WestEd
WestEd.org

Examination of Data Usability Options for Assessing Eligibility for Higher Education in California

Anthony Fong
Vanessa Barrat
Neal Finkelstein

May 2018

Suggested citation: Fong, A., Barrat, V., & Finkelstein, N. (2018). *Examination of data usability options for assessing eligibility for higher education in California.* San Francisco, CA: WestEd.

WestEd is a nonpartisan, nonprofit research, development, and service agency that works with education and other communities throughout the United States and abroad to promote excellence, achieve equity, and improve learning for children, youth, and adults. WestEd has more than a dozen offices nationwide, from Massachusetts, Vermont, Georgia, and Washington, DC, to Arizona and California, with headquarters in San Francisco.

Contents

List of Figures

List of Tables

Acknowledgements

We would like to thank the many people and organizations that made this report possible, as their time and energy are very much appreciated. We are grateful to the Governor's Office of Planning and Research for the financial support provided to conduct the study. The California Department of Education provided the data from the California Longitudinal Pupil Achievement Data System and reviewed the report for technical accuracy. The University of California and the California State University were also generous in providing data from their admissions data files, and they both provided technical support to understand and report on key elements of the data. Finally, we are very grateful to David Silver and his team at RTI International for working with us in comparing the data from the California Department of Education with the high school transcripts collected by his team.

Executive Summary

In 2015, consistent with California Senate Bill 103, the Governor's Office of Planning and Research commissioned an analytic study to determine the number of students in California who were eligible to attend the University of California (UC) system and the California State University (CSU) system under current policies. The study, *University Eligibility Study for the Public High School Class of 2015* (Silver, Hensley, Hong, Siegel, & Bradby, 2017), required the collection of individual transcripts from high schools across the state. WestEd was also commissioned to examine data alternatives for future eligibility studies by examining the comparability of extant data being collected for other primary purposes. In particular, WestEd was commissioned to examine whether federal reporting data submitted to the California Department of Education (CDE) by Local Education Agencies (LEAs) could be adequately utilized as a surrogate for individual transcript data in future UC/CSU eligibility studies. The CDE data to be examined were obtained from the California Longitudinal Pupil Achievement Data System (CALPADS), which is a federally funded student-level longitudinal dataset designed primarily for K–12 federal reporting purposes.

This study was commissioned because of the inherent differences between CALPADS course data and the transcript data maintained by high schools and other LEAs. CALPADS data are certified by each LEA to be accurate at specific points in time necessary to meet the federal reporting deadlines, but the data are not as current or specific as data submitted in the UC/CSU enrollment applications. Moreover, CALPADS course data for the three school years of 2011–12, 2012–13, and 2013–14 are known to be incomplete. During those years, LEAs submitted course data to CALPADS only on a voluntary basis, and not all LEAs chose to do so. The CDE has estimated that the voluntary submission rate for LEAs during those three years was 62%, 93%, and 92%, respectively, compared with 98–99% for the three years since, when submission was mandatory. However, in order to consider a full, four-year cohort, it was necessary for WestEd to examine the incomplete data that districts voluntarily submitted in the three years prior to the 2014–15 school year.

This report examines data similarities and differences across four data sources: CALPADS, UC admissions files, CSU admissions files, and raw transcript data files that were provided by a sample of California high schools in the Silver et al. (2017) study. The analysis looks at the student-level records across the different sources for the four school years between the fall of 2011 and the spring of 2015 to examine the patterns of discrepant information (e.g., different A-G course names and/or grades). It also examines both the availability and the alignment across sources of credits during this period.

Availability of courses

For the population of 428,410 California public school students who graduated from grade 12 in 2015, WestEd examined over 18 million course records corresponding to courses taken from school years 2011–12 to 2014–15. WestEd examined the availability of course records to report on any year or school

characteristic that might be related to underreporting in CALPADS records. The availability of the CALPADS course records for students varied across the four years beginning with its voluntary implementation in 2011–12 until 2014–15, when submissions were mandatory for the first time. Specifically, while 62% of the 2015 graduates had taken at least one course during each of the school years of the analysis, 20% of the students had no course records in CALPADS for school year 2011–12 but had course records for all other years. This percentage is much higher than that noted for the subsequent years, which was expected because only 62% of LEAs submitted records in this first year of CALPADS course collection. However, data confirm that in subsequent years the proportion of LEAs that voluntarily submitted course information increased significantly until submission was mandatory (CDE estimates a 98–99% submission rate).

The availability of courses in CALPADS also varied with some characteristics of the school of enrollment. The availability of course-level data for all four years of the analysis was lower for students enrolled in schools with a high percentage of students eligible for free or reduced-price meals (FRPM)[1] (56% versus 62% for all students), for students enrolled in small schools[2] (36%), and for students enrolled in nontraditional schools[3] (31%). These percentages have implications for eligibility calculations of students enrolled in these types of schools.

Variation in the availability of CALPADS course records by school characteristics and lower availability for the early years of CALPADS might preclude the utilization of 2011–2015 CALPADS course records to estimate the eligibility rates of the 2015 graduates. However, this report can help to better understand the possibility of using the CALPADS course records for future UC/CSU eligibility studies based on an evaluation of the latest years of CALPADS data available at the time of writing this report.

Alignment of units passed by subject area across sources

To this goal, WestEd examined A-G course completion, by A-G category and by year, using the data that districts submitted into CALPADS. WestEd then assessed how close these values were to those reported in the UC and CSU admissions data. WestEd also estimated grades 10 and 11 grade point average (GPA) using CALPADS course records and compared it to a similar estimate computed using UC and CSU admissions course data. Finally, a comparison examined A-G course completion, by A-G category and by year for the years 2012–13 and 2013–14, between CALPADS course data and transcripts collected by RTI International for their eligibility study (Silver et al., 2017).

WestEd received course-level data for 90,533 fall 2015 applicants to the UC system from California public high schools and was able to match 98% of those applicants, or about 88,000 students, with CALPADS records. In the CALPADS–UC comparison, the rates for which there was exact alignment

[1] Defined as greater than or equal to 75% of total school enrollment.

[2] Defined as enrollment less than or equal to 300 students.

[3] Nontraditional school types included schools classified in CALPADS as Alternative Schools of Choice, Continuation High Schools, County Community Schools, District Community Day Schools, Juvenile Court Schools, Opportunity Schools, Special Education Schools, State Special Schools, and Youth Authority Facilities.

averaged approximately 75% over the different A-G subject areas and across the years. However, the alignment rates varied by year, not surprisingly showing improvement over time from 2011–12 to 2013–14, and the alignment rates varied by A-G subject areas. By 2013–14, an exact alignment of the number of units passed with a grade of C or better is reported for 76–87% of students with a subject course recorded in either data source for **history and social science** (A), **English** (B), **mathematics** (C), **laboratory science** (D), and **language other than English** (E). Alignment of units for **visual and performing arts (**F) and especially for **college preparatory elective** (G) courses was notably lower. If there was not exact alignment, then CALPADS number of units being lower than the number of units in the UC admissions dataset or data only being in the UC admissions dataset was the most common occurrence. Matching rates were lower for small schools and nontraditional school types but not for schools with a high percentage of students who are eligible for FRPM. In addition, the GPA calculations between the CALPADS dataset and the UC dataset were similar for most students. For over 95% of the students, the difference was within three tenths of one point. When calculating whether a student's GPA was at or above 3 between the two datasets, the conclusion was the same for 97% of students.

