient for a robust longitudinal data system. It also emphasizes the need to link a state longitudinal system with employment and other data systems, to transfer records across states, and to protect student privacy.³

There is also growing interest within the federal government and among national organizations in developing statewide, and conceivably national, longitudinal SUR data systems. The federal government recently launched a program, the Statewide Longitudinal Data Systems Grant Program, to support the development of state student K–12 data systems. In November 2005, the U.S. Department of Education

² For details, see www.DataQualityCampaign.org (accessed on June 24, 2007).

³ Although states make fuller use of student data, students' privacy rights and the rights of parents to protect their children need to be preserved, as set out in the federal Family Educational Rights and Privacy Act (FERPA).

Institute for Education Sciences (IES) awarded three-year grants to 14 states (out of 45 applicants) to help them design, implement, or upgrade their longitudinal student data systems.⁴

National organizations have examined moving beyond individual state SUR systems to developing a federal system.⁵ Such a system would provide state SUR systems with a mechanism both for gaining information on students who transfer into their education systems from another state and for following students beyond their own borders. Because many students move from one state to another at some point during their educational career, accurately capturing student progression requires a national perspective.⁶ The National Center for Education Statistics (NCES) determined that it was feasible to develop a national SUR system,⁷ and recommendations to do so were featured in reports of both the National Commission on Accountability in Higher Education (2005) and the Secretary of Education's Commission on the Future of Higher Education (2006).

Many States Have Made Considerable Progress in Building a K–20 Data System

According to the 2006 National Center for Educational Accountability (NCEA) survey of states⁸ conducted for the Data Quality Campaign, only Florida currently has all ten essential elements of a state

⁴ Grants to additional states are expected to be made in subsequent years.

⁵ The organizations include the National Center for Higher Education Management Systems (NCHEMS), the American Association of State Colleges and Universities, the National Commission on Accountability, the National Commission on the Future of Higher Education, and State Higher Education Executive Officers.

⁶ Five states—Florida, Missouri, Kentucky, Ohio, and Washington—have partnered with neighboring states to share data about student cross-border mobility. A five-state data exchange demonstration coordinated by NCHEMS is currently in the planning stages (Ewell and Boeke, 2007).

⁷ Cunningham and Milam (2005).

⁸ The results of this latest annual survey can be found at www.DataQualityCampaign.org (accessed December 2006). Similar studies were conducted in 2003, 2004 and 2005.

longitudinal data system. But 78 percent of states currently have five or more of the essential elements. This suggests considerable progress from the previous year, 2005, when the NCEA survey of states indicated that only 58 percent of states had five or more of the essential elements. The 2006 NCEA survey also suggests that 18 states are able to match student records between the pre-kindergarten and postsecondary systems.

A recent NCHEMS study examined more extensively the experience of states in tracking student progression from the earliest years through postsecondary education (Ewell and Boeke, 2007). It found that 11 states can link college student data with high school records. Some of these linkages are in their early stages in terms of the comprehensiveness of coverage of data elements and frequency of merging data. But this will likely change quickly, as 16 additional states in the NCHEMS study stated that they are currently planning to develop linkages between P–12 and higher education.

California Has Many Similar Reasons for Creating a K–20 Data System

The California Master Plan for Education presents an ambitious vision for California's education system.⁹ The plan intends that school districts, county and regional entities, community-based organizations, postsecondary institutions, the private sector, and the state all collaborate to build an aligned education system that ensures that the resources to meet learner needs are available. Accurate and comprehensive data are essential if the state is to develop an implementation plan for this vision.

The California legislature clearly intends that a longitudinal data management system able to track students from kindergarten through college be developed and maintained. AB 1570 (Ch. 916, Statute of 1999), amending the charter of the California Postsecondary Education Commission (CPEC), states that

⁹ See www.ucop.edu/acadinit/mastplan/master_plan2002.pdf.

CPEC shall develop and maintain a comprehensive data base that: (1) ensures comparability of data from diverse sources; (2) supports longitudinal studies of individual students as they progress through the state’s postsecondary educational institutions; (3) is compatible with the California School Information System and the student information systems developed and maintained by the public segments of higher education.

A more recent bill, SB 257 (Section 52052.5, 2003), states even more clearly that the legislature’s intent is “to promote good data management practices with respect to pupil data systems and issues including . . . linking pupil data with data from other agencies and users, including a mechanism to monitor pupil progress in postsecondary education” (emphasis added).

A K–20 SUR system for California would do much to achieve this goal, providing more reliable and transparent information than currently available about how California’s students are progressing over time. Specifically, it could help the state, educational administrators, and faculty address questions that cannot currently be fully addressed, such as

- What are student dropout and college retention rates at all levels of California’s education system?
- Given their students’ performance in college, how can high schools strengthen their curricula and instruction to improve student readiness for college?
- How do high school students who take college preparatory courses perform in postsecondary education segments, as well as in different types of postsecondary institutions?
- How do high school students who pass (or earn a proficient score on) state assessments perform in college?
- What classes of K–12 or college students may need special intervention or attention?
- What factors contribute to students dropping out of high school or higher education segments?
- How can remediation rates in college be reduced?

- How effectively do remediation programs in college help students stay and succeed in college? And, more generally, how effective are programs designed to retain students, prepare them for college, or help them succeed in college?
- Where do California graduates work—whether in state or out of state—and how successful are they in the labor market and more generally in life?
- How adequate is the workforce preparation of students graduating at the various levels of education to meet the demands of California's changing economy? What are the state's returns on its educational investments?

Florida, which maintains the oldest and most mature integrated SUR, has been using its data for multiple purposes, including accountability at the K–12 and postsecondary levels; to evaluate programs such as its scholarship, reading, and charter schools programs; and to assess return on educational investments. Its student data system is also used by legislators, administrators, teachers, and faculty to engage in data-driven policymaking and program design such as pay-for-performance, improving articulation between segments and across campuses, and analyzing course-taking patterns. Finally, Florida uses its student data systems to support its strategic planning activities.

In brief, a SUR data system properly designed and maintained could be used for description, diagnosis, program design, and evaluation. It could increase public understanding of what the system is accomplishing, what its issues are, and how it could be improved. It could help formulate policies and assess the returns on programmatic investments in education.

California Currently Lags Behind Other States

Despite the potential benefits, California is currently lagging behind most other states in developing a data system able to help policymakers and others align, assess, and improve educational programs (Hansen, 2007). The 2006 NCEA survey indicates that the state has only three

of the ten elements deemed “essential” for a longitudinal data system, and only partially at that (Ewell and Boeke, 2007). The survey suggests that California is not one of the 18 states with the ability to match student records between the pre-kindergarten and postsecondary systems. Nor is California among the states currently planning or implementing linkages between K–12 and postsecondary databases.¹⁰

Indeed, only recently has the state committed to developing a K–12 California Longitudinal Pupil Achievement Data System (CALPADS). CALPADS was authorized by state legislation in 2002 and 2003 to enable California to develop a longitudinal K–12 SUR data system to comply with the requirements of No Child Left Behind.¹¹ In 2006, the California Department of Education (CDE) received a \$3.25 million federal grant to supplement the state funding of CALPADS development and implementation (Hansen, 2007). CALPADS, which will move California closer to where many other states currently stand, is now in its early stages of development.

The California School Information Services (CSIS) has legislative authority to transfer individual K–12 student transcripts to other K–12 institutions and to California’s three postsecondary education segments. Because of funding constraints, academic records transfer through CSIS is currently occurring on a limited basis through the Eligibility in the Local Context (ELC) Program.¹² However, the standards for transcript transmission have been developed and this work can be leveraged should more funding become available.¹³

¹⁰ The California Community College system is developing a statewide Internet-based system (CCCTran) for requesting, viewing, and transmitting academic transcripts among authorized educational institutions and their trading partners. Eventually, links between the K–12 and community college system will be formed, but likely on a fragmentary and voluntary basis.

¹¹ See SB 257 (2003) and SB 1453 (2002).

¹² ELC is one of three paths to freshmen eligibility for the University of California; the other two paths are Eligibility in the Statewide Context and Eligibility by Examination Alone. Under ELC, the top 4 percent of students in each participating California high school class are designated UC-eligible based on the coursework taken while in high school.

¹³ According to staff at CSIS, discussions are currently under way to design a central electronic transfer of transcripts to colleges.

Differences Between California and Other States Might Make Creating a K–20 System Challenging

Unlike any other state in the nation, the governance of California’s public education system is divided between four relatively autonomous segments:

- a K–12 segment, under the administration of an elected state Superintendent of Education and a State Board of Education
- the California Community Colleges, under the administration of a Board of Governors
- the California State University, under the administration of a Board of Trustees
- the University of California, under the administration of a Board of Regents.

In addition, the governor has recently added an appointed Secretary of Education in the governor’s office to coordinate state education policies.

What is particularly unique about California is its three post-secondary segments. Each has its own policies and practices for information collection and dissemination and each has its own SUR data system. Even more fundamentally, the major policy instruments and practices to effectuate change, such as finance, accountability, curriculum, and standards, are not coordinated across the segments. This is in marked contrast with the data practices of other states. Of the 40 states currently with postsecondary SUR systems,

- thirty have a single SUR covering students in all public higher education institutions (as well as, in some cases, certain private institutions)
- four have a single SUR covering students in all public and private institutions
- five have two SUR systems that cover the state university and community college systems separately¹⁴ (Ewell and Boeke, 2007).

¹⁴ North Carolina, Oregon, Washington, Wyoming, and New York.

California’s AB 1570 (1999) legislation requires that CPEC develop a data system integrating the state’s three segments of higher education in a way that is compatible with—and, by implication, linkable to—the CSIS that currently covers the state’s K–12 segment: “CPEC shall develop and maintain a comprehensive data base that . . . *is compatible with the California School Information Services* and the student information systems developed and maintained by the public segments of higher education” (emphasis added).

But the three independent SUR data systems for college-level students—and the four-way fragmented nature of education governance in California more generally—make it especially difficult to link K–12 and postsecondary data in California.¹⁵ For example, global responsibility for a complete set of longitudinal data would rest with four entities: the State Department of Education, the Governors of the California Community Colleges, the Trustees of the California State University, and the Regents of the University of California (First Amendment project, 2007, p. 16). To this point, the state’s legislature has not backed up its intent with the tools needed to carry it out, either by providing institutional authority to anyone to do so, by mandating that the segments cooperate, or by providing the funding required for development and maintenance of an integrated K–20 system.

Focus of This Study

Policymakers, personnel from education agencies, and the researchers who have tried to use California education data in their work know well the deficiencies in the state’s education data systems. But few, if any, have a comprehensive picture of the current status of the data sys-

¹⁵ In addition, it is difficult in California to measure student success in postsecondary education because students may have transferred from one segment to another. A dropout in one segment may complete a degree in another. Also, California has a number of students enrolled at more than one institution or segment at any one time. A more comprehensive approach to tracking student progress is needed, particularly in light of national studies showing that more than half of students attend more than one institution in pursuit of a bachelor’s degree (Ewell, Schild, and Paulson, 2003).

tems used by the state's four education segments. Nor do they have a thorough understanding of the technical, political, financial, and legal issues involved in integrating the state's various student data systems.

Recognizing these gaps in current knowledge, the Hewlett Foundation asked the RAND Corporation (1) to review the status of the SUR data systems presently available in California at all four segments of education, and (2) to assess the political, financial, and legal challenges and design issues that might stand in the path of developing a K–20 longitudinal SUR data system. It also asked RAND to identify steps that might be taken to make the development of such a system possible.

This report presents the issues and challenges involved and suggests possibilities for change in California's student data systems and, eventually, in the ability of the state, educational administrators, and faculty to formulate policies and design curricula to improve California students' performance using accurate and reliable information. As such, it will serve as a shared foundation for discussions among the diverse stakeholders who would need to be involved in any improvement effort. The study addresses the following questions:

- What are the principal characteristics—e.g., purposes, content, and quality—of existing longitudinal SUR systems in California?
- What are the challenges California would have to address to develop an integrated K–20 student-level data system?
- What are key initial steps that might be taken to enable California to develop such an integrated student information system?

Our primary focus is on the prospect of developing an integrated, longitudinal K–20 SUR system for California. We recognize, however, that the ability to link other kinds of information to student records would vastly expand the number of policy-relevant questions that analysts could address. For instance, data on the qualifications and experience of a student's teachers, the financial resources devoted to his or her education, and his or her participation in the workforce after graduation would be additionally valuable. With this in mind, a secondary

purpose of this study is to serve as the groundwork for future discussions of the possibilities for linking these types of data with information on students.

Study Approach

1. We conducted a selected review of literature on longitudinal SUR systems, as well as on the experience of states that built such systems. The purpose of this review was to understand the SUR systems available around the nation, the issues states had to overcome to build them, the costs of developing and maintaining them, and the uses made of them. We reviewed the literature including the results of the 2006 survey of state SUR data systems conducted by NCHEMS and the Data Quality Campaign (DQC), as well as the findings from case studies on state approaches to building longitudinal data systems conducted by the Data Quality Campaign.¹⁶ In addition, we reviewed recent work by the Education Commission of the States on P–16 Collaboration in the States.¹⁷

2. We conducted interviews with SUR administrators in four selected states. We selected four states—Florida, Kentucky, Oregon, and Texas—on the basis of multiple factors:

- the literature
- conversations with staff at NCHEMS
- the current capacity of these states to match K–12 and higher education student data¹⁸

¹⁶ The DQC conducted case studies of Florida, Utah, Virginia, and Wisconsin that describe those states' experience in building longitudinal data systems, including the costs, challenges, pitfalls, and other perspectives in hindsight. These case studies are available at www.dataqualitycampaign.org (accessed April 15, 2007).

¹⁷ See, for example, L'Orange and Ewell (2006) and www.ecs.org/ (accessed May 20, 2007).

¹⁸ As we learned in our interviews, Kentucky and Oregon were still in the early stages of linking their K–12 and postsecondary student data systems and, hence, we gained little information on the challenges they faced.

- the size of their public education systems
- the breadth of content of their SUR systems
- the number of their higher education governance systems.

Details on the four states selected are included in Appendix A.¹⁹

We interviewed the individual in charge of the student unit record systems in each state except in Oregon, where three individuals were interviewed. In these interviews, we inquired about the status of their student data system, its oversight and governance, the challenges encountered (and still being encountered) in its development, and access to the data.

Both the general information gathered from the literature review and the specific information obtained from the state interviews guided our subsequent California-focused tasks and illustrated how some of the challenges faced by California might be addressed.

3. We reviewed and documented the characteristics of the SUR data systems currently in use and under development in California. We obtained the data for this effort from the Web sites of the various systems, as well as from the collection of file dictionaries from the entities that maintain the systems. We complemented these data with interviews with staff of the offices that collect and maintain the systems. This task allowed us to document existing systems, identifying their purpose, coverage, content, and format. It also provided initial input into the compatibility of data systems and the potential to link across systems.

In parallel, we reviewed California state law as it relates to the development and authorization of SUR data systems. This included a review of literature on FERPA and, specifically, how it might affect SUR data systems (see, for example, Coleman, Palmer, and Winnick, 2006). Finally, we reviewed any Memorandums of Understanding that existed between education segments or between education segments and other organizations pertaining to the sharing of SUR data.

¹⁹ We initially also chose to conduct interviews with administrators in Tennessee and Washington State, but early conversations indicated that they were not yet linking their K-12 and higher education SUR in any significant way.

