California Postsecondary Education Commission

Implementing the Commission's Student Record System -- First Steps

June 2005
OP/05-05

This report serves as a progress report on the implementation of the Commission's longitudinal student data system. Recent data submission by the University of California and the California State University, along with the community college data already in-house, will make it possible for the Commission to conduct studies on time-to-degree, dropout patterns, transfer patterns, and concurrent enrollment.

This kind of information will enable policy makers to determine how well public colleges and universities are responding to State policy priorities and accountability goals.

Contents

The Commission advises the Governor and Legislature on higher education policy and fiscal issues. Its primary focus is to ensure that the state's educational resources are used effectively to provide Californians with postsecondary education opportunities. More information about the Commission is available at www.cpec.ca.gov.

Background

The California Postsecondary Education Commission (CPEC) has collected individual student records since the inception of its data collection program in 1976. Although the data provided to CPEC over the years has offered greater understanding of the dynamics of higher education in California, the Commission has been limited in its ability to track individual students as they progressed through the system. Recognizing the need for better data, in 1999 the Legislature passed AB 1570, a bill which directed the Commission 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.*"

The key to developing and maintaining a comprehensive database that supports the longitudinal studies called for in AB 1570 is the addition of a unique student identifier to the data records already sent to CPEC by each of the public segments. These data include an enrollment record for each student attending a public college or university and a degree or certificate record for every student who graduated. The social security number is the most common student identifier in use and will likely remain so unless, or until, the California School Information System (CSIS) identifier used by K-12 education is adopted by the higher education segments.

Three significant issues were raised during the process of enacting AB 1570: (1) the cost of developing and maintaining the system; (2) compatibility with K-12 education data (CSIS) so that information about student progress would be available throughout a student's educational experience; and (3) protecting the confidentiality of personally identifiable information about students.

Cost of a longitudinal data system

During hearings held on AB 1570, the Department of Finance and others raised concerns about the cost of implementing and maintaining a longitudinal database. Since the segments were already collecting a student identifier, usually a student's social security number, the costs to collect this information were deemed to be minimal. For its part, the Commission had also significantly reduced its data processing costs by moving all processing from the Teale Data Center to its in-house database servers. While the positions that were created as a result of the passage of AB 1570 were eliminated in the budget cutbacks of the past few years, the Commission is implementing the system within existing personnel resources.

Consequently, concern about the costs associated with building and maintaining a longitudinal student data system has not been an issue since the passage of the bill.

Compatibility with the California School Information Services (CSIS) System

The initial system envisioned use of the SSN as the common student identifier since it was already being collected by the higher education segments. The K-12 system recently developed an identifier known as the CSIS identifier. If and when the segments begin capturing the CSIS identifier and reporting it to CPEC, it will be integrated into the CPEC longitudinal database and CPEC will have the ability to follow through with the Legislature's intent to maintain a seamless database for students from kindergarten through higher education.

Confidentiality of student information

As enacted, the legislation mandated that CPEC comply with all federal Family Educational Rights and Privacy Act (FERPA) of 1974 (20 U.S.C. Sec. 1232g) regulations. The legislation itself also specified that no personally identifiable information be released. The Commission is committed to maintaining the confidentiality of individual student records and will not release any information that could be used to identify individual students.

In past discussions, the California State University (CSU) and the University of California (UC) contended that releasing data containing unique student identifiers would violate the confidentiality requirements of FERPA. Legal opinions obtained from both the State Attorney General and the U.S. Department of Education confirmed the legal basis allowing CPEC to collect data containing individual student identifiers and, in April 2005, both UC and CSU agreed to provide the data. The California Community Colleges have provided unique student data to CPEC for five years.

Protection against unauthorized access

To protect student SSNs from unauthorized access, CPEC developed a data handling procedure that makes it unnecessary for SSNs to be maintained in its data files. ***Use of this procedure means that no student social security numbers will be stored on any computer or server maintained by the Commission.***

CPEC created an algorithm that converts social security numbers to codes that have no obvious relationship to the original SSNs. The segments can use this algorithm to encrypt the SSNs before transmission to CPEC. The original SSNs cannot be recovered from these codes without knowledge of the structure of the algorithm and the parameters that are used when the algorithm is actually applied.

In order to further protect confidentiality, CPEC will recode the code from the first encryption to a second encryption. This recoding will be done using a list of unique codes, kept in a secure location in CPEC's offices. This double encoding ensures maximum protection of student privacy.

SSNs were included in the initial submission of data so that the Commission could fully test the algorithm to ensure that it worked as expected. The Commission will erase the SSNs and they will not be stored at CPEC. The segments will not send any student SSNs to CPEC after their initial submission. This procedure will also allow the Commission to perform data quality assurance as well as validating the process for future data submissions.

To further enhance security, the Commission will not connect any computer or server on which it is processing these data to its internal local area network. Any staff accessing the data will use a stand-alone workstation to process data before it is aggregated onto a database server. The workstation hard drive will be wiped clean after processing has been completed. This equipment is located in a room which is securely locked and access is recorded. The Commission also is using a logbook to maintain a "chain of custody" of the data; it will be immediately evident if anyone tampers with this logbook.

Data submission

On May 2, 2005, the University of California submitted its individual student longitudinal data to the Commission and on May 3, 2005, the California State University submitted its individual student longitudinal data. The California Community College Chancellor's Office is resubmitting individual student longitudinal data. The following section outlines the Commission's initial approach for validation and processing of the data.

SSN validation and assessment

The current submission of data by the segments with the social security number as part of the record provides a unique opportunity to validate and assess the use of this number as the foundation for a longitudinal student identifier.

Algorithm validation

In subsequent years, the segments will submit their data with the SSN converted to a random number that the Commission will not be able to trace to a student. Having the SSN this year will allow CPEC staff to test the encoding algorithm that will be turned over to the segments to use for future data submissions.

Double-encoding process validation

The Commission will also test and validate its re-encoding of student IDs to further protect student confidentiality. Using the re-encoded number ensures that the segments would not be able to identify any of the students.

Editing and feedback

Commission staff is developing the procedures to edit, validate, and provide feedback to sending segments about record counts for each data element reported to CPEC. It is important that the data are verified as being complete and accurate, and match what the segments believe they transmitted to CPEC before any of these data are used for evaluation and analysis.

Data set validity (universe vs. sample)

CPEC staff will make an assessment about how to treat these data based on the number of student records missing valid SSNs. It may be that CPEC staff will treat the resulting data set as "sample" data if

too many student identifiers are missing or invalid. It is important that any statistical biases of the data be identified when the data are used.

Rules for data handling

Inevitably, differences in the coding of various student characteristics by multiple campuses will require the development of business rules to deal with these choices. For example, a student with the same identifier may have different ethnic codes used by different campuses. All of these situations require data handling rules that will have to be published to provide transparency to the use and consistency of the data.

Collaboration

The Commission plans to establish a research advisory committee to provide advice to its staff on the various studies and analyses that will be conducted using these data.

Anticipated first projects

With the addition of a student identifier, the Commission is better positioned to study the movement of individual students into and through the public segments of higher education. Initially, the Commission anticipates it will conduct studies to better understand and report on time-to-degree, dropout and stop-out patterns, transfer patterns, and concurrent enrollment. The information reported through the longitudinal student system will provide information necessary for policy makers to determine how well public colleges and universities are responding to state policy priorities and accountability goals.