Once consolidated across the different campuses, the CSU application dataset included records for about 185,000 students. The study matched 80% of those applicants, or about 148,000 students, with CALPADS records. The general trends of alignment were similar to those observed for the UC records: the alignment rates varied by year (with improvement over time), and they varied by A-G subject areas, with **visual and performing arts** (F) and **college preparatory elective** (G) courses showing a lower rate of alignment. The low rate of alignment may be due to LEA student information systems (SISs) failing to recognize that extra A-F subject matter courses should be converted to a G course elective when the SIS data is downloaded to CALPADS. Generally, the alignment rates for the CSU records were lower by a few percentage points compared to the UC records. In addition, while matching rates were lower for small schools, they were comparable to the statewide estimate for nontraditional school types and schools with a high percentage of students eligible for FRPM. For close to 90% of the students the GPA calculations were within three tenths of one point between the two sources, and, when estimating whether a GPA was at or above 3, the conclusion was the same for 93% of students.

The alignment of the CALPADS records for 2012–13 and 2013–14 was generally higher with the school transcripts collected for the RTI International study than was observed with records from the UC and CSU admissions datasets. However, the alignment showed the same trends of lower units in CALPADS and high non-alignment for the **college preparatory elective** (G) courses.

Specific challenges to using CALPADS administrative records for eligibility studies

Based on the analysis, WestEd found several specific challenges related to the 2011–2015 CALPADS course record data that should be investigated before those records can be used for an eligibility estimate:

- *Allocation of A-G courses to the different A-G categories*. About 5% of the CALPADS courses marked as A-G courses are missing an A-G category and cannot be used to

evaluate eligibility criteria without further analysis of the course label itself, and this issue is likely related to the undercounting of CALPADS units reported. Furthermore, the categorization of courses into the electives category was problematic and led to a very low matching rate for that category. A review of the categorization of courses into A-G categories in CALPADS so that it matches, for each year and school, the categorization used by the UC/CSU system would solve that challenge.

- *Terms and marking periods conversion*. The combinations of marking periods and terms were particularly complex and often did not add up to a clear description of the course length of instruction. Implementing a series of checks to verify the integrity of the combinations of terms and marking periods submitted by the schools would allow a better estimate of the number of units passed each year.
- *Validation rules*. Application of the validation rules requires looking beyond the A-G classifications at the specific course codes and labels. While CALPADS might be used in the future as an alternative for collecting school records, transforming the different marking period systems and applying the set of validation rules requires deep knowledge and information about the specific courses.

Introduction

In 2015, consistent with California Senate Bill 103, the Governor's Office of Planning and Research commissioned an analytic study to determine the number of students in California who were eligible to attend the University of California (UC) system and the California State University (CSU) system under current policies. The study, *University Eligibility Study for the Public High School Class of 2015* (Silver, Hensley, Hong, Siegel, & Bradby, 2017), required the collection of individual student transcripts from high schools across the state.

Because the Governor's Office understood the significant primary data collection requirements at the time the Silver et al. (2017) report was commissioned, a question was posed by the staff about possible alternative methodologies to collecting the necessary data. The 2017 eligibility report, as have similar analyses in the past, required a comprehensive review of student transcript files that were gathered from a sample of California high schools according to a carefully determined sampling framework. The data collection took time and resources and required a comprehensive review of course patterns and sequences to align with precise eligibility requirements.

One alternative path to completing the eligibility analysis in the future might be to examine other data sources that track similar course patterns and course grade information. To that end, WestEd was commissioned to examine the comparability of extant data being collected for other primary purposes and to examine whether the data could be adequately utilized as a surrogate for individual transcript data in future UC/CSU eligibility studies.

This report examines data similarities and differences across four data sources: the California Longitudinal Pupil Achievement Data System (CALPADS), UC admissions files, CSU admissions files, and raw transcript data files that were provided by a sample of California high schools in the Silver et al. (2017) study. The analysis looks at the student-level records across the different sources for the four school years between the fall of 2011 and the spring of 2015 to examine the patterns of discrepant information (e.g., different A-G course names and/or grades). WestEd looked for student records, or partial student records, that appear in one of the data sets but not the other. WestEd analyzed the mapping between CALPADS and the UC/CSU data files; and additional comparisons were coordinated between WestEd and RTI International to respect each firm's applicable data-sharing agreements with their data providers.

This study was commissioned because of the inherent differences between CALPADS course data and high school transcript data. CALPADS was funded by the federal government to allow the CDE to collect and report required Local Education Agency (LEA) data to the federal government. Accordingly, CALPADS data are certified by each LEA to be accurate at those specific points in time necessary to meet federal reporting deadlines, but the data are not as current or specific as the data submitted for UC/CSU enrollment applications. Moreover, the CALPADS course data for the three school years of 2011–12, 2012–13, and 2013–14 are known to be incomplete. During those years, LEAs submitted course data to CALPADS only on a voluntary basis, and not all LEAs chose to do so. The CDE estimated that the voluntary submission rate for LEAs was 62%, 93%, and 92%, respectively, for those years, compared with 98–99% for the three years since (as reported by CDE), in which submission was mandatory. However, in order to consider a full, four-year cohort, WestEd examined the incomplete data that districts voluntarily submitted in years prior to the 2014–15 school year.

CALPADS Data Availability

For the population of 428,410 California public school students who graduated from grade 12 in 2015, WestEd received over 18 million course records corresponding to courses taken from school years 2011–12 to 2014–15. CALPADS course records included a unique student identifier, the academic school year, school code, school name, local course code, local course description, state course code, state course description, A-G indicator, A-G admissions requirement code (A-G category), instructional-level code (documenting UC-certified Honors and college credit courses), academic term, marking period, final grade, credits attempted, and credits earned.

As noted previously, the CDE has estimated that the voluntary submission rate for LEAs was 62%, 93%, and 92%, respectively, for 2011–12, 2012–13, and 2013–14, increasing thereafter to 98–99% in subsequent years. Accordingly, the availability of the CALPADS course records for students varied across years, as illustrated in Table 1.

Table 1: Course record availability in CALPADS

	Number of students	Percent of students
Courses in all four school years 2011–12 through 2014–15	261,354	62%
Three years of course data, with school year 2011–12 missing	83,410	20%
Three years of course data, with either school year 2012–13, 2013–14, or 2014–15 missing	30,751	7%
Two years of course data	27,572	7%
One year of course data	15,517	4%
Total	**418,604**	**100%**

Note: Of the 428,410 graduates, 3,892 students had no course information in CALPADS, and 5,914 students had course information but no registered A-G courses.

Specifically, while 62% of the 2015 graduates had records of at least one course for each of the school years of the analysis, 20% of the students had no course records in CALPADS for school year 2011–12 but had courses for all other years. Seven percent of the students were missing exactly one other year of data, while 11% had only one or two years of course records for the period under analysis. Given that LEAs were not required to submit course data into CALPADS until the 2014–15 school year (the final year of data collected for the current study), it is encouraging to note that 89% of students had three or more

years of course data available directly from CALPADS and 96% of students had two or more years of data available.