4. We conducted face-to-face interviews with key state-level stakeholders, education-segment leaders and system administrators, and researchers. The purpose of our interviews was to develop a deeper understanding of the opportunities for and the challenges involved in developing an integrated K–20 SUR data system in California. Overall, we conducted 29 interviews at the following institutions:

- California Partnership for Achieving Student Success (CalPASS)
- California Community Colleges Chancellor’s Office (CCCCO)
- California Department of Education
- California Department of Finance
- California Policy Research Center
- California Postsecondary Education Commission
- California School Information Services
- California Senate
- California State Assembly
- California State University Chancellor’s Office (CSUCO)
- Institute for Higher Education Leadership and Policy
- Legislative Analyst’s Office (LAO)
- National Center for Public Policy and Higher Education
- Secretary of Education
- State Board of Education
- University of California’s Office of the President (UCOP)

Interviews generally ranged from 30 to 60 minutes and followed interview protocols appropriate to the type of respondent. The interviews were semi-structured to give us the necessary flexibility to ask follow-up questions aimed at eliciting the reasons for key practices, decisions, or opinions. The interview protocols covered various topics, including²⁰

- data collection procedures
- data coverage, quality, and reliability of audit procedures

²⁰ See Appendix B for the interview protocol.

- authority to change data content and definitions
- financing, costs, and staffing of systems
- current use of data
- potential use of an integrated SUR system
- confidentiality procedures and issues
- access limitations and issues
- choice of student identifier and issues
- decisionmaking procedures to make changes to data systems
- perceived challenges—political, technical, governance, financial, legal, and administrative—in developing an integrated K–20 SUR system
- suggested solutions to address identified challenges.

The information we collected from the various respondents was systematically grouped by topic and compared within and across types of respondents to identify key issues. We also used the synthesized information to develop a set of recommendations on steps that could be taken in the short and long term to effectively move in the direction of developing an integrated K–20 SUR system in California.

Study Limitations

This report focuses on California’s current student data systems and on issues and challenges that would need to be addressed to improve their usefulness to policymakers, educational administrators, faculty, and the public. Although a K–20 SUR data system such as the one described here can help identify issues and opportunities for improving education in California, actually addressing these issues and effectuating change will also require a better coordination of the policies—finance, curriculum, accountability, and standards—that currently exist between the four quasi-independent segments of the California education system. Resources for this study did not permit us to address these broader but important issues.

Organization of the Report

Chapter Two describes the student-level data systems currently in use in California. Appendix C contains more details on the characteristics of each of these systems. The challenges and design issues that the state would face in endeavoring to develop an integrated longitudinal K–20 student data system are discussed in Chapter Three. The final chapter suggests steps that would have to be taken to lead to the actual development and maintenance of such a data system.

California's Current Student Data Systems

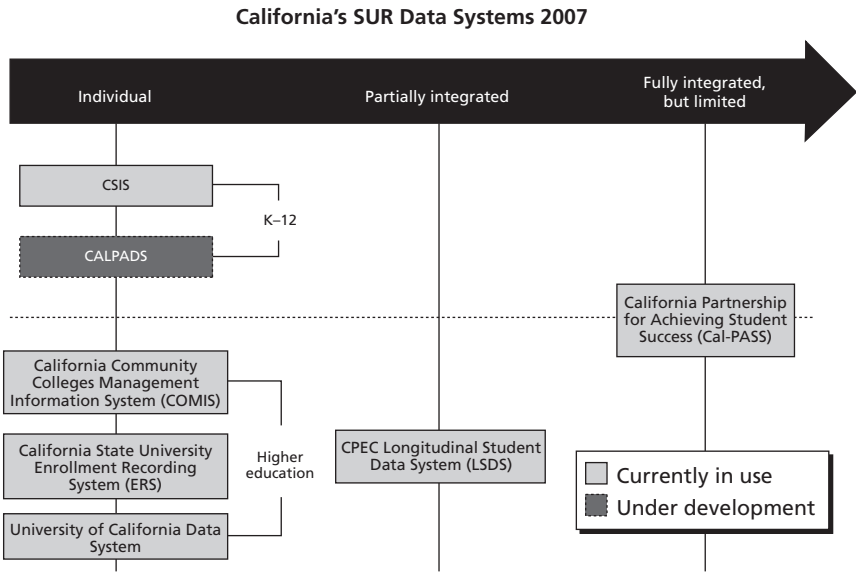
Although California trails other states in developing a K–20 SUR data system, it is not at square one. At present, the state has the embryonic architecture of a longitudinal system. Systems currently used by the state's individual education segments already collect many of the data needed to eventually develop and maintain a fully integrated file that would track individual students from kindergarten through graduation from college. In addition, California has several data initiatives under way that promise to fill some of the gaps in its existing ability to gather education data.

Of seven student data systems currently in use or being developed in the state (Figure 2.1), five are stand-alone systems:

- a mandatory K–12 system that contains limited student-level data on enrollment and demographics for all of California students and additional information on 60 percent of students in districts participating voluntarily—currently in use
- a new mandated comprehensive public K–12 system—under development
- one system for each of California's three public higher education segments—all currently in use.

A sixth system is partially integrated, bringing together student-level data from all three higher education segments. And another rapidly expanding system fully integrates student-level data from individual institutions from all four California educational segments, both K–12 and higher education, albeit on a still limited regional and voluntary basis.

Figure 2.1
California’s Seven SUR Data Systems in Use or Under Development



RAND MG695-2.1

In this chapter, we briefly describe each of these SUR data systems. Appendix C contains more detailed information.

Individual Student Databases for K–12 Students

California School Information Services

The CSIS was first established by the California legislature in 1997 (AB 107, Chapter 282) and was permanently authorized in 2002 (AB 1115). CSIS helps the 1,265 public K–12 Local Education Agencies (LEAs) assign and maintain the Statewide Student Identifier (SSID) and annually collects aggregated data on enrollment, grade, and primary ethnicity from each of these LEAs, as well as the official graduation and drop-out data. CSIS also administers a voluntary program called CSIS State Reporting. Of the 1,058 state LEAs, 263 are currently participating in CSIS State Reporting, covering about 60 percent of the state’s public

school pupils. Participating districts electronically transmit individual student data in the fall and spring of each school year. CSIS then aggregates these data for State Reporting LEAs and transmits them to the California Department of Education to meet reporting requirements for the California Basic Educational Data System (CBEDS), the Student National Origin Report (SNOR), and the Spring Language Census (R-30).

CSIS keeps the data it receives for up to seven years, if not needed for evaluations or audits.

Purposes of the System. CSIS is intended to (1) build the capacity of LEAs to implement and maintain (their own) comparable effective and efficient pupil information systems, (2) enable the accurate and timely exchange of pupil transcripts from LEAs to postsecondary institutions, and (3) help LEAs transmit state and federal reports to the California Department of Education.

Content: Categories of Information. The individual student information transmitted includes

- sociodemographic: gender, race/ethnicity, date of birth, primary language, parent education, English proficiency, disability status, socioeconomic status, and, since 2004, the SSID
- enrollment: institution, date, status, grade level, withdrawal date, and reason for withdrawal
- grades completed, degrees, and graduation date
- courses taken—including remedial, summer, advanced placement (AP) and honor courses.

Using the unique SSID that is assigned to each K–12 student since 2004, CSIS will be able to maintain a longitudinal enrollment history of all students. Although CSIS could be used as a vehicle for academic records transfer, transcript data including state assessment results, course completion and history, and grades are not stored in the CSIS longitudinal data base. The Fiscal Crisis and Management Assistance Team (FCMAT) founded in 1992 and the California Board of Education must approve any changes in data elements requested from the LEAs .

System for Tracking Students. California’s SB 1453 legislation in 2002 and SB 257 in 2003 mandated that CSIS assign a unique, non-personally identifiable student identification (ID) number to all students in California public schools from kindergarten to grade 12.¹ CSIS completed this assignment as of school year 2003–04 and continues to operate the SSID system. It has developed an algorithm to assign a non-personally identifiable number to a student when he or she first enrolls in a K–12 public school. CSIS has also developed and maintains a system for schools to locate the assigned SSID when a student transfers from one school to another so that the number follows the student throughout his or her K–12 career. This process for maintaining the integrity of the SSID includes quality checks to assure that an individual student is not assigned duplicate SSIDs, that two students do not share the same SSID, and that a student is not reported as enrolled in more than one school district in California.

Since 2004, CSIS has been collecting an Annual SSID Maintenance submission from LEAs in California. And, since 2006, the annual SSID Maintenance submission has been used to collect official enrollment data on individuals from all districts. Districts that do not comply with CSIS’s established standards for submission of enrollment data may lose categorical funding.

Current Quality of the Data. CSIS does some validity checks of the data it receives but, overall, the data are only as accurate as each individual Local Education Agency ensures them to be through the collection and auditing of data submitted by the schools. Overall, the accuracy of the CSIS data is unknown. It is generally recognized that there are great variations in each district’s capacity to collect, input, and verify its data.

Access. The data are directly accessible by authorized staff and users. They are not directly accessible to outside researchers. The legislative authority and mission for CSIS do not include research. Analytical staff members at CSIS work on system requirements, but they are not involved in research. Annual SSID Maintenance data and State

¹ This is the same legislation that established California’s other K–12 data system, CAL-PADS, currently under development (see below).

Reporting data collected by CSIS are transmitted to CDE. Researchers may gain access to the data via requests to CDE, not CSIS.

CSIS also sends student transcripts to colleges on behalf of schools when the schools request.

Staff. CSIS has 46 staff members serving 1,265 LEAs.

California Longitudinal Pupil Achievement Data System (CALPADS)

CALPADS will eventually become the K–12 student-level data system for California. Under development, it is not expected to be completed until 2009. In anticipation of this new, mandatory state data file, CSIS is no longer accepting new Local Education Agencies into its CSIS State Reporting. However in 2005–06, the legislature funded CSIS to administer the Best Practices Cohort Project, a project designed to help LEAs improve their local data management practices and prepare for CALPADS. All LEAs that did not participate in CSIS State Reporting are eligible for the Best Practices Cohort Project.

CALPADS was authorized by California SB 1453 (2002) and SB 257 (2003) mainly to meet the federal, state, district, and school reporting requirements under the No Child Left Behind Act of 2001.² All California public schools and districts are required by law to participate in CALPADS. Given that many of the state's schools and districts will have to develop the capacity to provide the data required—especially those that have not participated in the CSIS K–12 initiative—it can be expected that it will be some time beyond 2009 before CALPADS actually records all data elements on all students.³

CALPADS will retain information on individual students for up to 20 years, if needed for programmatic reasons.

Purposes of the System. The authorizing legislation calls for a database “to assess the long-term value of [the state’s] educational investments and programs and provide a research base for improving

² This same legislation mandated that the CSIS assign a unique, non-personally identifiable student ID to all students in California public schools.

³ Several of our respondents suggested that by 2009, the system will cover reliably about 80 percent of the LEAs in the K–12 system, although the proportion of students included is expected to increase over time as the smaller districts develop the capacity to provide the data electronically.

pupil performance.” The purposes of the system are to provide (1) a way to evaluate California’s educational progress and investments over time, (2) information to LEAs that they can then use to improve pupil achievement, and (3) an efficient, flexible, and secure way to maintain longitudinal pupil-level data statewide.⁴

Content: Categories of Information. The system is expected to initially contain only those data elements required by the No Child Left Behind Act. These comprise basic information on individual student status and educational progress, including

- sociodemographic: gender, race/ethnicity, date of birth, primary language, parent education, English proficiency, disability status, socioeconomic status, place of residence, and migrant status
- enrollment: institution, date, status, grade level, withdrawal date, and reason for withdrawal
- grades completed, degrees, and graduation date
- courses taken—including remedial, summer courses, AP, and honor courses
- test scores: STAR, CAHSEE, and CELDT (the system also collects information on students not tested); the system is not currently slated to include student grades, nor will it contain indicators of readiness for college, such as ACT and SAT test results.

System for Tracking Students. In preparation for CALPADS, in school year 2005–06 every student in California’s public K–12 education system was assigned a unique, yet non-identifiable SSID that will follow the student throughout his or her enrollment in the system.

CALPADS will contain only non-identifiable student data. Student identifiable information will be available only at the local educational agency level (i.e., only districts’ (LEA) authorized personnel will be able to match an SSID with a student).

Access. It is expected that this data system will be accessible to

⁴ The information in this subsection is derived both from California Department of Education (2004) and from interviews with CALPADS staff.

- all educational agencies in California
- their legislative and executive overseers
- researchers from established research organizations, on approval of the state Department of Education.

The public will not have direct access to CALPADS.

Individual SUR Databases for Postsecondary Students

California Community Colleges

The California Community Colleges Chancellor's Office maintains a consolidated longitudinal file, the California Community Colleges Management Information System. The system has usable individual data linked over time since 1992.

Purposes of the System. COMIS data are used (1) to prepare the report for Integrated Postsecondary Education Data System (IPEDS) and other required federal and state reports, (2) to conduct research, program evaluation, and statewide evaluation, and (3) to track student outcomes.

Content: Categories of Information. COMIS contains mostly standardized student-level information from all of the state's 109 community college campuses, including

- socio-demographic: gender, race/ethnicity, date of birth, disability status, socioeconomic status, citizenship, residence, zipcode, and student income
- high school of origin and date of attendance⁵
- enrollment: institution, status, grade level, degree-seeking status
- grade point average (GPA): term and cumulative
- units taken: term units attempted and units earned, cumulative units earned, and transfer credits
- degrees awarded and graduation date

⁵ COMIS includes information on the high school last attended, generally the school of graduation if the student graduated from high school and, if not, the school last attended. Students ages 21 or over need not provide this information.

- courses taken—including remedial and noncredit courses
- financial aid.⁶

It does not include ACT and SAT scores, which are not required for application to community colleges.

COMIS also contains data on staff that may be linked to students.⁷

System for Tracking Students. The social security number (SSN) is the unique student identifier used for matching students over time. Students who refuse to provide an SSN and students without one are assigned a campus-specific ID number that does not necessarily follow the student if he or she transfers from one college to another.

Reportedly, about 93–94 percent of students working toward an associate degree have an SSN associated with their records. The SSN reporting rate is much lower (about 80 percent) for students taking only noncredit courses, because registration in this sector takes place during the first class and collection of the number depends on the instructor.

Current Quality of the Data. Although the Chancellor’s Office makes validity checks of the data, each of the 72 local boards has autonomy.⁸ Data accuracy depends on the diligence of each campus. Data tied to categorical programs or to IPEDS are generally reliable. Other data elements are believed to be less so. The Chancellor’s Office has little leverage to require collection of data not mandated by the state, the federal government, or other entities.⁹

⁶ Taxonomy of Programs (TOP) is used by community colleges to collect and record information on programs and courses that have similar outcomes. See <http://www.curriculum.cc.ca.us/curriculum/RegulationsGuidelines/TOPCodes.htm> (accessed April 23, 2007).

⁷ In addition, COMIS includes faculty instruction work assignments for each term that is tied to the section and student. It also includes demographic information about the instructor.

⁸ The majority of campuses use Datatel, Sungard, Peoplesoft, or Santa Rosa as their operating system which, reportedly, gives some consistency to the data.

⁹ The Chancellor’s Office limited authority over individual colleges and university campuses is generally common to each of the three higher education segments. The central offices rely on mandates and funding tied to data elements to get the colleges and campuses to

The Chancellor's Office does not perform audits, although it is considering doing so eventually. Currently, it seeks to improve the comprehensiveness and quality of the data in two principal ways:

- by reporting the data by college, giving each institution an incentive to provide accurate data,
- by identifying in their data reports the colleges that do not supply any of the required data elements.

Reportedly, once college-specific data were put on the "data-Mart," they improved.¹⁰

Access. COMIS data on individuals are accessible electronically by authorized personnel at each of California's 109 community college campuses (but it is password-protected). On occasion, the Chancellor's Office has also made its non-identifiable data available to outside researchers.

The Community Colleges Chancellor's Office links the National Student Clearinghouse data with its own¹¹ and has occasionally linked those data with data from California's employment and wage records, the military services, Department of Social Services aid data, and the national student clearinghouse; among other purposes, this allows assessment of students' workforce experience and determination of which students go to private or out-of-state institutions.

The Community Colleges Chancellor's Office also has provided access to its student identifiable data to CPEC since 1993.

report specific items. As a result, changing or adding data elements is difficult and rarely succeed unless mandated by the respective segments' board or the state.

¹⁰ The data Mart provides college-specific aggregated information on student demographic characteristics, retention, financial aid, and the like. See <http://www.cccco.edu/> (accessed November 15, 2007).

¹¹ The National Student Clearinghouse is a nonprofit organization that serves member higher education institutions in degree, diploma, and enrollment verification.

Staff. The Chancellor’s Office management service that maintains this longitudinal data file includes five information technology (IT) specialists and five analysts.¹²

California State University

In fall 1973, the Chancellor’s Office for the California State University implemented its consolidated data archive, the Enrollment Recording System (ERS). The first entries were enrollment records for all students attending a CSU campus. In 1975, graduation records were added to the archive and in 1985, application records were also added. Longitudinal files can be created from the archive. Social Security numbers are used as the primary key to link student records across time. Today, all campuses but two—which continue to use legacy systems—are using the same software operating systems, PeopleSoft, to create their ERS files.

Purposes of the System. Cal State uses the data primarily (1) to track student retention and graduation and (2) to support regular term reports, IPEDS, and state budget requests. It also prepares various systemwide reports on academic performance such as graduation rates, as well as campus-specific data. To a limited extent, CSU uses the data for internal research.

Content: Categories of Information. The ERS contains standardized information on individual students from all of the 23 CSU campuses, including

- sociodemographics: gender, race/ethnicity, date of birth, English proficiency, disability status, citizenship, residence, parent education, and student income
- high school or institution of origin: year of high school graduation, high school GPA, college preparatory courses, AP and honor courses, and entry-level math proficiency

¹² The analytical staff size in the Office of Research and Planning was reduced in 2001 in the midst of the state budget crisis.

- enrollment: institution, status, grade level, major, degree-seeking status, and admission basis¹³
- GPA: cumulative
- units taken: term units attempted, cumulative units earned, and transfer credits
- degrees awarded
- courses taken—including remedial and summer courses only
- test scores: ACT, SAT,¹⁴ and others
- financial aid.

Unlike the Community College data system, the CSU file does not contain individual course and grade information. This information is available at the campus level but is not currently being transferred to the consolidated ERS.

System for Tracking Students. CSU campuses that use PeopleSoft assign their students a nine-digit random ID number that is not the social security number. The campus number does not move with the student if he or she transfers to another CSU campus. Consequently, it cannot be used to track students even within the CSU system.

The SSN is used to link records across campuses and over time (see below). Reportedly, about 5 percent of students do not have an SSN and another 5 percent provide an incorrect one. If a student provides an erroneous SSN, the record is not changed and the incorrect number is used to track the student only within the CSU system.

CSU is also collecting the SSID on its entrance application as an optional element.

Current Quality of the Data. Each CSU campus is responsible for the accuracy of the data transferred, although CSU IT staff members run validity checks on the data received. Respondents say that the data

¹³ Since the early 1990s, ERS has also contained information on students who applied but did not enroll.

¹⁴ This information is collected only for students who enter Cal State directly from high school. Other students, including transfer students, are not required to submit ACT or SAT scores.

are most accurate when tied to funding or reporting requirements and less so otherwise.

Access. ERS student-identifiable data are password-protected and accessible only to authorized staff at CSU headquarters and the campuses. Non-identifiable student-level data are accessible by authorized personnel in relevant state agencies and by other colleges and universities.

Outside entities may submit a request to use CSU SUR data. Requests are carefully reviewed by the CSU counsel's office. Reportedly, most requests for individual data are denied, because they can be met by information already contained in reports posted on the Cal State Web site.¹⁵ CSU also gives the state's Department of Labor access to its data file.

The California State University provides the California Postsecondary Education Commission with an extract of its identifiable data and indirect access to its data to the community colleges and the University of California.¹⁶

Staff. ERS is managed by four IT staff members. In addition, the office has five analytical staff members.

University of California

The University of California maintains a consolidated longitudinal student-level data file, the Corporate Student System (CSS), from all ten of its campuses. Records have been kept since the early 1980s but complete records on individual students have been available only since 1999, with somewhat less complete records dating back to 1994.

Content: Categories of Information. The CSS data include

- sociodemographic: gender, race/ethnicity, date of birth, English proficiency, primary language, disability status, citizenship,

¹⁵ <http://www.calstate.edu/AS/>

¹⁶ Indirect access means that one segment submits a request, such as tracking a specific subset of students, to another segment for specific information that is then processed by the latter.

residence, zipcode, parent education, and student and parent income¹⁷

- high school of origin: graduation date, high school GPA, and prior college attended
- enrollment: institution, status, grade level, major, degree-seeking status, admission, and withdrawal reason
- GPA: term and cumulative
- units taken: term units attempted, term and cumulative units earned, and transfer credits
- test scores: ACT and SAT
- degrees awarded and date
- financial aid.

The file does not contain information on individual courses taken and student grades, although this information is available at the campus level.¹⁸

Starting in 2008, UC will collect the SSID as an optional data element on its application form.

Purposes of the System. The data are used primarily by the University of California's President's Office to meet management and operational needs related to student enrollment and performance.¹⁹ They are also used for planning, financial reporting and analysis, and reporting to governmental and nongovernmental agencies.

The University of California has used the National Student Clearinghouse to track which students go out of state and is currently using CCCCCO data to examine points in the pipeline of transfer students to UC.

System for Tracking Students. Each University of California campus assigns each student a unique and arbitrary ID number. That

¹⁷ About 25 percent of students do not provide information on their parents' income.

¹⁸ One respondent indicated that UC's academic senate is interested in collecting data on courses and grades but that the cost in both time and dollars was felt to outweigh the benefits. He said "we don't like to overburden our campuses with requests for data unless the data is critical for our operations or decisionmaking."

¹⁹ Various offices within the president's office have an analytical capability.

number does not move with the student if he or she transfers to another UC campus or to any other postsecondary public institution. In the case of a transfer, a new ID number is assigned.

As with the community colleges and CSU, the University of California uses the SSN to link individual student records over time and across campuses. It estimates that about 96 percent of its students have an SSN. Students without one cannot be tracked either within or outside the UC system.

Current Quality of the Data. The University of California has no way to verify the accuracy of the data that individual campuses provide, but it does run field validation and cross-validation checks.²⁰ It has also started to check whether new students were previously in the system.²¹

Access. Only authorized personnel and officials can access student-identifiable data from the CSS file. Some other users have Web site access to aggregate data or can request ad hoc reports and data extracts through the Information Resources and Communication Office (IR&C).²² Other individuals, including researchers, may obtain authorization to access non-identifiable data by contacting the UC office of the President's Data Center and following a predetermined protocol and approval process.

The University of California also gives the state's Department of Finance its non-identifiable student data (i.e., excluding SSNs). UC provides an extract of its identifiable data to the California Postsecond-

²⁰ Reportedly, the most frequent errors occur because campuses do not follow the rules, submit the wrong data elements, and make input errors.

²¹ In interviews, staff members indicated that campuses occasionally make changes in the data they collect or change their coding sequence; UC corporate has no systematic way to detect such changes. For instance, one campus stopped providing the SSN, making it impossible to link students longitudinally. The problem was eventually fixed. To improve data quality, UC has begun to put the data on a public Web site, which should encourage the campuses to be more accurate in their reporting.

²² UC is planning to launch a new Web-based "UC Statefinder" in fall 2007. The UC Statefinder will allow users to query longitudinal undergraduate outcomes (beginning with admissions and later retention, graduations, and time-to-degree) on an aggregated basis. It will also permit queries on admission and performance by individual high schools and California Community Colleges.

ary Education Commission and indirect access to its data to the Community Colleges Chancellor's Office and CSU.

Staff. Ten IT staff members and analysts manage and use the University of California corporate data system.

Partially Integrated SUR Databases

California Postsecondary Education Commission

In 1999, CPEC was directed by the California legislature (AB 1570, 1999) to “develop and maintain a data-collection system capable of documenting the performance of postsecondary education institutions in meeting the post high school education and training needs of California’s diverse population.” In 2001, California’s three postsecondary education segments—California Community Colleges, California State University, and the University of California—signed a Memorandum of Understanding with CPEC that stated the conditions under which it would serve as the clearinghouse for data on postsecondary education in California, using personally identifiable data provided by each of the three segments. This agreement gives authorized personnel from each segment access to the others’ centralized data files.²³ Access is for the “purposes of evaluation or educational research aimed at improving instruction and documenting the success of California’s public colleges and universities in meeting policy and legislative expectations.” The agreement specifies that the partners may review and approve research projects and review the final drafts and results of those projects.

The Community Colleges Chancellor’s Office has made its data available to CPEC since 1993. But it was not until mid-2005 that CPEC received the first set of data from the California State University and the University of California. The delay in transferring the information to CPEC stemmed primarily from these two segments’ strict interpretation of FERPA and state laws on student confidentiality, which they

²³ The memorandum is dated December 11, 2001. To get the AB 1570 legislation, CPEC agreed to limit the data it would receive to those data elements it had already been collecting plus a student identifier. This was agreed on to avoid triggering mandated state costs.

understood as preventing them by law from sharing student identifiable data with a third party.

CPEC has had some discussion with the Association of Independent California Colleges and Universities (AICCU) regarding its participation in the CPEC data-collection effort, but no progress has been made to date.

Purposes of the System. CPEC uses its data file primarily to do analyses requested by the legislative branches, to do its own analyses, and to write reports on specific issues.²⁴

Content: Categories of Information. CPEC accesses the COMIS database of the California Community Colleges Chancellor’s Office directly and retrieves the data it wants. In contrast, the identifiable data that the California State University and the University of California now send CPEC annually (typically in June or July) is an extract of their individual SUR data systems going back to 2000.

The resulting CPEC Longitudinal Student Data System contains information on students in each of California’s three public postsecondary education segments, including

- sociodemographic: gender, race/ethnicity, date of birth, residence, and disability status
- high school of origin and date of graduation
- enrollment: institution, status, grade level, major, and degree-seeking status,
- units taken: cumulative units earned
- degrees awarded.

CPEC has also added aggregate school-level data such as high school Academic Performance Index (API), the percentage of students eligible for subsidized school lunches, and census data on the sociodemographic and labor characteristics of the surrounding communities of California high schools.

CPEC has been seeking further data elements from the postsecondary segments, such as reasons for admission—believed to be an

²⁴ CPEC reports can be seen at <http://www.cpec.ca.gov>.

important consideration in assessing student progress—and students' zipcodes (although this information is available only for community college students).

System for Tracking Students. CPEC receives the data with the student SSN. It then converts the SSN to a random number. For all practical purposes, CPEC's data file contains data that are nonidentifiable and should meet FERPA requirements for use in research by third and fourth parties.²⁵

Current Quality of the Data. CPEC seeks to standardize the definitions of the data elements during the edit phase of its data processing. It reports that the data are "imperfect and incomplete and that care must be taken in [their] use and interpretation." Our respondents from the higher education segments strongly echoed this warning. It reportedly took CPEC one year to clean up the first submission of data received from the segments. For instance, data were missing or not coded reliably and there were gaps in the data provided. Other than checks for internal validity, neither CPEC²⁶ nor the segments have the capacity to verify and validate the accuracy of the data provided by the 109 community colleges, 23 state universities, and ten University of California campuses.

For several reasons, it is proving difficult to address data completeness and quality issues in the CPEC SUR data system. First, the segments do not have access to CPEC's aggregated file and, hence, cannot use it for analytical purposes. Per the Memorandum of Understanding between CPEC and the segments, the segments cannot view each other's files.²⁷ Second, the Department of Finance's insistence on keeping the costs of data collection to a minimum seems to have hampered efforts to improve quality. Finally, strained communications between CPEC and some of the segments also seem to stand in the way. Although the agreement requires that the segments and CPEC

²⁵ Of course, eventual users of this file would still have to demonstrate that they meet all FERPA requirements and cells that have few observations should not be reported.

²⁶ CPEC also compares data from the most current year to data from prior years.

²⁷ As noted above, the segments have "indirect access to each other's file," meaning that any request by one segment from another is processed by the latter.

meet regularly, our respondents indicated that no such meetings have taken place in a “long time.” There is a perception among the segments that CPEC receives the data but does not consult with them about the data, or about interpretations of the data, or about research projects based on the data. They perceive this as a breach of the Memorandum of Agreement, in which it is stated that the partners will approve research projects and review final drafts and results.²⁸ There also is a perception that some of the studies done by CPEC have been unfairly critical of the segments’ performance.

Access. Only designated CPEC staff can access the identifiable data in the system. CPEC said that the non-identifiable data are accessible to anyone, including to the segments and researchers.²⁹

Staff. CPEC currently has six IT staff members and five analysts.

Fully Integrated SUR Databases

California Partnership for Achieving Student Success³⁰

Since 2003, Cal-PASS has maintained non-identifiable, longitudinal student-level data files on behalf of 11 regional consortia of K–12 schools, community colleges, Cal State campuses, and University of

²⁸ CPEC indicated, however, that segments are given an opportunity to review and comment on its studies. Nonetheless, there seems to be confusion on whether the agreement is still in effect. CPEC told us that the agreement expired one year after it was signed but the segments indicated that they still operated as if it were still in effect.

²⁹ CPEC indicated that the prohibition on sharing its identifiable student information with the segments came from a ruling of the U.S. Department of Education’s Family Policy Compliance Office (FPCO).

³⁰ Cal-PASS is a partnership between the California Community Colleges Chancellor’s Office and the Grossmont-Cuyamaca Community College District. Its funding is allocated to the Community Colleges for IT projects. There is no dollar cost to participating institutions other than staff and faculty time to provide the data and participate in the councils.

California campuses throughout the state.³¹ It has also provided the consortia with analytical support.

Participation in a consortium is voluntary. A school or college signs a Memorandum of Understanding agreeing to share its student data with all other members of the consortium³² and maintain the confidentiality of “any and all data exchanged by each member of the consortium.” Participants can pull out of a consortium at any time. Data are to be destroyed either when no longer needed or no later than ten years from the date Cal-PASS first receives them. Consequently, Cal-PASS typically stores a rolling ten-year history of student data.³³

As of this writing, 96 K–12 schools, 44 community colleges, ten Cal State campuses, five University of California campuses, and four private colleges are participating in one of Cal-PASS’s 11 consortia.³⁴ The number of participants differs across consortia. The number of consortia and participants within each consortium has been growing and is expected to continue to grow. So far, only two of the early partners have pulled out of a consortium, reportedly for workload reasons.

Depending on the type of institution, Cal-PASS initially obtains a minimum of five years of longitudinal data and obtains data annually thereafter. For K–12 schools, Cal State campuses, and University of California campuses, it obtains the data directly from the institution, which first encrypts them.³⁵ For community colleges, Cal-PASS obtains extract data from the Community Colleges Chancellor’s Office

³¹ Apart from Cal-PASS, some community colleges, such as San Jose Community College and Santa Ana College in Orange County, have formed data connections with selected high schools in their regions.

³² It may also designate institutions from other consortia with which they are willing to share their data.

³³ This window is in keeping with the State of California protocol for destruction of electronic data. See www.calpass.org/consortia/default.aspx (retrieved 8/2/2007).

³⁴ None of the regional consortium includes all schools, state, and UC campuses in the region. Reasons given for non-participation include workloads, the difficulty of changing to another data system, and a “wait and see” attitude on how the data will be used. See www.calpass.org/consortia/default.aspx (retrieved 8/2/2007).