The percentage of students with missing data for school year 2011–12 (20%) is much larger than for the subsequent years and validates the fact that this early year of CALPADS course collection was not yet complete. Course records for 2011–12 were incomplete and not certified and would therefore be problematic to use for eligibility studies. However, even though data for the two subsequent years are also incomplete, the submission rate (per CDE) significantly increased from 62% to over 92%. Because submission rates are now over 98%, the missing data issue may be nearly resolved.

The availability of course records in CALPADS also depended on some characteristics of the school of enrollment. Table 2 presents the course record availability in CALPADS for specific subsets of students based on the characteristics of their school of enrollment as of graduation. Table 2 reports availability of courses for schools with a high percentage of students eligible for FRPM (at least 75%), small schools (less than or equal to 300 students enrolled), and students enrolled in nontraditional school types. Nontraditional school types included schools classified in CALPADS as Alternative Schools of Choice, Continuation High Schools, County Community Schools, District Community Day Schools, Juvenile Court Schools, Opportunity Schools, Special Education Schools, State Special Schools, and Youth Authority Facilities.

Table 2: Course record availability in CALPADS for students enrolled in schools with a high percentage of students qualifying for FRPM, small schools, or nontraditional school types

Course data availability	All schools	High percentage of FRPM-eligible students (greater than or equal to 75%)	Small schools (enrollment less than or equal to 300 students)	Nontraditional school types*
	(Total students = 418,604)	(Total students = 115,169)	(Total students = 29,270)	(Total students = 36,435)
Courses in all four school years 2011–12 through 2014–15	62%	56%	36%	31%
Three years of course data, with school year 2011–12 missing	20%	24%	12%	18%
Three years of course data, with either school year 2012–13, 2013–14, or 2014–15 missing	7%	8%	18%	18%
Two years of course data	7%	8%	21%	20%
One year of course data	4%	5%	12%	12%

Note: Column percentages may not add up to 100% due to rounding.

***Nontraditional school types included schools classified in CALPADS as Alternative Schools of Choice, Continuation High Schools, County Community Schools, District Community Day Schools, Juvenile Court Schools, Opportunity Schools, Special Education Schools, State Special Schools, and Youth Authority Facilities.**

The rate of availability of course-level data for all four years of the analysis was lower for students enrolled in schools with a high percentage of students eligible for FRPM (56% versus 62% for all students). This result corresponded to a higher percentage of students missing course data for 2011–12. However, general availability of course data was similar to course availability of the whole population for the subsequent years (e.g., similar percentages with only one or two years of data). In contrast, for students enrolled in small schools or nontraditional schools, about 20% had no course data available for two years of the period under analysis, about three times the rate for the whole population, and 12% had only one year of course data available from 2011–12 to 2014–15 (this is also three times the rate for the whole population). These results suggest that the lower availability of course data may have persisted beyond the low reporting of school year 2011–12.

Variation in the availability of course records and the lack of record certification during the early years of CALPADS might preclude the utilization of CALPADS course records to estimate the eligibility rates of the 2015 graduates. However, an important goal of this report is to estimate the potential of using the CALPADS course records based on an evaluation of the latest years of CALPADS data available.

To determine if federal reporting data submitted to the CDE by LEAs could be adequately utilized as a surrogate for transcript data in future UC/CSU eligibility studies, it would be optimal to be able to compare the eligibility rate using different sources of data, including the UC and CSU admissions data as well as transcript data collected by RTI International for its recent eligibility report (Silver et al., 2017). However, UC and CSU admissions data do not include a full set of grade 12 course records because students apply to the UC and CSU in the fall of their senior year. In addition, RTI International course-level data collected for each student are currently protected from being accessed across agencies.

To estimate if the CALPADS course data could be adequately utilized as a surrogate for transcript data in the future, WestEd examined A-G course completion, by A-G category and by year, using the data that districts submitted into CALPADS; then WestEd assessed how close these values are to the UC and CSU admissions data. WestEd also estimated grades 10 and 11 GPA using CALPADS course records and compared it to a similar estimate computed using UC and CSU admissions course data. Finally, a third analysis compared A-G course completion, by A-G category and by year for the years 2012–13 and 2013–14, between CALPADS course data and transcripts collected by RTI International for its eligibility study.

Alignment between CALPADS data and UC data

WestEd received course-level data from the UC for 90,533 fall 2015 applicants to the UC from California public high schools. Over 84% of the students who applied to the UC submitted a self-reported Statewide Student Identifier (SSID). Using that unique identifier, WestEd matched directly about 80% of the UC applicants with the CALPADS records. Then, using a fuzzy matching process, WestEd was able to increase the matching rate to 98% of the UC applicants, or about 88,000 applicants. Additional details about the matching process are included in the appendix.

Focusing on the UC applicants for whom WestEd could find course information in CALPADS for all years of the study (60,849 students), WestEd examined the alignment of the number of A-G courses with a

grade of C or better, by A-G subject area and by school year, between the CALPADS and the UC course records.

Tables 3 through 5 summarize the results by school year and show variation in alignment by year and A-G category.

Table 3: Comparison between the CALPADS and UC datasets of the number of units (in years) with a grade of C or better by A-G subject area, 2011–12

		In both CALPADS data and UC data			In UC data only (percent)	In CALPADS data only (percent)
Subject area	Number of students*	CALPADS number of units = UC number of units (percent)	CALPADS number of units > UC number of units (percent)	CALPADS number of units < UC number of units (percent)		
A	24,617	60.8	0.2	12.5	22.7	3.8
B	60,724	80.9	0.8	13.9	4.3	0.1
C	60,775	82.1	1.9	11.8	4.0	0.3
D	54,207	75.7	0.2	14.3	6.2	3.6
E	50,553	85.4	0.6	9.2	4.5	0.4
F	30,114	71.1	2.0	14.1	10.9	2.0
G	20,774	22.6	4.3	8.3	40.9	23.9

Note: Row percentages may not add up to 100% due to rounding.

***Number of students who passed at least a course with a grade of C or better according to at least one of the data sources in the corresponding category. Students for whom the data sources align on reporting that they did not take a course in a given category in 2011–12 (60,849 minus the population reported in the table for each category) are not included in the table.**

Table 3 provides data on the percentage of students in the sample, for each A-G subject area, for each of the five scenarios: 1) exact alignment of the number of units passed with a grade of C or better between the CALPADS and UC datasets, 2) higher number of units in the CALPADS dataset than in the UC dataset, 3) lower number of units in the CALPADS dataset than in the UC dataset, 4) units only recorded in the UC dataset, and 5) units only recorded in the CALPADS dataset.

Table 3 shows that about 40% of UC applicants in the sample (24,617 out of 60,849 students) had passed **history and social science** (A) courses with a grade of C or better according to the CALPADS dataset or the UC admissions dataset in 2011–12. The number of units passed with a grade of C or better in the two data sources was the same for about 61% of these students, higher in CALPADS for less than 1% of the students, and lower in CALPADS for 13% of the students. For 23% of the students, no records of **history and social science** (A) courses could be found in CALPADS, but there were records of these courses in the UC admissions system. This result suggests that a certain number of **history and**

social science (A) courses may not have been identified as A-G courses in CALPADS. In comparison, 4% of the students had no records of **history and social science** (A) courses in the UC dataset but did have units recorded as such in the CALPADS dataset.