³⁵ Cal-PASS expects that once CALPADS is up and running, data will be obtained from CALPADS rather than from individual K–12 schools.

of Management and Information Systems (MIS) and then encrypts them. It stores the identifiable information in a separate database with its own security.

Purposes of the System. Designated members (staff and faculty) of the participating institutions and Cal-PASS use the longitudinal data file to track students from elementary school through university to improve educational outcomes. To this end, each consortium is required to establish a “Professional Learning Council” made up of teachers and staff from the participating schools and campuses. These councils use the Cal-PASS data (1) to gain insights into how their respective students are doing and how well prepared they are for the next grade or transition to college, and (2) to identify trouble spots in such areas as course articulation between segments and the transition between high school and college. The councils are then meant to take the next step and design approaches to address the issues they have identified. Eventually, faculty could use Cal-PASS data to evaluate the effectiveness of the changes that have been implemented.

Initial experience with the Professional Learning Councils indicated that their members did not have either the time or the ability to do analysis themselves. To address this problem, Cal-PASS is developing its own analytical capability to support the Councils.

Content: Categories of Information. The content of the Cal-PASS files is the most complete of all the California SUR data systems. Since it receives its student data directly from the participating schools, colleges, and universities, it has access to all the data included in the segments’ data files described above, plus data that may be available only at the school or campus level. In particular, it can access data on individual courses taken and grades in high school and college via participating Cal State and UC campuses.

Cal-PASS maps the Community Colleges Chancellor’s Office courses “TOP codes” to “CIP codes”³⁶ from the federal classification of instructional programs used by every other state except California.

³⁶ CIP is a taxonomic coding scheme of instructional programs developed by the U.S. Department of Education. For details, see <http://www.nces.ed.gov/pubs2002/cip2000/ciplist.asp>.

Coding of the Cal State and University of California courses is not standardized across campuses.

Cal-PASS is currently negotiating with the California Employment Development Department (EDD) to add wage and employment data to its longitudinal files.

System for Tracking Students. Cal-PASS uses a “derived key” for students for whom it does not have a unique ID number. It uses the first three letters of the student’s first and last name, his or her date of birth, gender, and school of origin to create a unique identifier for that student. Reportedly, use of the “derived key” has resulted in a better than 90 percent match. For privacy protection, the data are then encrypted, generating a new ID number that is non-identifiable. Cell sizes of fewer than ten observations are not released.

Current Quality of the Data. The data are reportedly uneven and Cal-PASS does “a lot of first-level screening.” As Cal-PASS gains experience analyzing its data, it is encountering issues, especially with coding of courses and grades. It is seeking to work with the campuses to improve the quality of the data.

Access. Designated personnel of the participating schools, colleges, and universities who have signed a confidentiality agreement have access to Cal-PASS data. Access to the Cal-PASS data file is available only to partners who have agreed to share their data.

Analytical information produced by and for a consortium is primarily for the internal institutional use of the consortium’s members. Consortium members must approve the content and form of any data that may be publicly released. Cal-PASS, however, is expanding a Web site, myCal-PASS, to encourage faculty and staff of member institutions to share with one another what they have learned. Access to any of the consortia’s longitudinal data by outside researchers would require individual approval by each partner in the given consortium.

Staff. Cal-PASS has an IT staff of five (including two contractors) to maintain the data and a staff of three analysts to provide analytical support. Reportedly, it could use twice as many. It has six support staff in addition.

Summary

California’s seven educational data systems differ along many dimensions. For example, they were developed for multiple and differing purposes. Some are used almost exclusively to improve education internal to the segments; others are used primarily as tools to produce aggregated reports for the state and the federal governments.

The internal use of these data systems for research purposes has been limited (perhaps with the exception of the Community Colleges Chancellor’s Office). Some reasons stated are

- lack of staff time
- lack of training in how to use the data at the campus level
- lack of access to other datasets that would be useful for research purposes
- not seeing a clear benefit to additional research with current data.

External researchers rarely use the systems. A main reason for limited outside research is that all of California’s education segments typically interpret state and federal laws on student confidentiality conservatively. The Community Colleges Chancellor’s Office has been more open than other segments to letting outside researchers have access to its non-identifiable student-level data.

All of the data systems include information on student sociodemographics, enrollment, cumulative units earned, GPA, degrees awarded, and financial aid,³⁷ although they differ on the specific data elements included in these categories. Except for Cal-PASS and the Community Colleges’ COMIS, they do not include individual courses taken and grades; COMIS does not include test scores. Even when systems include similar data elements, the definition of these elements may differ across segments.

The accuracy of the data differs across data elements also. Overall, managers and selected users of the data rated the quality of the data as

³⁷ The Cal State file contains a financial aid flag.

good for elements whose funding depended on reporting. They rated the quality of data from other elements as uneven and, in some cases, poor. Individual schools and campuses place little emphasis on the accuracy of the data collected from elements without funding incentives or any negative consequences for providing poor quality data. Although the managers of the centralized individual SUR files conduct internal consistency audits on the data they receive from individual schools and campuses, they do not have the capacity to conduct local audits. Consequently, the completeness and accuracy of the data collected—both overall and for specific data elements—are unknown for any of the current longitudinal systems.

California's education segments differ on their use of a unique student identifier to track students over time and across schools and campuses. The K–12 CALPADS data file will use the State Student Identification first assigned to students in 2004. It is not yet clear how the SSID would be carried into the higher education system—CSIS is currently encouraging schools to include it on high school transcripts, but there is no mandate to ensure that it is transmitted from K–12 to postsecondary institutions.

Currently, CSU and the University of California include the SSID as an optional field on their applications. In contrast, the Community Colleges have no plans to do so, in part because many of their students are not traditional and will not have been assigned one. All three postsecondary segments use the Social Security number to track students. The 5 to 8 percent of students (depending on the segment) who do not have an SSN are given a campus-generated ID number that is not necessarily carried across institutions, even within the same segment.³⁸ To match its students between K–12 and postsecondary institutions, the integrated Cal-PASS system uses a “derived key,” which reportedly produces a 90 percent match.

The staffing requirements to manage the current separate SUR systems are relatively low, ranging from four to six IT staff members per system. Additional analytical staff may use the data to prepare required reports; answer requests from legislators, the Executive branch, and

³⁸ As many as 80 percent of noncredit students in community colleges have no SSN.

their respective administrative oversight boards; and conduct limited research for primarily internal audiences. The size of the analytical staff differs from three to five. CSIS's large total staff of 46 is due seemingly to the fact that it provides technical and other information-technology support to 1,265 LEAs and over 8,000 schools throughout California.

Major Challenges and System Design Issues

In addition to the embryonic architecture that would enable it to develop and maintain a K–20 longitudinal SUR data system, California has the necessary technical expertise. Our respondents indicated that there are no major technological barriers to developing such a system. However, our respondents suggested that a number of important issues would still need to be addressed first. These range from matters of educational governance and funding to issues of legal authority, individual privacy, administrative authority, access, and content; resolution of these issues is needed to overcome the reluctance of the California quasi-independent educational segments to let others use their student data. Capability aside, the climate that might make the development of a California K–20 SUR feasible is not currently in evidence.

We discuss these issues in this chapter. The first section discusses a set of major challenges—primarily issues of support, governance, and cost. The second section discusses a set of system design issues, including content, standardization, access, and student privacy.

Major Challenges

Overcoming the Culture of Independence and a Protective Mindset Among the Stakeholders

According to our respondents, the four governance segments are powerful, relatively independent, and have developed strong separate cultures and identities. They do not have a tradition of collaboration and are protective of their independence and control of the data they let out. As

a result, each segment has developed its own policies and administrative practices, largely independently. Few efforts at coordination across segments are being made. For instance, even though State Superintendent O'Connor created a 64 person K–20 council, this forum has been lacking focus, according to an informed respondent. Also, an Intersegmented Coordinating Committee of the California Education Roundtable was formed with representation from all four segments. But this committee has no real power to make decisions. One of our respondents put it as follows: “The segments are not siblings, but more like neighbors who do not talk to each other or rarely do so.”

With the possible exception of the Community Colleges Chancellor's Office, the segments are suspicious of outside organizations evaluating their systems, apparently believing that these organizations may intend to embarrass them rather than to provide facts and information. “There is a lack of trust” said one of our respondents. A couple of the higher education segments mentioned instances in which publicly released information had embarrassed them because, in their opinion, it had misrepresented the performance of their institutions. A few respondents pointed to instances where “erroneous” findings were made public, requiring that the segments dispute the findings and causing embarrassment.

The segments are also concerned that outside evaluators do not understand their data in part because data elements are not standardized across campuses and segments and in part because some data elements may not be reliable. As a result, outside people may draw the wrong conclusions from the data. The tendency to be protective and independent will need to be overcome before the systems can move forward.

Gaining Support and a Consensus from Stakeholders for a K–20 SUR File

Under these circumstances, getting all stakeholders to agree on the desirability and on an approach to developing, accessing, and using a California K–20 student data system will be difficult. A legislative or executive mandate would go only so far. In the end, cooperation across segments and stakeholders will be required and it can be gained only

over time. Experience in other states suggests that cooperation will come only if all stakeholders see the eventual benefits to using an integrated data system. A majority of our respondents similarly emphasized the importance of showing the eventual long-term benefits of such a system to obtain buy-in.

Assessing the support that might currently exist for developing an integrated K–20 longitudinal SUR system is inevitably speculative. Generally, respondents in the California legislature, the legislative office, and the governor’s office took a positive view of such a system. But they qualified that their level of support would depend on the use made of the data, its costs, and its benefits. Our respondents emphasized the following potential benefits:

- better articulation between the segments
- more appropriate allocation of students across the segments
- potential cost reductions.

With few exceptions, respondents also thought that the business sector would be supportive and that teachers and faculties would not oppose it.¹

Likewise, the leadership of California’s four education segments generally also expressed support for the development of a K–20 SUR database.² But the staff members directly overseeing and operating each segment’s current individual student data systems were more cautious. Some were skeptical of the benefit from it, especially in the short run. Staff members at the California State University and the University of California campuses saw themselves as being at the end of the educational process, the success of their institutions depending primarily on how well students had been prepared by the K–12 system and the com-

¹ Although they would probably oppose it if the system permitted linking students with teachers and faculty.

² By the time this report was completed, we had been unable to schedule a meeting with the President of the University of California or his designee.

munity colleges. As one respondent at the University of California put it, “we are at the end of the food chain.”

Individual staff members interviewed suggested that support from their segment might depend on who would operate the system and who would have access to the data. Individual staff members also felt that the eventual development of such a system was inevitable, as pressures for accountability continued to increase.

Although gaining the broad support of and a consensus among stakeholders in California is a sizable challenge, there are some positive signs. The rapid increase in membership in the consortia served by the fully integrated Cal-PASS data system (see Chapter Two) and the success of that system show that there is growing interest at the ground level—from schools and campuses and their faculties across the state’s four education segments—in using data-driven analysis to enhance student preparation and improve course articulation across segments.³ Cal-PASS’s use of Professional Learning Councils to define the questions to be addressed ensures that this data system directly benefits each participating consortium and its respective institutions.

Cal-PASS and other states’ experience with integrated K–20 systems indicate that to gain institutional cooperation, the K–20 system must be designed consciously to make data useful to schools and individual colleges and universities. In Florida, for instance, teachers and administrators have access to and frequently use the data. Florida also uses a quid-pro-quo system for negotiating with researchers who request the state’s education data. In exchange for access to data, the research questions have to be of interest to the state (or fees are assessed for acquiring the data) and “in-house” staff must be allowed to review the findings before they are released. Texas provides another example of a data system that meets the needs of educators: There, school staff appreciate the fact that they can find out whether students who have dropped out are enrolled at another school or have gone to a community college. In Kentucky, individual institutions can also access the system and get comparison data with other schools, colleges, or univer-

³ Multiple respondents attributed the success of Cal-PASS to its policy of not making its reports public (see Chapter Two for details).

sities. At Cal-PASS, the Professional Learning Councils are critical in determining that the data are used in ways that benefit the participating institutions.

To gain cooperation, states now developing such systems are making early and frequent use of advisory committees that include representatives from stakeholders in addressing specific system design issues.

Assigning Responsibility for Operating the K–20 Data System⁴

FERPA does not permit the release of student data beyond a third party. Under this federal law, students are the first party (the owners of their data) and school districts, colleges, universities, the California Department of Education, and the administrations of the community colleges, state universities, and University of California are the second party (responsible for maintaining their respective student data).⁵ FERPA allows sharing of identifiable education records from second parties to third parties only under a narrow set of exceptions and third parties are not allowed to re-disclose identifiable data to anyone else.

These limitations present a challenge for developing a K–20 longitudinal dataset because of the four-way fragmented nature of education governance in California. In this system, global responsibility for a complete set of longitudinal data rests with four entities: the California Department of Education, the California Community College Board of Governors, the California State Board of Trustees, and the Regents of the University of California (First Amendment project, p. 16).

There are four options to potentially address this fragmentation, each with its pros and cons.

Create an anonymous K–20 longitudinal dataset. An anonymous file would contain no personally identifiable information. Such a system could be operated by any state, local, or private institution. Under

⁴ This section and other sections that refer to FERPA and state privacy laws draw heavily from the following two legal analyses of California and federal statutory and regulatory schemes pertaining to the release of educational data to researchers: (1) Greene and Perkins (2007), and (2) Coleman, Palmer, and Winnick (2006). We recognize that the interpretations contained in these two analyses may be more liberal than others might be.

⁵ There is no unanimity on this interpretation of the law among the segments.

FERPA, anonymous data are disclosable without consent.⁶ Developing such a system would require the creation of a unique random identifier that would be neither the Social Security number nor the SSID number but could be linked to an individual student by a linking key that would itself be protected from disclosure by FERPA.

A drawback is that each of the four education segments would have to run the algorithm that creates the unique anonymous identifier for each individual record.⁷ An advantage is that the resulting anonymous file could be administered by any existing state and local agency and even a private institution. It could be accessible to anyone, including researchers not affiliated with any of the four segments.

Although this option could be implemented under a voluntary agreement between the segments, it is likely that the needed cooperation would have to be mandated through either legislative action or an Executive Order.

Have each of the four segments continue to operate a stand-alone data system and share their respective longitudinal student data with each other. In other words, each segment would have its own integrated, identifiable K–20 data system. A drawback of this option is that each segment would become a third party to the other three, meaning that the data received could be analyzed only by the segments themselves and could not be re-disclosed to fourth parties such as unaffiliated researchers. The latter could have access to anonymous data from any of the four segments.

Designate a new or existing state institution that would take responsibility for the K–20 longitudinal data on behalf of each of the four segments. Under this arrangement, the designated institution would be considered a second party acting on behalf of each segment. This option would most likely require legislative action to authorize an existing or new state institution to play this role.

⁶ However, information cannot be released if any individual might be identifiable from the information. This is discussed under Student Privacy, below.

⁷ It is planned that the CDE will receive anonymous data files and that school districts will be responsible for maintaining the cross walk between the SSID and student identifiable data.

The advantage of this option is that the identifiable data could then be made available to third parties within the limitations of FERPA. A potential model for such an option is the Florida Education Data Warehouse (EDW).⁸ Similarly, Kentucky's state legislature is currently considering a bill that would authorize the development of such an education data warehouse.⁹

Change the system of governance of the California education system and place all four education segments under a single K–20 form of governance. This would be the most drastic option. It would provide the most flexibility in terms of developing and using longitudinal data, but it would clearly require legislative action that is unlikely to be feasible, at least in the short or medium term. As one respondent said “you cannot put the genie back in the bottle.”

Both options 1 and 3 offer the advantage of having a single integrated K–20 data system instead of four separate ones, but both would require that an existing or new organization be given the responsibility of operating the system. Any of the segments or a university or college within a segment could potentially house the K–20 data system, but, according to some of our respondents, it “would be difficult.” Similarly, our respondents felt that eventually adding this responsibility to CPEC would not be desirable without first clarifying that organization's role and considerably augmenting its capabilities.

States house their K–20 and other SUR in various types of institutions. The Florida K–20 Education Data Warehouse is lodged in the Accountability, Research, and Measurement Office of the Florida Department of Education.¹⁰ In Florida, governance of the public K–20 education system is consolidated within the State Board of Education. In Texas, the PK–20 Public Education Information Resource (TPEIR)

⁸ <http://edwapp.doe.state.fl.us/doe/>.

⁹ Interviews with state staff, April 24, 2007. The Kentucky Department of Education, Post-secondary Council, and Education Professional Standards Board submitted a proposal for a “P-16 Seamless Data Warehouse Project,” seeking an initial \$3 million in the 2006–2008 budget period, to be followed by \$4 million in the subsequent budget period. However, the proposal has yet to be funded.

¹⁰ <http://edwapp.doe.state.fl.us/doe/> (accessed July 15, 2007).

is housed in the Texas (K–12) Education Agency (TEA).¹¹ Statutory authority provided the basis for cooperation and coordination of data resources between the TEA and the Texas Higher Education Coordinating Board (THECB). THECB maintains both its own postsecondary data and most of the TEA K–12 data as well as allowing the latter to do its own analyses. By contrast, Kentucky is considering eventually establishing an independent data repository for its P-20 warehouse.

Assigning Decisionmaking Authority for the K–20 Data System

Given the independence of the administration of the four California education segments, shifting some of the decisionmaking authority on data system design issues to another party will be controversial. Making a decision about which organization should be responsible for developing and maintaining an integrated K–20 SUR data system is different from assigning an entity the decisionmaking authority to address and resolve those issues that the legislature either has or may have left unaddressed (see Chapter Four). These issues may range from developing and overseeing the process of obtaining access to the data to addressing ongoing FERPA related concerns, determining the common set of data elements, standardizing data elements, defining data auditing requirements, and making IT decisions.

Currently, the only mechanism to address this kind of issue across the four education segments is the IT Inter-segmental Council. IT staff members from each segment are represented on this unofficial council. However, our respondents told us that because it has no decisionmaking authority, issues rarely get resolved, and if they do, it takes a long time.

Other states with K–20 linkable or integrated K–20 student data systems are using different forms of governance to make decisions about them. These range from a single K–20 form of governance, as in Florida's State Board of Education, to a coordinating board with representation from each of two state education segments, as in Texas and Kentucky.

¹¹ <http://www.texaseducationainfo.org/tea.tpeir.web/tpeirpage/index.html> (accessed July 15, 2007).

Costs

Legislative as well as other respondents expressed two concerns about the costs of developing a K–20 SUR for California. First, experience with developing other type of data systems in the state had shown that initial estimated costs tend to be exceeded. Second, system development may impose additional requirements on individual institutions, which schools and community colleges may consider as unfunded mandates. The costs of developing a California K–20 student data system will depend on whether

- it relies exclusively on data currently available in existing centralized student data systems in each of the four segments
- it adds data that are already being collected by individual schools, colleges, and universities
- it will require new data collection.

In the first instance, the incremental costs of creating an initial K–20 SUR data file by merging the K–12 CALPADS with the three existing postsecondary SUR systems would be relatively low. Because the system would rely on data already being collected, there would be no additional costs for data collection. The main expenditures would be for staff and operations of the “warehouse.”

All of the entities currently maintaining or developing SUR databases manage their systems with a relatively small staff, typically four to six IT staff, up to five analysts, and four to six administrative support staff. All indicated, however, that they could use more staff, especially analysts.¹² The Florida warehouse has a staff of 11 programmers and analysts to maintain and analyze its K–20 student data system.¹³ Hence, it can be inferred, given California’s larger student population, that integrating the four California educational segments’ data systems as they now exist (or will exist, in the case of CALPADS) would

¹² CSIS has a staff of 46; however, this includes staff to train and help schools in their data collection and transmittal activities.

¹³ Email communication from the warehouse, on August 11, 2007.

require a somewhat larger staff—say ten IT staff, ten analysts, and six support staff for a total of 26.

To the costs for staff should be added expenses for leasing a secure facility to house the data system. Cal-PASS currently pays about \$13,000 a year to lease such a space. One time costs for hardware and software would also be incurred. We estimate a need for ten to 15 servers and software at \$9,000 per server, backup hardware at \$14,000–\$20,000, storage racks at \$10,000–\$15,000,¹⁴ and the cost of furniture.

In total, a K–20 SUR dataset formed by integrating existing longitudinal data files would require an estimated one-time cost of up to \$500,000, plus annual operating costs of about \$4 million to 5 million annually.¹⁵ This latter estimate, even if doubled, is not a large sum relative to the more than \$45 billion spent annually on education in the state of California.¹⁶

But as tempting as it may be to minimize startup and operating costs by simply merging existing longitudinal data files, the resulting integrated K–20 SUR system would not provide the maximum possible benefits. Ideally, the system should be expanded to include data already being collected at the campus level but not currently included in what schools or campuses forward to the central data system of each educational segment—for example, data on courses taken, grades, ACT and SAT test results, and participation in special programs. These data elements could possibly be added at a minimal one-time cost that would cover setting up the processes for transferring the data from the institution-based data systems to the central data system.

Nonetheless, several issues remain. The obsolescence of some of the systems may require more difficult and costly reprogramming; estimating the cost of this was beyond the scope of this study. Another

¹⁴ Based on Cal-PASS recent experience.

¹⁵ Assuming an average of \$150,000 to \$200,000 annual costs per staff.

¹⁶ These estimates do not include the costs of developing the CALPADS and providing schools with the support to upgrade their data collection capabilities. The state is already committed to developing this K–12 individual student data system. The development costs of the Florida data warehouse amounted to about \$7 million over three to four years (Hansen, 2007). In addition, it has a support and maintenance IT staff of 12 and 11 analysts; these numbers are well within the range of the estimates discussed above.

issue is that placing additional data requirements on schools and community colleges (although not on Cal State and the University of California) may be considered an unfunded requirement and, hence, may require additional state funding.

It may take several years before experience with either the minimal or the expanded K–20 student data systems suggests the need for collecting data elements currently not being collected at all. The costs of collecting new data would inevitably be high and a decision to do so would have to be weighed very carefully against the extra benefits it would provide.

System Design and Technical Issues

Technical Feasibility

According to our respondents, the technology for developing an integrated SUR data system in California is available. Indeed, a few other states have already done so. The technological issue is therefore not one of feasibility but one of standardization of the technological systems used by individual schools and campuses and often serviced by different private companies. Some of the current hardware and operating systems—particularly in California’s higher education segments—are old and no longer up to today’s standards. It may not be easy to modify them to add or standardize data elements. For instance, the University of California’s data system still uses a main frame about 20 years old and only a few people know its “in and outs.” The software language it uses, COBOL, is also outdated. Moreover, different systems may have different programming requirements for transferring data.

Given that CPEC and Cal-PASS have been able to create partially integrated data systems within existing operating systems, we believe that these issues need not delay the development of a K–20 SUR data system in California. But it is likely that upgrades and, perhaps, standardization of operating systems across schools and campuses may be desirable. We expect that upgrades and standardization will take place whether or not the state develops an integrated K–20 data file. But if it

succeeds in doing so, that may accelerate the need to upgrade current systems.

Incompleteness and Uneven Quality of the Student Data

Incompleteness, lack of standardization, and unknown reliability of some of the data elements within and across education segments are issues for the development of a reliable K–20 individual student longitudinal data system.

Incomplete Data Elements. Ideally, all four of California’s education segments should collect the minimum set of data elements necessary to describe the educational paths and success of different groups of students and to assess the effectiveness of educational policies and programs. Such a set of data elements would include a student’s

- sociodemographic information
- enrollment status
- reason for enrollment
- school or campus of origin
- individual courses and grades
- participation in special programs including summer school, remedial programs, and high school-to-college transition programs
- results of state K–12 tests, SAT, ACT, and similar tests
- financial aid.

Each segment currently falls short on collecting all of these data elements. CALPADS may compile most of them, including information on individual courses in high school and results on state tests. But at the moment there appear to be no plans to include in its system grades and results of SAT and ACT tests—data elements that can be used to assess students’ degree of preparation for college, their needs for remedial college programs, and their probability of success in different types of colleges and majors.

Within Cal State and the University of California, the most critical information missing is individual courses and the grades obtained in these courses and, in addition for UC, participation in remedial and support programs. This information is available at individual cam-

puses, but because it is not needed for federal and state reporting, it has never been requested, nor is it transferred to the central data files of these two segments.

Our respondents at the three postsecondary segments indicated that because of the costs and staff time that may be involved, they do not want to overburden individual campuses with requests to transfer data already being collected to the central file, and they certainly do not want to ask that they collect new data. In addition, the campuses are somewhat independent, and absent a requirement from their respective oversight boards, the segment of which they are a part would have to convince them that a given additional data element is more important than handling the many ongoing requests for data they receive.

Lack of Standardization. Because the K–12 CALPADS data file currently under development will be a new system built from the top down, it can be expected that the data elements it will incorporate from individual schools will generally follow standardized definitions.¹⁷ However, California’s three postsecondary data systems were developed from the bottom up by hundreds of quasi-independent colleges and universities. Consequently, there is no reason to believe that the same data elements are being collected in the same way according to the same definitions across all the campuses within and across segments. This problem is likely to be more acute for data elements that are not typically forwarded to the central data system, such as in the labeling of courses. However, data elements required to prepare federal and state reports, such as IPEDS, are believed to be standardized across campuses.

Experience in other states suggests that the issue of standardization requires a great deal of attention to ensure the credibility of the data system and the analytical results that may be derived from it. Experience in other states also suggests that standardization is a process that takes time. This may particularly be the case for course content. At the postsecondary level, course content is left to the discretion of the professor: A course labeled Algebra 101 can cover a narrow or a

¹⁷ CDE has been working to standardize data elements across schools for several years (Hansen, 2007).

broad spectrum, and anything in-between, and can be taught at various levels of difficulty. Also, each college and university has the discretion to define what and how many levels of remediation courses they will offer. Standardization of course content and level of difficulty cannot be achieved by decree but only by discussions around articulation of courses between high school and college and between colleges and universities—a process that would most likely take years. In the end, however, variations in the content of courses that carry the same label across campuses will be unavoidable.¹⁸

Unknown Quality of the Data. Just as the extent of the lack of standardization is unknown, so too is the extent of the quality of the available data. Users of the current California data say that information is missing, data elements are inconsistently coded, and the like. Because no one campus uses its respective segment-wide data system and the data it sends rarely comes back to it, the data at the source rarely get cleaned up. Overall, our respondents rated the quality of the data from “good” to “uneven” to “getting better with use.” This is a common issue highlighted by all our respondents in other states.

Experience in other states with developing and maintaining a SUR data file suggests that quality improves as the data are used for resource allocation and other public reporting at the school and campus levels, that is, when the schools and campuses have a stake in the accuracy of the data released. Quality reportedly also depends on how well trained and knowledgeable are the school or campus-level staff collecting and transferring the data. The need for adequate ongoing training was mentioned by respondents in the states we selected for interviews, as well as by our respondents in the California segments.

¹⁸ Florida was able to develop a common course numbering and coding system. It also has standardized its data elements definitions and has a common data reporting calendar. Reportedly, there was initially much resistance to standardization, but over time the system gained acceptance and today is seen as indispensable. Florida has the advantage of centralized governance of its education system, which made the task easier than it would be in California. In addition, Florida’s education system is considerably smaller than California’s. For instance, Florida has 67 districts; California has 1,000.

Coordination of Data Submission by the Segments

The postsecondary segments collect data at the beginning of the year and every term thereafter. This seems to be adequate to capture information on students who start or transfer in the middle of the year¹⁹ and students who may drop out during the academic year. At the K–12 level, current plans are for schools to submit student-level data for CALPADS at the beginning of the school year and thereafter during the school year as deemed necessary by the LEA or by the California Department of Education to prepare its reports. High levels of student mobility in elementary and secondary schools, however, suggest that it might be best to require that schools report students who transfer in or leave a school at the time the event takes place. This practice would help to identify reliable retention rates by institutions and to track transfers from one institution to another so that individual segments and institutions within segments are appropriately credited.

All four segments will have to define the academic year uniformly, possibly from the beginning of the school year for K–12 and from the beginning of the fall term to the end of summer school about one year later for higher education institutions.

A Common Student Identifier

A common student identifier that follows a student from his or her first entry into the California education system to exit would greatly facilitate the merging of student data over time and across all four education segments. Most other states across the country that currently have a K–20 SUR system, or are developing one, use the SSN as the common identifier. California has four options to consider, each with advantages and disadvantages. Regardless of the choice, in practice there is no such a thing as a comprehensive common identifier.

Maintain the use of the SSID for K–12 and the SSN for the postsecondary segments. This is probably the least attractive option. Matching K–12 with postsecondary individual student data would

¹⁹ According to a national estimate, about 20 percent of students do not enter postsecondary education in the fall quarter or semester.

require the development of an algorithm based on data elements such as portion of the first and last names, date of birth, high school of origin, and the like. As we noted above, Cal-PASS is using such a procedure that reportedly results in about a 90 percent match.

Require or gain the cooperation of the three postsecondary segments to use the SSID currently assigned to students attending California high schools.²⁰ Eventually the current K–12 identifier, the SSID, could track students who enter a California college from a California high school or public postsecondary campus. This option would require either that the SSID be included on student transcripts²¹ or that the postsecondary segments require that students provide it on their applications. Either approach would call for changes in current practice and has drawbacks. The first alternative would be the most efficient and reliable. But the legislature would have to mandate that the SSID be placed on high school transcripts.²² As of now, no decision has been made to do this.²³ The issue is being discussed, however, in the informal Inter-Segmental Data Coordinators Group among IT specialists from California’s four education segments.²⁴

Also, unlike CSU and the University of California, California’s community colleges do not require high school transcripts from their students. This is unlikely to change because many of their students

²⁰ Florida uses a unique K–20 state identifier other than the SSN. Since 1999, it also has been using the Florida Automated System for Transferring Educational Records (FASTER), which allows the automated transfers of transcripts across Florida institutions (Data Quality Campaign, 2006a).

²¹ As noted above, CSIS sends electronic transcripts to UC for the Eligibility in the Local Context program. These transcripts include the SSID.

²² Such a requirement might be considered an unfunded requirement by schools and community colleges.

²³ A number of districts and schools now put the SSID on their students’ transcripts voluntarily and, starting with 2007, CALPADS will require that LEAs that accept funds help them in the transition to CALPADS. However, in other districts that do not receive CALPADS-related state funds, such a requirement might be deemed an unfunded mandate.

²⁴ This unofficial council has representatives from all four education segments. It meets three to four times a year to exchange information on IT activities across the segments; however, its members do not have decisionmaking authority.

take noncredit courses; that is, they do not pursue an associate degree or intend to transfer to a four-year university.

The second approach would require that a student be given an SSID and report it. Currently, CSU and UC have added a field on their respective applications to collect the SSID voluntarily.²⁵ The Community Colleges Chancellor's Office is considering doing so. But at California State University, at least, there are no plans to transfer students' SSID numbers to its central ERS file.²⁶ Moreover, our respondents indicated that this approach may be unreliable because students may not know their SSID or may refuse to report it. Furthermore, because the students may not know their SSID, it may be necessary to implement in the postsecondary segments a similar algorithm to locate student's SSID based upon the student's name and demographic information, as CSIS has done.

In this option, the SSN would continue to be used for the large share of students entering from either private schools in California or from postsecondary segments in other states. Another ID would have to be assigned to foreign students as well.