For **English** (B), **mathematics** (C), and **language other than English** (E), at least 80% of the students had the same number of units in both the CALPADS dataset and the UC dataset. For **laboratory science** (D) and **visual and performing arts** (F) courses, the exact alignment rate was between 71% and 76%.

Finally, **college preparatory elective** (G) courses had a much lower degree of alignment between the two datasets. Only 23% of the students had the same number of annualized units in CALPADS and the UC dataset. A high percentage of records corresponding to **college preparatory elective** (G) courses in the UC admissions dataset were not found in CALPADS (41%) for the same students; likewise, a high percentage of these records in CALPADS were not found in the UC dataset (24%). The fact that any course in the A-F categories could also be used to fulfill the elective requirement at the time of admission could explain the high rate of non-alignment of the G courses. In addition, the low rate of alignment may be due to an LEA's student information system (SIS) failing to recognize that extra A-F subject-matter courses should be converted to a G course elective when the SIS downloads data to CALPADS. For example, if a student takes an extra **mathematics** (C) course as an elective, CALPADS allows the LEA to identify it as meeting both C and G requirements by designating it as GC. However, the SIS may not have that capability or may report it as an additional C course rather than as a G course.

Overall, if annualized units were observed in both CALPADS and the UC dataset and if the units did not align across the two datasets, then the count in the CALPADS dataset was likely lower than that of the UC dataset. It is also noted that among the subject courses with larger amounts of enrolled students (particularly B through F), there are higher rates of exact alignment.

Finally, it is worth noting again that 2011–12 was the first year for the voluntary submission of course records in CALPADS, and that the quality of A-G course records in CALPADS has improved since the 2011–12 school year, as submission has moved from voluntary to mandatory.

Table 4: Comparison between the CALPADS and UC datasets of the number of units (in years) with a grade of C or better by A-G subject area, 2012–13

		In both CALPADS data and UC data			In UC data only (percent)	In CALPADS data only (percent)
Subject area	Number of students*	CALPADS number of units = UC number of units (percent)	CALPADS number of units > UC number of units (percent)	CALPADS number of units < UC number of units (percent)		
A	57,825	74.5	0.5	18.1	6.6	0.3
B	60,808	77.9	0.7	16.7	4.5	0.1
C	60,666	81.6	1.9	12.3	4.0	0.3
D	59,301	76.4	0.6	18.1	4.5	0.4
E	55,857	83.4	0.4	11.4	4.4	0.4
F	23,389	64.0	2.2	19.0	12.5	2.4
G	14,325	20.9	3.1	6.1	39.3	30.6

Note: Row percentages may not add up to 100% due to rounding.

***Number of students who passed at least a course with a grade of C or better according to at least one of the data sources in the corresponding category. Students for whom the data sources align on reporting that they did not take a course in a given category in 2012–13 (60,849 minus the population reported in the table for each category) are not included in the table.**

With respect to the 2012–13 school year, the general overall patterns remain from the 2011–12 school year. Specifically, with the exception of the **college preparatory elective** (G) courses, the rates of exact alignment between the CALPADS and UC dataset are at least 64%. The rate of exact alignment for the G subject area is the lowest among the different subject areas at 21% in 2012–13. Also similar to the 2011–12 school year, if records were available in both the CALPADS and UC dataset and if the two datasets differed, usually the UC number of units was larger than the CALPADS number of units. Finally, **language other than English** (E) courses had the highest rate of exact alignment at 83%, followed by **mathematics** (C) at 82%.

In 2013–14, all the measures of alignment improved, perhaps not surprisingly, as shown in Table 5.

Table 5: Comparison between the CALPADS and UC datasets of the number of units (in years) with a grade of C or better by A-G subject area, 2013–14

		In both CALPADS data and UC data			In UC data only (percent)	In CALPADS data only (percent)
Subject area	Number of students*	CALPADS number of units = UC number of units (percent)	CALPADS number of units > UC number of units (percent)	CALPADS number of units < UC number of units (percent)		
A	59,327	76.2	4.0	14.9	4.2	0.7
B	60,692	81.1	1.2	13.9	3.6	0.2
C	59,932	82.9	2.4	10.1	4.0	0.6
D	57,593	79.4	1.6	13.7	4.3	1.0
E	45,411	87.3	0.5	6.5	5.1	0.5
F	30,554	70.6	1.9	14.7	10.9	1.9
G	24,850	29.9	3.1	12.4	41.5	13.2

Note: Row percentages may not add up to 100% due to rounding.

***Number of students who passed at least a course with a grade of C or better according to at least one of the data sources in the corresponding category. Students for whom the data sources align on reporting that they did not take a course in a given category in 2013–14 (60,849 minus the population reported in the table for each category) are not included in the table.**

For the 2013–14 school year, with the exception of the **college preparatory elective** (G), all of the subject areas have exact alignment rates above 70%. More specifically, **history and social science** (A), **English** (B), **mathematics** (C), **laboratory science** (D), **language other than English** (E), and **visual and performing arts** (F) have exact alignment rates of 76%, 81%, 83%, 79%, 87%, and 71%, respectively. However, the exact alignment rate for **college preparatory elective** (G) courses remains low at 30% (although it did increase from 21% in 2012–13).

Overall, about 300,000 A-G courses were compared for each school year of the analysis. Table 6 shows that across all categories, about 75% of the courses taken had the same annualized number of units between the UC admissions dataset and the CALPADS record, with a small improvement across years.

Table 6: Alignment rate of the number of A-G units (with a grade of C or better) by year between CALPADS data and UC data

	Total number of A-G courses compared	CALPADS number of units = UC number of units	CALPADS number of units > UC number of units	CALPADS number of units < UC number of units	In UC data only	In CALPADS data only
2011–2012	301,764	74.3	1.2	12.3	9.3	3.0
2012–2013	332,171	75.2	1.0	15.2	6.8	1.8
2013–2014	338,359	76.4	2.1	12.3	7.6	1.6

Note: Row percentages may not add up to 100% due to rounding.

Focusing on the last year of data, WestEd examined the variation in matching rates by some characteristics of the school of graduation (Table 7). Few students coming from small schools (defined as enrolling 300 students or less) applied to UC; WestEd compared about 3,000 records from 2013–14 for these students. The overall percentage of courses with the same number of annualized units obtained with a grade of C or better was lower for students applying from small schools than the overall rate. This result corresponded to a higher percentage of courses with a lower number of units in CALPADS than in the UC dataset as well as a higher percentage of courses that could not be found in CALPADS. The number of records available for comparison for students applying from nontraditional schools (including schools classified in CALPADS as Alternative Schools of Choice, Continuation High Schools, County Community Schools, District Community Day Schools, Juvenile Court Schools, Opportunity Schools, Special Education Schools, State Special Schools, and Youth Authority Facilities) was even lower, about 900 records, and showed a slightly lower percentage of exact match as well as a higher percentage of courses with a lower number of units in CALPADS. In contrast, for students applying from schools with FRPM eligibility making up 75% or more of total enrollment, the percentage of courses that matched exactly was actually higher than the population percentage (79% versus 76%).[4]

[4] With respect to the percentage of graduates who applied to the UC from the three categories in Table 7, among those graduates with course information in CALPADS for all years of the study, the percentages are as follows: 1% of the graduates graduated from small schools, 19% graduated from schools with a high percentage of FRPM-eligible students, and less than 1% graduated from nontraditional school types.

Table 7: Comparison of the number of passed A-G units for students enrolled in small schools, schools with a high percentage of FRPM-eligible students, or a nontraditional school type between CALPADS data and UC data, 2013–14

	Total number of A-G courses compared	CALPADS number of units = UC number of units	CALPADS number of units > UC number of units	CALPADS number of units < UC number of units	In UC data only	In CALPADS data only
All schools	**338,359**	**76.4**	**2.1**	**12.3**	**7.6**	**1.6**
Small schools (enrollment less than or equal to 300 students)	3,007	63.4	3.2	17.7	12.6	3.1
Schools with a high percentage of FRPM-eligible students (at least 75%)	64,153	78.6	2.2	9.4	7.4	2.5
Nontraditional school types*	924	71.8	2.4	15.6	8.0	2.3

Note: Row percentages may not add up to 100% due to rounding.

***Nontraditional school types included schools classified in CALPADS as Alternative Schools of Choice, Continuation High Schools, County Community Schools, District Community Day Schools, Juvenile Court Schools, Opportunity Schools, Special Education Schools, State Special Schools, and Youth Authority Facilities.**

A GPA estimate was computed based on all courses taken in grades 10 and 11 by assigning the following values to each course: A=4 points, B=3 points, C=2 points, and D= 1 point. Extra points were allocated for up to eight semesters of approved Honors, International Baccalaureate (IB), and Advanced Placement (AP) courses with a grade of C or better. A GPA was calculated based on CALPADS course records and UC course records separately, and the difference between the two GPAs were computed for each student. The distribution of that difference is presented in Figure 1.

Figure 1: Alignment of grades 10 and 11 GPA between the CALPADS data and UC data

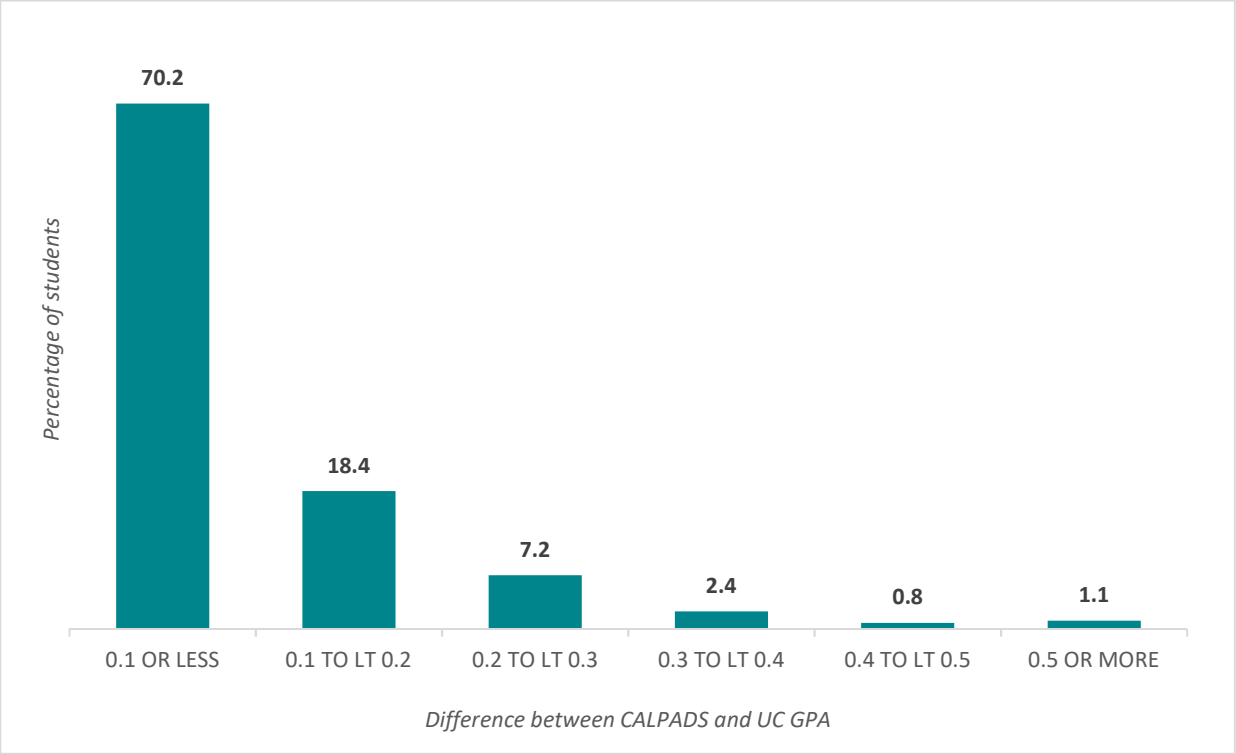


Figure 1 provides the results from an estimate of GPA differences between CALPADS and UC data computed for all students with courses in grades 10 and 11 (school years 2012–13 and 2013–14). The GPA calculations between the CALPADS dataset and the UC dataset were similar for most students. For 70% of the students the difference in GPA between the two datasets was within one tenth of a point. Moreover, for over 95% of the students the difference was within three tenths of a point.

In addition, when calculating whether a student's GPA is at or above 3 between the two datasets, the conclusion was the same for 97% of students, with 90% of GPAs estimated to be at or above 3 and 7% estimated below 3. The conclusion was different for 3% of the students.

Alignment between CALPADS data and CSU data

Once consolidated across the different campuses, the CSU application dataset included records for about 185,000 students. About 36% of the students who applied to the CSU submitted an SSID that matched the CDE records. In addition, a fuzzy matching process was used to identify CALPADS high school records for 80% of the CSU applicants, or about 148,000 applicants. Additional details about the matching process are included in the appendix.

Focusing on the CSU applicants for whom WestEd could find course information in CALPADS for all years of the study (101,343 students), WestEd examined the alignment of the number of A-G courses with a

grade of C or better, by A-G subject area and by school year, between the CALPADS and the CSU course records.

Tables 8 through 10 summarize the results by school year and show variation in alignment by year and A-G category.

Table 8 shows that the alignment of A-G courses between CALPADS and CSU in 2011–12, or grade 9 for the cohort of analysis, varied by A-G subject area. 2011–12 is a relatively early year for course records in CALPADS, and the course allocation to A-G courses may have changed since that year.

Table 8: Comparison between the CALPADS and CSU datasets of the number of units (in years) with a grade of C or better by A-G subject area, 2011–12

		In both CALPADS data and CSU data			In CSU data only (percent)	In CALPADS data only (percent)
Subject area	Number of students*	CALPADS number of units = CSU number of units (percent)	CALPADS number of units > CSU number of units (percent)	CALPADS number of units < CSU number of units (percent)		
A	34,274	55.7	0.3	13.7	23.7	6.7
B	101,017	77.3	1.0	17.2	4.2	0.4
C	101,041	77.8	2.5	13.9	4.3	1.6
D	90,421	66.6	0.3	14.7	15.4	3.0
E	77,279	82.7	1.1	10.4	4.2	1.6
F	46,488	67.7	3.2	14.3	11.0	3.9
G	33,187	19.5	4.8	6.1	29.6	40.0

Note: Row percentages may not add up to 100% due to rounding.