Use the SSID for California public school students and also require that an SSID be assigned to students who enter one of California's postsecondary segments from in-state private schools or postsecondary institutions, from another state, or from abroad. This approach would ensure that a common and unique identifier is used for all students in the California education segments. However, it would still be necessary to collect the SSN at the postsecondary levels to track students as they enter the labor market and use government programs.

Require use of the SSN as the common identifier, including at the K-12 level. Doing so would be particularly useful for the eventual linking of students' educational experience with their outcomes in the labor force including those of high school dropouts and students that do not continue to college. The drawback of this option is that it would require reverting to an approach that was rejected by the legislature

²⁵ It is not known how often students fill-out this SSID item.

²⁶ CSU's campuses started doing so for academic year 2007-08.

when the latter required that a random number be used for identifying and tracking K–12 pupils. It would also require that postsecondary institutions allocate another ID for foreign students.

Determining Who Should Have Access to the Data

California’s education segments disagree about what student-level data may be released to whom and how. Each segment has its own counsel responsible for protecting it from potential lawsuits. Fear of legal action, some of our respondents said, was a primary reason for the “conservative” interpretation of the federal and state student privacy statutes by some of the segments.

Currently, designated staff members from each segment have access to their respective individual student data for the purpose of developing reports and conducting statistical and education improvement analyses. In contrast, in the K–12 segment, non-designated staff or researchers either within or outside the California Department of Education, CSIS, and Cal-PASS do not have access to individual student data,²⁷ even if non-identifiable. Although at the postsecondary level, the three segments have procedures to allow outside researchers access to their respective SUR data under specified circumstances, only the community colleges allow access to their data more often than just in rare instances.

The legislation that authorized CPEC to integrate the student data from California’s three postsecondary segments is largely silent on who should have access to the data, saying only that it should be available to the “educational segments” (AB 1570, Chapter 916, Sec. 2,m1 E). But in the legislation designed to guide the development of the K–12 CALPADS, the California legislature specifically declared its intent that CALPADS should be accessible not only to school districts, charter schools, and state agencies with responsibility for education, but also to legislative policy analysts and researchers:

²⁷ It is not uncommon for school districts in California and other states to allow access to their longitudinal individual student data to outside researchers if the proposed research meets the FERPA’s exception requirements as discussed further below. The data provided to outside researchers are nearly always non-identifiable.

It is the intent of the legislature to . . . establish state data management practices that require the development of specific categories of users and uses for pupil data and establish responsibility for establishing and servicing users as well as responsibility for establishing and posting protocols, criteria, and procedures for use that are developed in a manner that is consistent with recommendations of the State Department of Education's advisory committee on privacy and data protocol. Approved users should include schools, charter schools, state agencies with responsibility for education, legislative policy analysts, evaluators of public school programs, and education researchers from established research organizations.²⁸

Making student data and information widely available and providing feedback to faculty, administrators, and legislators is in the interest of public transparency, efforts to improve education, and assessments of the effectiveness of education programs. There appear to be no barriers to doing so in federal FERPA and California state legislation, as long as the individual data provided are non-identifiable. For most analyses of educational policy, improvement, and program effectiveness, anonymous individual student-level data are all that are required. Of course, any users of the data would still have to meet the FERPA requirements. In particular, they would need to take the necessary measures to ensure individual student privacy (discussed in "Student Privacy," below).

Under the FERPA "studies exception," personally identifiable data may also be provided to a broad set of third-party users, including researchers of state and local educational institutions. An educational agency (or institution acting on behalf of an educational agency) may release personally identifiable data to an organization conducting studies on its behalf. However, the purpose of the studies must be specifically to "develop, validate, or administer predictive tests, administer student aid programs, and improve instruction."²⁹ Most interpretations of the FERPA studies exception also allow for independent research to be conducted so long as precautions are taken:

²⁸ SB 257, Chapter 782, section 1 (c).

²⁹ First Amendment Project, p. 14.

This authority may be implemented to include disclosing personally identifiable information to organizations for *independently initiated studies* by promulgating state regulations or guidelines to provide access for authorizing studies initiated by third parties in which the local education agency or school has a clear interest (whether or not it funds the study) and including strong privacy protections against re-disclosure, consistent with FERPA (emphasis added).³⁰

In brief, subject to the appropriate procedural precautions and purposes specified for studies, there appear to be no privacy barriers encoded in national or state law that would prevent third parties from accessing individual student data, identifiable or not—including researchers from organizations outside California’s four education segments. Other states provide examples. Florida, as noted above, allows outside researchers access to its data on a quid pro quo basis, requiring that their research questions be of interest to the state and retaining the right to review their methodology and findings. Under these terms, Florida has provided access to its data to researchers from Harvard University, the Manhattan Institute, the University of Chicago, and Florida University, among many others.

Texas, on the other hand, is more cautious. The state provides access to its education data primarily by doing analyses on behalf of state users and education institutions. But it has also provided student-level data to researchers from the University of Texas at Dallas and to the state demographer.³¹ In addition, the Texas legislature has recently passed a bill (House Bill 1, May 2006) to provide funding to establish up to three education research centers. These centers will be expected to provide outside researchers, as well as their staff, access to anonymous student-level data.

It is fair to say, however, that states have generally been very cautious in allowing outside researchers access to student-level data.

³⁰ First Amendment Project, p. 15.

³¹ Interview with state staff, April 24, 2007.

Student Privacy

Whether or not an eventual longitudinal K–20 individual data file for California contained identifiable data, FERPA and state law require that precautions be taken to ensure that identifiable data are not released inadvertently. This could happen when the data are transferred from agency to agency or from agency to users, and when results are being reported.

Release of individual grades and student test results, even without student identifiable data, is of particular concern because of potential litigation. Indeed, when Kentucky released student grades to the overseer of its SUR, a FERPA complaint was filed by a school district with the federal privacy office. This office reportedly took two years to make a decision, in the end ruling that “FERPA could not be used as an excuse by individual institutions for not providing the data”³² because the overseer of the SUR was a state authority.

The technology and procedures to protect student privacy are well established. Those in California who maintain or use the data files described in Chapter Two know and apply them:

- Data are encrypted before being sent from one agency to another.
- Access to individual data is accessible only by using a password.
- When the data are in use, student identifiable data are kept separately from the analytical file (and in a locked cabinet).
- When reporting data, care is taken not to report data for cells that contain few observations or that otherwise might lead a reader to identify a specific person. The segments have set different rules in this regard.

Even these precautions do not guarantee that identifiable student data will never be accessed by unauthorized individuals. Precautions to minimize this type of incident require trained personnel, adequate funding, and continuous attention to security procedures, with diligent application and frequent updates.

³² Interview with state staff on April 23, 2007.

Summary

Many interrelated issues will have to be addressed before an integrated K–20 SUR data system becomes a reality in California. The culture of independence and protective mindset among the four public education segments will need to be overcome, and a consensus that building such a system is desirable will have to be developed. Similarly, developing a consensus regarding who should have decisionmaking authority over the development and maintenance of the system and who should operate it is likely to be difficult as it will require shifting some decision-making authority on these matters away from the overseers of the four California public education segments. And given the state budget woes and its experience with cost overruns in the development of some previous state data systems, providing the necessary funding, even though minimal compared to the billions spent on education in California, may well be a roadblock.

In addition, a consensus on various design issues will have to be developed, ranging from content to standardization of data elements across segments improving the quality of the data being collected, selecting a common student identifier across segments, defining who should have access, and protecting student privacy.

Although each of these issues is controversial, all can be overcome as they have been in other states that have implemented and are operating K–20 (and beyond) student data systems.

What Next for California?

We conclude that developing an integrated K–20 SUR data system in California is feasible. But many interrelated issues—from major challenges concerning cooperation among stakeholders, governance, and system “ownership” to questions of accessibility, data content, and standardization—need to be resolved, with divided stakeholders reaching a consensus. This will be a daunting task that may take several years.

In our view, several steps will need to be taken before these issues can be resolved, the initiative can move forward, and maximum use can be made of the newly available data:

- Complete the design and implementation of CALPADS.
- Identify a champion to be an advocate for a K–20 SUR data system in California.
- Obtain legislative authority.
- Begin building the K–20 data system incrementally.
- Develop an “objective” analytical capability and expertise.

Complete the Design and Implementation of CALPADS

The first step toward eventually building a K–20 student data system for California is, of course, to complete the design and implementation of CALPADS. Much remains to be accomplished before the system is operational. The Department of General Services (DGS) has yet to select a vendor for developing the system, for which a request for proposal was released on September 8, 2006. Development work is sched-

uled to be completed by the end of 2008 and the system is to become operational during the 2009–10 school year.¹ However, lack of funding may delay implementation of CALPADS. Although the DGS has received a federal grant of \$3.25 million to develop and implement the system, additional state funding will be required to help smaller LEAs develop sustainable data management practices. Funding for this purpose was deleted from the state 2007–08 budget.

Identify a Champion to Be an Advocate for a K–20 Data System in California

We believe that getting the process started will require an influential individual or group of individuals to champion the development of an integrated K–20 SUR data system, to gather the necessary support from stakeholders and the public, to develop a consensus on the system’s governance and design, and to sponsor the legislative action that would mandate and define its capabilities and limitations.

Several of our California respondents held the view that if a K–20 student data system were to be developed in California, it would need to have this sort of strong and authoritative advocate. Potential candidates include the governor, a well connected business leader, or a legislator. Some suggested that if a legislator became the advocate, he or she should have been elected relatively recently, and therefore would not be about to hit the term limits, and should be coupled with a more senior and powerful figure. Alternatively, the case for a K–20 SUR could be championed by a commission that might include a representation by committed stakeholders and potential users.

Obtain Legislative Authority

Even though new legislative action is not needed to develop and maintain a K–20 SUR (see the previous chapter), our respondents were unanimous in their view that such legislation would be required,

¹ <http://www.cde.ca.gov/ds/sp/cl> (accessed on August 6, 2007).

because of the fragmented nature of California’s education system, the traditional lack of cooperation between its education segments, and disagreement among stakeholders on most of the issues outlined in this report. Many of these issues are sensitive and important and need to be aired, debated, and eventually resolved in an open public debate. It is not likely that the disparate views about governance of the data system (i.e., who has decisionmaking authority and who owns the data system) and access to the data could be resolved without legislative authority. Experience in other states, including Florida and Texas, suggests that legislative action or an Executive Order has typically paved the way to guaranteeing the necessary cooperation.

The California legislature would need to address the following questions:

- What purposes would a K–20 SUR data system primarily serve?
- Who should have decisionmaking authority over the design of the data system?
- Who should operate the data system and where should it be located?
- Who should have access to the data it contains?
- What common student ID should be used?
- What should be the minimal content of the data file?
- What level of funding should be allocated to develop, maintain, and use the data system?

In this report, we have presented a number of possible answers to these questions—various options that the legislature may consider, some that would lead to the development of a K–20 SUR system that differs in content, access, and other characteristics.

Begin Building the K–20 Student Data System Incrementally

Once CALPADS is operational and the legislature has set the ground rules and provided its guidance, the building of the K–20 data system itself could take place incrementally. Fully developing a K–20 data

system that maximizes its utility for policy and administrative decisions at the state and at the segmental and individual institution levels will take several years. But in the immediate future, central files from all four segments could be integrated as soon as the K–12 CALPADS file is operational in 2009. Even this simple merger of the three current postsecondary files with CALPADS would provide useful and more accurate information than is currently available on dropout and transfer rates, student mobility, and educational progress. However, it would stop short of including all the data elements—course and grade information at the higher education level, grades for individual courses at the high school level, SAT and ACT test results—that may be desired for maximum benefits. For the most part, these additional data elements, needed to address issues of articulation across segments and questions of outcomes and effectiveness, are already being collected at the school and campus levels and could be added later.

Linkages to education-related files and post-high school and higher education data files, such as those containing employment and wage, health, welfare, foster care, prison, military and similar data, could also be added later after some experience using the integrated data system has been gained. Such linkages would allow outcomes to be tracked.

A possible incremental development sequence over a period of four to five years might be to

1. Integrate the four existing segmental data systems “as is.”
2. Add data elements already being collected at the individual school and campus level, but not currently forwarded to the central student database of the respective segments.
3. Link the K–20 data file to other institutional administrative state and federal data systems.
4. Add data elements not currently being collected by one or more of the segments—but only if deemed desirable and cost-effective after several years of experience using the data system.

A long-term perspective will have to be taken toward improving the quality and standardization of the data collected by schools and

campuses. In the short term, those responsible for data collection and entry could receive training to ensure they use appropriate quality control practices. Experience suggests that this will go only so far, however, and that data quality will improve primarily through use and users' feedback to schools and campuses.

A Note on Linking the K–20 Student Data System to Other Individual Systems. The ultimate outcome of education is the extent of the success that students eventually realize in their lives after leaving school by contributing to the economy and society's welfare. Such outcomes have been measured only in the aggregate, if measured at all. The existence of a K–20 student-level longitudinal data file combined with individual administrative records from various state agencies would offer the opportunity to track individuals not only throughout their student life, but also through their adult life. Most administrative records, including data on employment and wages, welfare, Medicaid, foster care, health, and corrections, include the SSN so that merging these data with individual student records presents no issues beyond those already discussed in this report.² Eventual inclusion of data on students in pre-kindergarten would also be eventually desirable.

In addition to linking the K–20 system to the administrative records of the state's agencies, links could be developed with the postal service, the Department of Defense, and the Office of Personnel Management to track California students into the military and federal employment. Another potential future linkage could be made with the National Student Clearinghouse to track students leaving California colleges and universities to attend postsecondary institutions in other states.³ This linkage is particularly important because of the high mobility of the student population across state lines: Among college students, 20 to 25 percent will cross state lines and one out of five

² California state laws restrict how SSNs may be used. We have not reviewed the relevant statutes, but legislative amendments may be required to enable the kind of individual data matches described in this subsection.

³ The Florida student data "warehouse" links its student data with data from its various state agencies, the National Clearinghouse, the Department of Defense, and the Office of Personnel Management.

matriculating in one state will have attended a college or university in another state (Edelman, 2007).

Including California's preschools and private colleges and universities in the state's K–20 SUR data file would also be desirable. Their participation would have to be voluntary. In some states, private post-secondary institutions have found it in their interest to participate and to have access to their state's SUR. For instance, Kentucky private educational institutions participate because they receive state aid and can use the state SUR data system to provide the information required to receive that aid. In Texas, private institutions participate because they receive state equalization funds and in exchange are required to produce graduation rates.

Just as legislative authority would be needed to integrate the various segmental student data systems in California, it may also be necessary to secure the cooperation of the state's agencies. Several in the education segments said that they had attempted occasionally to make linkages with other agencies, such as the Employment Development Department, to obtain employment and wage data, but were unable to secure cooperation. They attributed this failure to the same privacy issues discussed in the preceding chapter.

In the long term, it may also be desirable to integrate the K–20 SUR with existing data about individual teachers,⁴ faculty, and institutional financial information. Integration of these data with student data raises union and confidentiality issues that could not be explored in this study.

⁴ A parallel longitudinal data system for teachers, the California Longitudinal Teacher Integrated Data Education System (CALTIDE), was authorized by the legislature in 2006 (SB 1614). As with K–12 students, teachers are to be assigned a unique California ID. CALTIDES is meant to contain information on teacher credentials including teacher preparation program, certification including NCLB highly qualified status for specific subject matters, selected demographic characteristics, course assignment, and in-service education and staff development. It will not include teacher name, SSN, or other information that would allow individual teachers to be identified. Although the legislation specifies that the data “may not be used, solely or in combination, with data from CALPADS, for purposes of pay, promotion, or personal evaluation of an individual teacher or groups of teachers . . . (SB 1649, Sec. 3.c),” it does not seem to prevent its eventual use in combination with CALPADS to assess the effectiveness of teacher preparation and development programs.

Develop an “Objective” Analytical Capability and Expertise

Making maximum use of the K–20 student-level data and effectively responding to the needs of individual schools, colleges, universities, and state officials would require the development of an analytical capability, adequately funded, either within the organization operating the K–20 system or independently. Currently, California’s education segments individually (with perhaps the exception of the community colleges) have only a limited internal capability to analyze their own student data. They use their own data primarily for descriptive and administrative purposes. Similarly, professors within the segments and, certainly, outside researchers have only rarely made analytical use of individual student data. Even though the higher education segments have processes for researchers to request student-level information, they are cumbersome and have rarely been used. Several of our respondents indicated the need to develop an adequate analytical capability in parallel with the K–20 data system itself. As one respondent said, “California has data, but not the information needed for policy decisions.”

States that have developed K–20 (or beyond) student data systems have recognized this need. In Texas, for instance, the Houston Endowment has provided a \$2.6 million, four-year grant to the agency housing the Texas K–20 SUR to create an internal “think tank.” In addition, Texas House Bill 1 (2006) establishes and initially funds up to three Education Research Centers in the state. These centers will have access to the K–20 SUR data and will be expected to provide outside researchers access to it as well. The centers will receive state funding in the first year and, thereafter, will be expected to raise their own funds through grants, contracts, or other fundraising activities.⁵ In contrast to Texas, Florida has developed an analytical capability in its K–20 Education Data Warehouse, which has a staff of 11; in addition, Texas provides access to its K–20 data SUR to outside researchers under a specified set of rules (see Chapter Three). Also, Florida reports

⁵ State staff interview, April 24, 2007.

heavy use of its warehouse by teachers and faculty as well as by districts' staff.

Developing a capability to analyze the longitudinal K–20 student data in California would also address one principal concern of respondents from the state's higher education segments: the possible misinterpretation of their data. Such misinterpretations or analytical errors can often be prevented. They may occur because analysts may be unaware of variations in the definitions of data elements across colleges, misunderstand data elements' definitions, or simply make a processing or operational error. Misinterpretation of the data may also occur if a change in policy or practice in one segment that affects another segment has taken place unnoticed by the analysts and, as a result, they attribute the change to the wrong cause. A sophisticated analytical capability and rigorous professional training will significantly reduce the frequency of such problems.

Experience with CPEC and Cal-PASS suggests that even five or so qualified analysts would not provide the needed capability. Ten to 20 analysts in total might be required to provide the level of analytical capability and expertise desired.

Is There a Better Alternative?

If California wants to track all students in the state from kindergarten through their college years and perhaps beyond, is there a better or cheaper way of achieving this goal than developing a SUR data file that cuts across all of the state's education segments?

Some of our respondents suggested that a more efficient alternative would be to collect data directly from cohorts of samples of students. The concept is to draw a sample of students at entry into the system (say in kindergarten) and to follow this sample over time until they leave school or college and, thereafter, follow them into the labor force. Information would be collected via direct surveys of students or their parents and from the administrative records of the schools and colleges these students attend. The U.S. Department of Education is

conducting such cohort surveys at the national level, although they usually start with a cohort entering middle or high school.⁶

The main advantage of this approach is that information not typically kept in administrative records—information relevant to understanding student choices and behavior—can be collected from the students or parents. Such information may include physical and mental health status, family income and status, student expectations, language spoken at home, and the like.

The main drawback of the cohort approach is that it cannot provide a systemwide view of progress across grades and across time. Take, for instance, the desirability of calculating an accurate retention rate in high school and college every year, or at least every two or so years, to assess whether progress is being made. With the cohort approach, for a cohort starting in kindergarten, the information would not become available for as many as 10 years. And, even then, it would supply a measure of the retention rate only of those who started in the system and not of the many students who entered at any time after kindergarten. This would give a potentially misleading reading of the “true” retention rate.

Conceivably, several cohorts could be started at the same time at different grade levels—say, one entering kindergarten, one entering middle school, one entering high school, and one entering college. Then, cohorts entering at these levels two to four or so years later could be added. This approach would assure more timely information and would capture potential changes in trends over time. However, it would not entirely solve the problem of obtaining systemwide information on key indicators of education progress annually, and it would rapidly add so much to the costs as to make it infeasible. The Department of Education spends \$12 million annually on surveying just one cohort. With four to five cohorts, the costs increase to \$60 million or more annually, not including the costs for analysis. In addition, there are reliability problems because cohort samples become smaller over

⁶ See *High School and Beyond* at <http://www.nces.ed.gov/surveys/hsb>, and the *National Longitudinal Study of 1988* at <http://www.nces.ed.gov/surveys/nels88>.

the years as students move out of the state, cannot be found, or get tired of responding to an annual survey.

Given these drawbacks, the cohort approach to tracking students over time does not appear to be a cost-effective alternative to the development of an integrated K–20 SUR in California. However, it can be a supplement to a K–20 SUR if and when more detailed information about student motivations, behaviors, and circumstances is desired.

In conclusion, development of a K–20 SUR for California promises to offer the most effective way to assess student educational progress over time and to provide the information policymakers, educational administrators, faculty, and the public need to improve educational opportunities for all California's students as well as meet the future labor demand of its economy.

Selected Characteristics of States Selected for Interviews

Table A.1
Characteristics of Student Data Systems of States Selected for Interviews

System Size	Multiplicity of Higher Education Systems	Breadth of Content of the SUR Systems
Florida		
67 county-wide public school districts encompassing almost 4,000 schools in which over 2.6 million students were enrolled in fall 2005.	System of public higher education consists of the State University System and the Community College System. The primary policymaking and governing body for public K–20 education in Florida is the State Board of Education.	The Florida EDW provides a single repository of data extracted from multiple sources on students, education facilities, curriculum, and instructional staff in the P-20 public education system. The EDW allows longitudinal data analysis at the student and staff levels from 1995–96 forward. Student-level data include demographics, enrollment, course completion, assessment results, financial aid, and employment.
Kentucky		
175 public school districts and 1,243 public schools enrolling about 665,000 students in fall 2005.	System of public higher education consists of the State University System and the Community and Technical College System. The Council on Postsecondary Education is a coordinating board for all	Kentucky is developing a K–12 SUR, the Kentucky Instructional Data System (KIDS). The Council on Postsecondary Education maintains a SUR database that includes information on students in all public post-

Table A.1 (continued)

System Size	Multiplicity of Higher Education Systems	Breadth of Content of the SUR Systems
	<p>public higher education in the state, having regulatory and advisory powers over postsecondary education.</p> <p>It is the only entity empowered to present institutional budgets to the legislature. It is not a governing board; Kentucky universities are governed by several multicampus and individual institutional boards.</p>	<p>secondary institutions and all private institutions that are members of the Association of Independent</p> <p>Kentucky Colleges and Universities. Some proprietary schools also house student data there. Student-level data include demographics, enrollment status, financial aid, credits attempted and earned, GPA, and degree awarded. Kentucky has an Education Cabinet that includes both the Kentucky Department of Education, which oversees K–12 education, and the Council on Postsecondary Education. The Cabinet plays a key role in moving a P–20 data system forward.</p>
Oregon		
174 public school districts enrolling about 560,000 students in fall 2005	Oregon’s University System is led by a chancellor who reports to the State Board of Higher Education. The Community Colleges are supervised and regulated by the Department of Community Colleges and Workforce Development, which is overseen by the State Board of Education.	The Oregon Department of Education (ODE), which oversees K–12 education, is in the process of developing a statewide K–12 SUR, called KIDS. The Oregon University System (OUS) has had a SUR data system for about 16 years. The Community Colleges established a SUR system in 1994–05. Each higher education SUR includes information on demographics, enrollment status, financial aid, credits attempted and earned, GPA, and degree awarded. OUS currently has an appropriation request pending with the legislature for pilot testing a PK–20 Education Information System.

Table A.1 (continued)

System Size	Multiplicity of Higher Education Systems	Breadth of Content of the SUR Systems
Texas		
1,227 public school districts and open enrollment charter schools enrolling over 4.5 million students in fall 2005	A single agency, the THECB, provides leadership and coordination of the Texas higher education system. That system consists of six multicampus systems with their own governing boards, three institutions with individual governing boards, and 50 community college districts.	The THECB SUR system contains data from all public higher education institutions, independent colleges and universities, and career schools that offer associate degrees. By legislative mandate, the Texas Education Agency (K–12) and the THECB cooperate in producing data for the TPEIR. The TPEIR is a statewide database that can follow students from pre-kindergarten through high school into public higher education and into educator certification.

Illustrative Interview Protocol

Protocol

Respondents 201, 202, 301, 401, 402

Thank you for your willingness to participate in our study assessing the feasibility and opportunities for developing an integrated K-20 SUR data system. I am XXX and this is YYY with the RAND Corporation. We are conducting this study which is funded by the Hewlett Foundation.

Your participation in this interview is voluntary and you may refuse to answer any of our questions. All information you give us today will be kept strictly confidential. We will not provide or publish any information that identifies you to anyone outside of the study team, except as may be required by law.

We also would like to ask you permission to record the interview under the same strict confidentiality. [If the respondent says no, do not turn on the recording device and place back in the briefcase and proceed with the interview. If yes, proceed with recording with the interview.]

Okay, let's start, can you tell us . . .

A. Currently Link SUR Data over Time?

1. Since when are reliable individual student records kept over time?

What approximately is the proportion of students in your SUR that don't have a SSN?

What proportion of students has multiple identifiers from different campuses?

How do you match individual student data overtime or across institutions when you don't have a SSN for these students?

Are there any plans to use an ID other than SSN in the near future?

Do you expect to continue to provide the SSN to CPEC?

When CALPADS is up and running do you expect the CALPADS student ID to be used then? As a substitute or in addition to the SSN?

In practice, how would you go about getting the CSIS ID?

Are there any discussions about this in the inter-segmental coordination committee?

B. Other Data Included in Segment SUR?

2. Are you currently including or have sought to include into your SUR data system data from other (external) data systems? If so, from which of the following data systems:

- a. Other education segments? *[Ask for each segment; For CALPADS as in the future],*
- b. UI [unemployment insurance]/wage records?

- c. Professors?
- d. Driver's license records?
- e. College Board?
- f. Other?

What data elements are included or *have you sought* from these data systems?

If you were unable to obtain the data you have sought, what were the reasons?

For data received:

How often are the data incorporated into your SUR system?

For what purposes are the data used?

Are there agreements/MOU with each one of these data sources? Do these agreements impose any restrictions on the type of data elements they give you and on how these data elements may be used and by whom?

- a. Why these restrictions?
[Get copies of the agreements]

What internal or external political, legal challenges did you have to overcome to get these data elements?

Who is doing the matching with the other institutions' data elements?

What are the advantages (disadvantages) of this arrangement?

Were you planning to add additional data from other data systems in the future?

If so, from which one(s)?

What data and for what purposes?

What difficulties do you expect encountering in getting this information?

How do you plan to address them?

3. Are you currently providing data from your SUR data system for inclusion in other institutions' SUR data systems? If so, to which institutions:

- CPEC?
- CAL-PASS
- Other education *segments*?
- Other?

If so, what data? How often? For what purposes?

Are there agreements/MOU with each one of these data systems?

Do these agreements impose any restrictions on data elements to be transferred, who can use your data elements, and/or for what purpose can they be used?

- a. Why these restrictions?
[Get copies of the agreements]

What internal or external political, legal challenges did you have to overcome to provide these data?

Who is doing the matching with the other institutions' data elements?

Why was this arrangement chosen?

Have you been asked by other institutions to provide some data elements from your SUR which you have refused to provide?

If so which ones and why?

Do you have access to the data systems to which you provide data [Probe specifically re. CPEC data]

If not, why not?

Would you like to have access? For what purpose? And What would it take to gain access?

If yes, for what purposes do you access the data?

Are the analyses that the receivers of your data make of it useful to you and in what ways? If not, why not?

What would you like to see happening instead?

Do you expect to be asked to share/provide additional SUR data elements from your system in the future?

If so, by whom:

- CALPADS?
- CPEC?
- Other *education* segment?

What data and for what purposes?

Do you expect to provide this information? If not, why not?

Are there data elements in your SUR that would not be shared under any circumstances? What are they?

Let's now focus on your SUR per se:

C. Reliability/Quality of the Data?

4. What is the reliability of the data elements in your SUR data? What is the range of error rates you have within your data elements? What data elements tend to be less reliable?

What are some of the causes of the error rates and what are you doing to address them?

What quality checks do you make? How often?

D. Who Has Direct Web Access to the Data and How Can the Data Be Used?

5. Who has direct Web site access to:

- a. The identifiable data?
 - i. Your office only
 - ii. Selected personnel at the campuses, which ones?
 - iii. Professors or researchers at the campuses
 - iv. Outside researchers
- b. The non-identifiable data?
 - i. Same list as above

Who has to approve access to the individual data and who sets the policy for such access?

What restrictions are there in the use of the data? What are the reasons (political, legal or others) for them?

We hear that there much reticence about sharing data across segments or with researchers? Why is it so?

What guarantees or other assurances would be needed to overcome this reticence?

E. Linking the Higher Education Segments and CALPADS Files

I will now ask you to consider two alternative ways that data from the four California education segments could be linked.

The first alternative would simply allow each segment to periodically (say yearly) merge data elements from the other three segments into their respective SUR, at their discretion. We refer to this alternative as the linkable SURs alternative.

The second alternative would centralize and maintain all segmental SURs into a central location to be managed by a government or nonprofit organization. We will refer to this as the centrally managed California SUR alternative.

6. What do you think of the idea of allowing each higher education segment and eventually CALPADS to link their respective SUR with the others?

[Whether respondent thinks it is a good or not so good idea, probe why]

Hypothetically,

What would be required to make this possible?

Would there any specific legal, political, and/or technical issues that would have to be addressed?

What internal and external group(s) would be likely to support such a linkage?

And which group(s) would be likely to oppose it and why?

Which common identifier would you recommend to permit the linkage to be done?

If a linking capability existed, would you use it? Why? For what purposes?

Currently, is there sufficient statutory authority to allow such linkages to be done?

If not, what would be needed?

F. Integrating the Higher Education Segments and CALPADS?

Taking this hypothetical one step further:

7. What do you think of the idea of eventually integrating all segments and CALPADS into one centrally managed SUR?

[Whether respondent thinks it is a good or not so good idea, probe why]

What would be required to make this integration possible?

Would there be any legal, political, and/or technical issues that would have to be addressed?

What group within or outside your segment would be likely to support such an integration?

And which group would oppose it and why?

What restrictions would you place on data elements to be included?

How would you recommend controlling access to the data?

Who would ideally have access to the identifiable data?

The non-identifiable data?

Who would have to approve access to the individual data and set the policy for such access?

Would anyone have access to the data without your authorization?

Where and by whom would you recommend that the SUR be managed:

AT CSIS, CPEC, in a new governmental organization, in a nonprofit organization?

What would be the advantages of this arrangement? The disadvantages?

Would you actively support the development of this capability?

Currently, is there sufficient statutory authority to allow such a centrally managed SUR to be develop?

If not, what would be needed?

8. Which of the two hypothetical would you prefer, the linking capability across individual segments or the centrally managed SUR and why?