***Number of students who passed at least a course with a grade of C or better according to at least one of the data sources in the corresponding category. Students for whom the data sources align on reporting that they did not take a course in a given category in 2011–12 (101,343 minus the population reported in the table for each category) are not included in the table.**

In 2011–12, about a third of CSU applicants in the sample (34,274 out of 101,343 students) had records of passing **history and social science** (A) courses in CALPADS or the CSU dataset with a grade of C or better. The number of units obtained in the two sources was the same for slightly over half of the students (56%), higher in CALPADS for less than 1% of the students, and lower in CALPADS for 14%. For almost a quarter of the students, no records of passing **history and social science** (A) courses could be found in CALPADS for 2011–12, but such records had been submitted in the CSU admissions system. These alignment measures suggest that a certain number of **history and social science** (A) courses may not have been identified as A-G courses in CALPADS and that those courses that were identified as such

were reported with a combination of terms and marking periods that led to an undercounting of the annualized units passed for 14% of the students.

In contrast, for **English** (B) and **mathematics** (C), most students had a record of passing such course in both data systems, and the number of units obtained with a grade of C or better matched between the two data sources for approximately 77% of the students. A lower number of units was reported in CALPADS compared to CSU for 17% of students in **English** (B) and 14% in **mathematics** (C). **Language other than English** (E) courses followed about the same alignment pattern with a higher rate of matching (83%) and a lower rate of undercounting of the units in CALPADS (10%). For **laboratory science** (D) and **visual and performing arts** (F) courses, the alignment was lower at approximately 66%, with a higher rate of courses not found in CALPADS at 15% for **laboratory science** (D) and 11% for **visual and performing arts** (F).

Finally, **college preparatory elective** (G) courses had a much lower rate of alignment. Only 20% of the students had the same number of passed annualized units in CALPADS and CSU datasets. A high percentage of passed elective courses in the CSU dataset were not found in CALPADS (30%), and a high percentage of passed elective (G) courses in CALPADS were not used to fill the elective requirement in the CSU dataset (40%). The difficulty in classifying the G courses in CALPADS and the fact that any course in the A-F categories could also be used to fulfill the elective requirement at the time of admission could explain the high rate of non-alignment of the G courses.

Table 9 shows that, while the alignment of A-G courses between CALPADS and CSU datasets in 2012–13 still varied by A-G subject area, the measures of alignments in **history and social science** (A) and **laboratory science** (D) increased to be more comparable with the other A-E courses. Exact alignment for **visual and performing arts** (F) courses and especially **college preparatory elective** (G) courses remained low.

Table 9: Comparison between the CALPADS and CSU datasets of the number of units (in years) with a grade of C or better by A-G subject area, 2012–13

		In both CALPADS data and CSU data			In CSU data only (percent)	In CALPADS data only (percent)
Subject area	Number of students*	CALPADS number of units = CSU number of units (percent)	CALPADS number of units > CSU number of units (percent)	CALPADS number of units < CSU number of units (percent)		
A	97,725	71.6	0.5	19.8	7.3	0.9
B	101,272	74.3	0.7	20.3	4.4	0.3
C	100,987	76.5	2.4	15.4	4.0	1.8
D	98,379	71.3	0.5	20.5	5.6	2.0
E	89,987	80.7	0.8	12.8	4.3	1.4
F	34,874	59.3	3.8	19.0	13.7	4.1
G	20,595	16.6	3.0	5.3	34.3	40.8

Note: Row percentages may not add up to 100% due to rounding.

***Number of students who passed at least a course with a grade of C or better according to at least one of the data sources in the corresponding category. Students for whom the data sources align on reporting that they did not take a course in a given category in 2012–13 (101,343 minus the population reported in the table for each category) are not included in the table.**

In 2012–13, the number of passed units in A-E courses exactly aligned between the two data sources for 71% to 81% of the students who took those courses. For A-E courses it was higher in CALPADS for 2% or less of the students and lower in CALPADS for 13% to 20%, depending on the category.

While in 2011–12 **history and social science** (A) and **laboratory science** (D) courses showed a high percentage of students with elective courses in the CSU dataset that were not found in CALPADS, those rates are now at 7% or below for all A-E courses, suggesting a better identification of those courses in CALPADS. Alignment stayed markedly lower for **visual and performing arts** (F) courses and especially **college preparatory elective** (G) courses.

In 2013–14, all the measures of alignment improved as shown in Table 10.

Table 10: Comparison between the CALPADS and CSU datasets of the number of units (in years) with a grade of C or better by A-G subject area, 2013–14

		In both CALPADS data and CSU data			In CSU data only (percent)	In CALPADS data only (percent)
Subject area	Number of students*	CALPADS number of units = CSU number of units (percent)	CALPADS number of units > CSU number of units (percent)	CALPADS number of units < CSU number of units (percent)		
A	100,379	74.4	3.5	16.4	4.9	0.9
B	101,161	77.8	1.2	16.9	3.7	0.4
C	99,755	78.8	3.0	12.6	3.9	1.8
D	94,406	73.8	1.3	16.5	5.4	3.0
E	72,596	83.6	0.9	8.4	5.5	1.7
F	51,666	67.4	3.0	15.2	11.2	3.2
G	36,755	27.0	3.5	10.7	38.3	20.5

Note: Row percentages may not add up to 100% due to rounding.

***Number of students who passed at least a course with a grade of C or better according to at least one of the data sources in the corresponding category. Students for whom the data sources align on reporting that they did not take a course in a given category in 2013–14 (101,343 minus the population reported in the table for each category) are not included in the table.**

For **history and social science** (A), **English** (B), **mathematics** (C), **laboratory science** (D), and **language other than English** (E), 74% or more of the students showed the same number of annualized passed units in the CALPADS and CSU datasets for school year 2013–14. The number of units was lower in CALPADS for 8% to 17% of the students, depending on the category.

While improved compared to previous years, exact alignment stayed lower for **visual and performing arts** (F) courses and especially for **college preparatory elective** (G) courses.

Overall, about a half million A-G courses taken by applicants to the CSU system were compared for each school year of the analysis. Table 11 shows that across all categories about 70% of the courses had the same annualized passed number of units between the CSU admissions dataset and CALPADS, with improvement across the years.

Table 11: Alignment rate of the number of passed A-G units (with a grade of C or better) by year between CALPADS data and CSU data

	Total number of A-G courses compared	CALPADS number of units = CSU number of units	CALPADS number of units > CSU number of units	CALPADS number of units < CSU number of units	In CSU data only	In CALPADS data only
2011–2012	483,707	69.8	1.6	13.7	10.1	4.8
2012–2013	543,819	71.6	1.2	17.5	6.8	3.0
2013–2014	556,718	73.1	2.2	14.3	7.4	2.9

Note: Row percentages may not add up to 100% due to rounding.