One last question:

9. How many staff do you have to collect the information from the campuses and maintain your SUR system?

APPENDIX C

California Student Record Data Systems

**Table C.1
Common Data Elements and Other Characteristics**

K–12		Postsecondary				K–12 and Postsecondary
CSIS	CALPADS	CCC (COMIS)	CSU (ERS)	UC (CSS)	CPEC	Cal-PASS
Build the capacity of LEAs to implement and maintain comparable, effective, and efficient pupil information systems.	Provide school districts and CDE access to data necessary to comply with federal NCLB reporting requirements.	Eliminate hardcopy reporting (IPEDS). Consolidate data collection. Integrate data.	Provide admission, enrollment data, and other demographic information for each enrolled student each term of the college year.	Provide information to meet the management, analytical, and operational needs of the UC Office of the President related to student enrollment and performance.	Develop an ongoing statewide plan for the operation of an innovative and coordinated system of postsecondary education.	Collect, analyze, and share student data to track performance from elementary school through university and improve success.
Enable the accurate and timely exchange of pupil transcripts between LEAs and to postsecondary institutions.	Provide better ways to evaluate educational progress and investments over time.	Provide longitudinal tracking abilities (also across institutions).	Track student retention/graduation.	Support planning, financial reporting, and analysis.	Integrate policy, fiscal, planning, data, and programmatic analyses about issues concerning education beyond high school to the legislative and the executive branches of California government and to the general public.	Transfer and exchange student academic data for use in studies conducted to improve instruction.
Help LEAs transmit state and federal reports to CDE.	Provide LEAs information that can be used to improve pupil achievement.	Provide flexibility to answer ad hoc requests.	Support regular term reports, IPEDS, state budget requests.	Support reporting to governmental and nongovernmental agencies.		
	Provide efficient, flexible, and secure ways to maintain longitudinal statewide pupil-level data.	Provide common accepted data definitions, and data for research, program evaluation and statewide accountability.	Respond to federal information requests. Support research requirements.			

Table C.1 (continued)

Data Element	CSIS	CALPADS	CCC (COMIS)	CSU (ERS)	UC (CSS)	CPEC	Cal-PASS
Student ID	SSID, which is anonymous. Student demographic elements, such as legal name, gender, and birthdate are used by the LEAs to match student records with the SSID. Social Security number is collected only when voluntarily provided.	Will use SSID (see CSIS).	Social Security number. If student does not have one, the college assigns a unique local identifier.	Social Security number.	Social Security number.	Social Security number. CPEC converts Social Security numbers to unidentifiable ID numbers.	Collects Social Security number and SSID, if available. Uses a derived key based on parts of first and last name, birthdate, gender, and school of origin to create a unique ID for students who do not have either a Social Security number or an SSID.

Demographic Information

Name	Yes	Yes	Partial name	Yes	Yes		Partial name
Age/birthdate	Yes	Yes	Yes	Yes	Yes	Yes (age category)	Yes
Place of birth	Recommended						
Ethnicity	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Gender	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Citizenship			Yes	Yes	Yes		Yes (CC)
Residence		Yes	Yes	Yes	Yes	Yes	
Zipcode			Yes		Yes		Yes (CC)
Primary language	Yes	Yes			Yes		Yes
Parent education level	Yes	Yes		Yes	Yes		Yes

Table C.1 (continued)

Data Element	CSIS	CALPADS	CCC (COMIS)	CSU (ERS)	UC (CSS)	CPEC	Cal-PASS
Attendance status (full- or part-time)	N/A	N/A	Yes	Yes	Yes	Yes	
Enrollment date	Yes	Yes					Yes
Withdrawal date	Yes	Yes					
Withdrawal reason (including graduates and dropouts)	Yes	Yes	Yes		Yes		
Enrollment status (e.g., transfer student)	Yes	Yes	Yes	Yes	Yes	Yes	Yes (CC, CSU, UC)
Gifted and Talented (GATE)	Yes	Yes					Yes
Educational options (e.g., international baccalaureate)	Yes	Yes		Yes			Yes
Entry level mathematics proficiency				Yes			
Grade level	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Program or major	N/A	N/A	Yes	Yes	Yes	Yes	Yes (CC, CSU, UC)

Table C.1 (continued)

Data Element	CSIS	CALPADS	CCC (COMIS)	CSU (ERS)	UC (CSS)	CPEC	Cal-PASS
Degree-seeking status	N/A	N/A	Yes	Yes	Yes	Yes	Yes (CC, CSU, UC)
Admissions basis	N/A	N/A		Yes	Yes		Yes (CSU, UC)
High school of origin	N/A	N/A	Yes	Yes (institution of origin)	Yes	Yes	Yes (CC, CSU, UC)
High school graduation date	Yes	Yes		Yes (year)	Yes	Yes	Yes (K-12)
High school GPA				Yes	Yes		
Prior college attended	N/A	N/A		Yes (institution of origin)	Yes		Yes (CSU, UC)
Courses taken	Yes	Yes	Yes				Yes
College preparatory courses	Yes	Yes		Yes			Yes (K-12)
Remedial courses	Yes	Yes	Yes	Yes			Yes
Summer courses	Yes	Yes		Yes			
Non-credit courses			Yes				Yes
AP and/or honors courses	Yes	Yes		Yes	Yes		Yes (K-12)
Grades obtained			Yes		K-12		Yes (K-12, CSU, UC)
Term data collected			Yes	Yes	Yes		Yes
Term GPA			Yes		Yes		Yes (CSU, UC)

Table C.1 (continued)

Data Element	CSIS	CALPADS	CCC (COMIS)	CSU (ERS)	UC (CSS)	CPEC	Cal-PASS
Term/course units attempted			Yes	Yes	Yes		Total units enrolled
Term/course units earned			Yes		Yes		Yes (CSU, UC)
Cumulative GPA			Yes	Yes	Yes		Yes (CC, CSU, UC)
Cumulative units earned			Yes	Yes	Yes	Units completed	Yes (CC, CSU, UC)
Cumulative units attempted					Yes		Yes (CC)
Transfer credit			Yes	Yes	Yes		Yes (CC, CSU, UC)
Academic standing			Yes	Yes (student standing)			Yes (CC)
Degrees awarded	Yes	Yes	Yes	Yes	Yes	Yes (degree type)	Yes
Awards date	Yes	Yes	Yes		Yes		Yes
Multiple ethnicities flag	Yes			Yes			
Multiple major degree flag				Yes			
Assessment/Testing							
K-12 state tests	Test vendors submit test data directly to CDE	Yes (STAR, CAHSEE, CELDT)		Yes	Yes		Yes (K-12)
Not tested		Yes					Yes (K-12)
ACT				Yes	Yes		

Table C.1 (continued)

Data Element	CSIS	CALPADS	CCC (COMIS)	CSU (ERS)	UC (CSS)	CPEC	Cal-PASS
SAT				Yes	Yes		
Other tests				e.g., EPT, TOEFL			
Other							
Teacher/staff data	Yes—not linked to student file	Yes—not linked to student file	Yes—not linked to student file (but could potentially be linked)			Yes—not linked to student file	
Institution data	Yes—not linked to student file	Yes—not linked to student file	Yes—not linked to student file			Yes—not linked to student file	

**Table C.2
Other Characteristics of the Data Systems and Their Use**

K-12		Postsecondary				K-12 and Postsecondary
CSIS	CALPADS	CCC (COMIS)	CSU (ERS)	UC (CSS)	CPEC	Cal-PASS
Coverage						
263 LEAs, which is 25% of total 1,058 LEAs (as of July 2006); participation in CSIS state reporting system is voluntary.	Not yet implemented; participation of all LEAs will be mandatory.	109 Community Colleges.	23 campuses.	10 campuses.	CC, CSU, UC central systems.	159 elementary schools, high schools, community colleges, colleges, and universities.
Frequency of Data Collection						
Annual	Enrollment, demographic and program participation data at fall enrollment; as often as deemed necessary by the LEA throughout the school year.	Financial aid, student assessment, student award, and some employee data annually, otherwise at the end of each term, three times a year.	Mixed system; census data shortly after the start of each term; otherwise, data are collected in accordance with deadlines timed to support systemwide reporting requirements.	Registration and financial support data are updated quarterly.	Annual (usually in June or July).	Annual (November).

Table C.2 (continued)

K–12		Postsecondary				K–12 and Postsecondary
CSIS	CALPADS	CCC (COMIS)	CSU (ERS)	UC (CSS)	CPEC	Cal-PASS
Authority to Change Data Elements						
Proposed changes to data elements are submitted to the FCMAT Board and the State Board of Education for approval.	To be determined.	Chancellor’s Office in collaboration with all relevant parties; if the changes are federally mandated, the Chancellor’s Office notifies others accordingly.	Changes are made based on discussions with the Chancellor; generally try to get buy in from all campus presidents; if the changes are federally mandated, this office notifies others accordingly.	Collaborative decision – want buy in from each sector (graduate, undergraduate, financial aid, and registration).		Representatives of participating schools/campuses meet annually or as needed to agree on the specific data elements to be shared among the institutions.
Link or Direct Access to Other Databases						
No.	No.	No, but Cal-PASS and CPEC can access COMIS data.	No.	No.	No.	No.
Access to Identifiable Individual Student Data						
Data are not identifiable.	Data are not identifiable.	Authorized Management Information Services (MIS) staff.	Authorized campus-level personnel and authorized personnel in the Chancellor’s Office who have signed confidentiality agreements.	Authorized personnel and officials.	Authorized CPEC data management personnel.	Designated consortium members, staff, and faculty who have signed a confidentiality agreement.

Table C.2 (continued)

K-12		Postsecondary				K-12 and Postsecondary
CSIS	CALPADS	CCC (COMIS)	CSU (ERS)	UC (CSS)	CPEC	Cal-PASS
Access to Non-Identifiable Individual Student Data						
Authorized CSIS staff only; group of participating LEAs approved by the FCMAT board; currently only aggregate student data are released to the CDE.	Will be given to approved users; intends to provide access to CDE, county offices of education, school districts, charter schools, state agencies with responsibility for education, legislative policy analysts, evaluators of public school programs, education researchers from established research organizations.	California Department of Labor, CC, CSU, UC, and Cal-PASS (available through password-protected Web site).	Authorized individuals who need the data to fulfill their responsibilities.	Users may access CSS themselves or they can obtain ad hoc reports and data extracts through the IR&C information retrieval service on a recharge basis; individuals who wish to access corporate data themselves must obtain authorization.	Does not grant access to their data.	Authorized users may access consortium partners' data in accordance with the Memorandum of Understanding (MOU); information produced through this consortium is for internal institutional use; when deciding to release data that are not for internal purposes, each consortium member first reviews the data and determines what data will be released and in what form.

Table C.2 (continued)

K–12		Postsecondary				K–12 and Postsecondary
CSIS	CALPADS	CCC (COMIS)	CSU (ERS)	UC (CSS)	CPEC	Cal-PASS
Reports Generated from the Data						
CBEDS and Language Census (R30) spring submission reports; approximately 30 additional state reports will be transitioned to CSIS over the next several years.	Federal and state reporting.	IPEDS reporting; other regular reporting, e.g., collective bargaining; data Mart; ARCC Project (accountability reporting for community colleges).	Statistical reports and abstracts, academic performance reports, proficiency reports, graduation rates, system-wide summary tables, campus-specific tables, IPEDS (fall enrollment, financial aid survey).	Annual statistical summary of students and staff, student data warehouse, information digest, IPEDS.	Reports and analysis to the legislative and executive branches of California government and to the general public.	None.
Data Quality/Monitoring: Prevent Use of Duplicate ID						
Yes	Yes	No	No	No	No	No
Data Quality/Monitoring: Validation Checks						
Yes	To be determined	Yes	Yes	Yes	Yes	Yes
Data Quality/Monitoring: Audit of Data Source						
No	To be determined	No	No	No	No	No

Table C.2 (continued)

K-12		Postsecondary				K-12 and Postsecondary
CSIS	CALPADS	CCC (COMIS)	CSU (ERS)	UC (CSS)	CPEC	Cal-PASS
State/System-Level Staffing						
46 total staff	Approximately 20 CDE staff will be trained on system maintenance and operations; CDE plans to outsource the maintenance and operation of CALPADS; information seems somewhat conflicting.	5 IT specialists, 4 researchers.	4 IT staff, 5 analysts.	10 CSS staff.	6 IT staff, 5 analysts.	5 IT staff (3 full-time, 2 part-time) (2 contractors doing special projects), and 3 research staff; in addition, faculty coordinators and administrators.
Web Sites						
http://www.csis.k12.ca.us/library/reporting-requirements/	http://www.cde.ca.gov/ds/sp/cl/	http://www.cccco.edu/divisions/tris/mis/dedmain.htm	http://www.asd.calstate.edu/ers/index.shtml	http://css.ucop.edu/	http://www.cpec.ca.gov/OnLineData/DataDictionaryMenu.asp	http://www.cal-pass.org/Documents/CalPassDedJuly2005.pdf

References

Adelman, Clifford, *Making Graduate Rates Matter*, Inside Higher Education. As of April 30, 2007:

<http://insidehighered.com/views/2007/03/12/adelman>

California Department of Education, *California Longitudinal Pupil Advancement Data System (CALPADS): Feasibility Study Report*, Technical Serial Division, Sacramento, Calif., 2004.

Coleman, Arthur, Scott R. Palmer, and Steven Y. Winnick, *State Longitudinal Data Systems and Student Privacy Protections Under the Family Educational Rights and Privacy Act*, November 2006. As of November 27, 2007:

<http://www.hklaw.com/content/whitepapers/StateLongitudinalsDataSystems2.pdf>

Cunningham, A. F., and J. Milam, *Feasibility of a Student Unit Record System Within the Integrated Postsecondary Education Data System (NCES 2005-160)*, Washington, D.C.: U.S. Department of Education, National Center for Education Statistics, U.S. Government Printing Office, 2005.

Data Quality Campaign, *Florida Case Study: Building a Student-Level Longitudinal Data System*, August 2006a. As of November 27, 2007:

http://www.dataqualitycampaign.org/files/State_Specific-Florida_2006_Site_Visit.pdf

———, *Building Student-level Longitudinal Data Systems: Lessons Learned from Four States*, 2006b. As of November 27, 2007:

http://www.dataqualitycampaign.org/files/State_Specific-2006_Site_Visit_Synopsis.pdf

Education Commission of the States, *P-16 Collaboration in the States*, 2006. As of November 27, 2007:

<http://www.ecs.org/clearinghouse/69/26/6926.pdf>

Ewell, Peter, and Marianne Boeke, *Critical Connections: Linking State's Unit Record Systems to Track Student Progress*, Indianapolis, In.: Lumina Foundation for Education, 2007.

Ewell, Peter, Paula R. Schild, and Karen Paulson, *Following the Mobile Student: Can We Develop the Capacity for a Comprehensive Database to Assess Student Progression?* Indianapolis, In.: Lumina Foundation for Education, 2003.

Greene, David, and Paula Perkins, Legal Memorandum to Pamela Burdman, Orogam Officer at the William and Flora Hewlett Foundation, Frist Amendment Project, Oakland, Calif., February 9, 2007

Hansen, Janet S., *Education Data in California: Availability and Transparency*, Palo Alto, Calif.: Stanford University: Institute for Research on Education and Practice, 2007. As of August 7, 2007:

[http://irepp.stanford.edu/documents/GDF/STUDIES/15-Hansen/15-Hansen\(3-07\).pdf](http://irepp.stanford.edu/documents/GDF/STUDIES/15-Hansen/15-Hansen(3-07).pdf)

L'Orange, Hans P., and Peter Ewell, *P–20 Data Systems: An Alignment Status Report*, January 2007. As of December 11, 2007:

http://www.dataqualitycampaign.org/files/Meetings-DQC_Quarterly_Issue_Brief_061306.pdf

Mills, Jack, *State Data Systems and Privacy Concerns: Strategies for Balancing Public Interests*, Boston, Mass.: Jobs for the Future, 2005.

National Commission on Accountability in Higher Education, *Accountability for Better Results: A National Imperative for Higher Education*. Boulder, Colo.: State Higher Education Executive Officers (SHEEO), 2005.

The Secretary of Education's Commission on the Future of Higher Education, *A Test of Leadership: Charting the Future of U.S. Higher Education*, Washington, D.C.: U.S. Department of Education, 2006.