Focusing on the last year of data, the variation in matching rates by some characteristics of the school of graduation was examined. About 6,500 records reported in 2013–14 from students applying from small schools were compared. As noted for the CALPADS/UC comparison on a smaller number of courses, the overall percentage of courses with the same number of annualized units obtained with a grade of C or better was lower for students applying from small schools than the overall rate. It again corresponded to a higher percentage of courses with a lower number of units in CALPADS as well as a higher percentage of courses that could not be found in CALPADS. However, for students applying from nontraditional schools and students applying from schools with FRPM eligibility making up 75% or more of total enrollment, no difference in rates of matching were observed between those applicants and the rates for the total population of CSU applicants.[5]

[5] With respect to the percentage of graduates who applied to the CSU from the three categories in Table 12, among those graduates with course information in CALPADS for all years of the study, the percentages are as follows: 1% of the graduates graduated from small schools, 25% graduated from schools with a high percentage of FRPM-eligible students, and less than 1% graduated from nontraditional school types.

Table 12: Comparison of the number of passed A-G units for students enrolled in small schools, schools with a high percentage of FRPM-eligible students, or a nontraditional school type between CALPADS data and CSU data, 2013–14

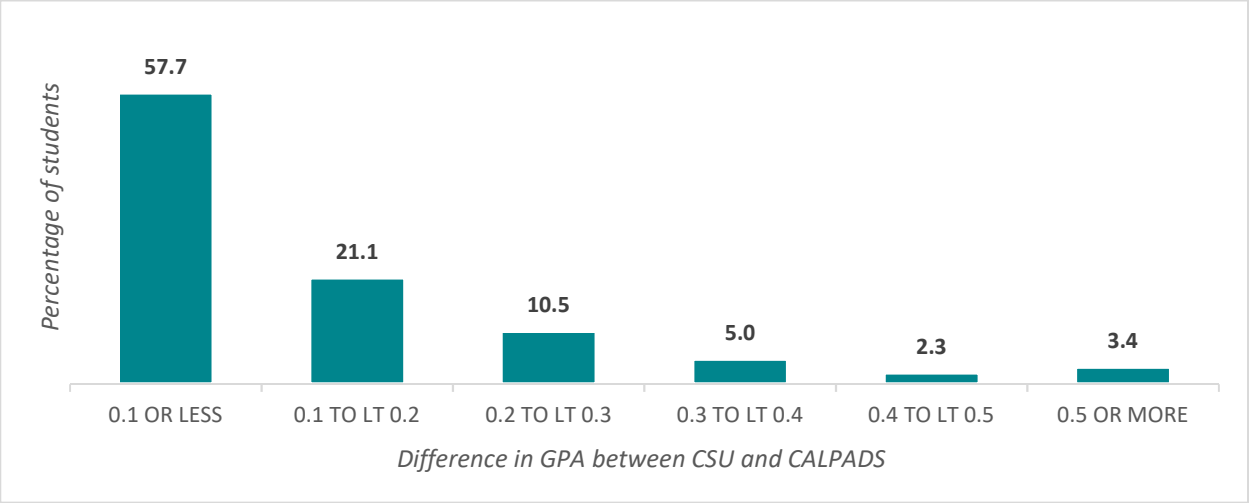
	Total number of A-G courses compared	CALPADS number of units = CSU number of units	CALPADS number of units > CSU number of units	CALPADS number of units < CSU number of units	In CSU data only	In CALPADS data only
All schools	**556,718**	**73.1**	**2.2**	**14.3**	**7.4**	**2.9**
Small schools (enrollment less than or equal to 300 students)	6,416	63.4	3.2	20.1	9.5	3.7
Schools with a high percentage of FRPM-eligible students (at least 75%)	138,417	74.0	2.4	13.1	6.8	3.7
Nontraditional school types*	2,049	73.2	2.2	14.3	7.4	2.9

Note: Row percentages may not add up to 100% due to rounding.

***Nontraditional school types included schools classified in CALPADS as Alternative Schools of Choice, Continuation High Schools, County Community Schools, District Community Day Schools, Juvenile Court Schools, Opportunity Schools, Special Education Schools, State Special Schools, and Youth Authority Facilities.**

An estimate of GPA was computed for all students with courses in grades 10 and 11 (school years 2012–13 and 2013–14). The GPA estimate was computed using the same method as described for the UC data. A GPA was defined based on CALPADS course records and CSU course records separately, and the difference between the two GPAs was computed for each student. The distribution of that difference is presented in Figure 2.

Figure 2: Alignment of grades 10 and 11 GPA between the CALPADS data and CSU data



The GPAs computed using the two data sources were close, and for the majority of students (58%) the difference between the two GPAs was within a tenth of a point. For close to 90% of students the two GPAs were within three tenths of a point.

As a result, when estimating whether a GPA is at or above 3 between the two sources, the conclusion was the same for 93% of students, with 67% of GPAs estimated at or above 3 and 26% lower than 3. The conclusion conflicted for 7% of the students.

Alignment between CALPADS data and high school transcript data collected by RTI International

WestEd worked with RTI International to compare records between the CALPADS dataset that WestEd possessed and the transcripts that RTI International had collected from a sample of high schools in California. The population of students that were compared were 1) students who attended any of the approximately 160 high schools from which RTI International collected high school transcripts[6] and 2) students for whom there was course data (at least one course) in both the 2012–13 and 2013–14 school years. To compare the records between the CALPADS dataset and the transcript files, WestEd sent to RTI International a list of the SSIDs from CALPADS that met the two previous criteria and, for each SSID, the number of units passed with at least a grade of C for each of the A-G subjects. RTI International then matched the students from WestEd's CALPADS dataset to the transcript data that it had in its possession. Next, RTI International calculated the proportion of students (for the 2012–13 and 2013–14 school years each) for which 1) CALPADS and the transcript data showed the same number of courses passed, 2) there was a higher number of courses passed in the CALPADS dataset compared to the

[6] Data for 22 schools or about 9,000 students are not included in the analysis. Those data were received and stored separately and were not available from RTI International for this analysis.

transcript file, 3) there was a higher number of courses passed in the transcript file compared to the CALPADS dataset, 4) there were no courses passed according to the CALPADS dataset, and 5) there were no courses passed according to the transcript file. The results of this analysis are presented in Tables 13 and 14.

Table 13: Comparison between CALPADS data and transcript files of the number of units (in years) passed by A-G subject area, 2012–13

Subject area	Number of students*	CALPADS number of units = transcript file number of units (percent)	CALPADS number of units > transcript file number of units (percent)	CALPADS number of units < transcript file number of units (percent)	In transcript file only (percent)	In CALPADS data only (percent)
A	44,247	85.9	2.7	8.9	1.9	0.7
B	45,826	82.8	3.0	12.0	1.6	0.6
C	45,438	79.9	5.4	11.2	1.7	1.8
D	46,937	75.3	2.3	8.4	1.3	12.6
E	32,458	85.6	3.2	8.4	2.3	0.5
F	14,277	71.7	3.4	13.1	5.4	6.4
G	5,014	29.9	2.7	1.4	19.0	47.0

Note: Row percentages may not add up to 100% due to rounding.

***Number of students who passed at least a course with a grade of C or better according to at least one of the data sources in the corresponding category. Students for whom the data sources align on reporting that they did not take a course in a given category in 2012–13 (46,937 minus the population reported in the table for each category) are not included in the table.**

Overall in 2012–13, the rates of exact alignment between the CALPADS dataset and the transcript file varied between 71% and 86% for all A-G categories except **college preparatory elective** (G) courses, which had the lowest rate of alignment at 30%. In comparison to the analyses between the CALPADS versus UC applications dataset and the CALPADS versus CSU applications dataset, the analysis between CALPADS and the transcript file shows the same patterns with a generally better rate of exact alignment. For instance, the CALPADS–transcript file analysis showed higher rates of exact alignment for most A-G subject areas in 2012–13 compared to the CALPADS–UC and CALPADS–CSU analyses for the same school year; the two exceptions concern the **mathematics** (C) and **laboratory science** (D) subject areas for the CALPADS–UC comparison, in which the exact alignment is higher (by only about 1%).

If there were courses passed in both the CALPADS dataset and the transcript file but there were differences between the two sources, then usually the transcript file reported more passed courses

(between 8% and 13% of the students depending on the category, with the exception of **college preparatory elective**). In that aspect, the CALPADS records were closer to the transcript file than the UC or CSU records where the percentage of students with fewer units in CALPADS was between 11% and 20%. And instances in which either the CALPADS dataset or the transcript file reported no courses passed usually amounted to fewer than 2% of the students; the two exceptions are the **laboratory science** (D) courses for which there were no courses passed in the transcript file for 13% of the students and the **college preparatory elective** (G) courses for which there were no courses passed in the transcript file for 47% of the students, a pattern also noted for the CALPADS–UC and CALPADS–CSU comparisons.

Table 14: Comparison between the CALPADS data and transcript files of the number of units (in years) passed by A-G subject area, 2013–14

Subject area	Number of students*	CALPADS number of units = transcript file number of units (percent)	CALPADS number of units > transcript file number of units (percent)	CALPADS number of units < transcript file number of units (percent)	In transcript file only (percent)	In CALPADS data only (percent)
A	46,068	79.5	6.7	11.2	1.7	1
B	46,071	80.4	3.3	14.1	1.6	0.6
C	43,924	79.7	4.9	11.8	1.8	1.8
D	46,920	65.5	3.2	8.5	2.0	20.9
E	24,244	88.2	2.7	5.9	2.5	0.7
F	23,307	77.5	4.6	8.8	4.8	4.4
G	10,522	11.2	2.6	1.7	14.2	70.3

Note: Row percentages may not add up to 100% due to rounding.

***Number of students who passed at least a course with a grade of C or better according to at least one of the data sources in the corresponding category. Students for whom the data sources align on reporting that they did not take a course in a given category in 2013–14 (46,920 minus the population reported in the table for each category) are not included in the table.**

Table 14 provides the comparison between the CALPADS dataset and the transcript files for the 2013–14 school year. Compared to the 2012–13 school year, the rates of exact alignment across the A-G subject areas were similar. Overall, the rates of exact alignment ranged from 66% to 88% across the A-G subject areas except **college preparatory elective** (G) courses, which had the lowest rate of alignment at 11%.

Similar to 2012–13, if there were courses passed in both the CALPADS dataset and the transcript file but there were differences between the two sources, then the transcript file more often reported more

passed courses. As shown for the 2012–13 comparison, instances in which either the CALPADS dataset or the transcript file reported no courses passed usually amounted to 5% or fewer of the students; the two exceptions were again the **laboratory science** (D) courses for which there were no courses passed in the transcript file (representing 21% of the students) and the **college preparatory elective** (G) courses for which there were no courses passed in the transcript file (representing 70% of the students) and no courses passed in the CALPADS dataset (14%).

Limitations and Main Challenges to Compute Eligibility Using CALPADS Course Records

Based on the analyses, WestEd found several specific challenges related to the use of incomplete and voluntarily submitted CALPADS course record data that should be investigated before those or subsequent CALPADS records are used for an eligibility estimate: 1) allocation of A-G courses to the different A-G categories, 2) rules to translate the various combinations of terms and marking periods into a number of units passed, 3) application of the validation rules, 4) inclusion of grades 7 and 8 course records, and 5) inclusion of test scores. These challenges are described in this section. Moreover, because three years of the CALPADS course record data analyzed in this study were voluntarily submitted and incomplete (CDE estimates 62–93% complete), these data are not representative of the potential for using mandatory CALPADS course record data collected during and after the 2014–15 school year (CDE estimates 98–99% complete).

1. Allocation of courses to A-G categories

In addition to student and school information, the CALPADS course records included local course code and description, state course code and description, A-G indicator and admission requirement code (A-G category), instructional-level code (including UC-certified Honors and college credit courses), academic term, marking period, final grade, credits attempted, and credits earned. A course was considered an A-G course based on the A-G indicator or if an A-G category was provided for the course. Three issues related to using the A-G categories in CALPADS can introduce bias in the estimate of A-G completion:

a. About 5% of the courses were identified as A-G courses, but no A-G category was provided. Those courses could not be identified as matching in the tables that compare the count of A-G courses by category presented above. The definition of an A-G category for each A-G–approved course is required to eliminate the undercount of the number of units passed for each A-G category when using the CALPADS data.

b. Students might use any A-G course to fill the **college preparatory elective** (G) requirement. When looking at the count of courses by year, the availability of an extra A-G course that could be used to meet the G requirement at the time of application to the UC or CSU is not taken into account in the present analysis.

c. The A-G categories in CALPADS included with the course records might differ from the ones used for the same courses by the college admissions offices. The allocation of courses to the **college preparatory elective** (G) category in CALPADS is described in Table 15.

Table 15: A-G admissions requirement codes included in the CALPADS records

A-G code	Number of courses	Percentage	Label	Description
A	1,110,251	8.6	History/social science	The course meets UC/CSU requirements for history/social science.
B	1,785,590	13.8	English	The course meets UC/CSU requirements for English.
C	1,810,956	14.0	Mathematics	The course meets UC/CSU requirements for mathematics.
D	1,033,921	8.0	Laboratory science	The course meets UC/CSU requirements for laboratory science.
E	905,393	7.0	Language other than English	The course meets UC/CSU requirements for a language other than English.
F	1,058,754	8.2	Visual and performing arts	The course meets UC/CSU requirements for visual and performing arts.
GA	965,580	7.5	History/social science elective	A preparatory elective in history/social science
GB	888,466	6.9	English elective	A preparatory elective in English
GC	481,273	3.7	Mathematics elective	A preparatory elective in mathematics
GD	842,296	6.5	Laboratory science elective	A preparatory elective in science
GE	483,716	3.7	Foreign language elective	A preparatory elective in a foreign language
GF	300,288	2.3	Visual and performing arts elective	A preparatory elective in visual and performing arts
GO	689,390	5.3	Other elective	A preparatory elective in any other subject area
All G	4,651,009	36.0	Any GA-GO code above	
Missing	585,942	4.5	Missing admissions requirement but marked as A-G course	

To check for matches with the college admissions datasets, a category A-G had to be defined for the GA, GB, … GO courses. WestEd looked at the two following possibilities: 1) categorizing all the GA, GB, … GO courses as preparatory electives (G courses) or 2) allocating back the GA courses to the A category, GB courses to the B category, and so on (but keeping the GO course as other elective). A best solution would be to analyze the state course code provided in CALPADS and, for example, determine whether the GA course can be used either only to meet the G requirement or both the A and G requirements.

A preliminary examination of the state description of courses categorized as GC (preparatory elective in mathematics) showed a large overlap between those courses and the courses categorized as C. As a

result, when all GA, GB, ... GF courses are coded as electives, a very low rate of matching by A-G category is obtained, with many courses coded as subject content electives in CALPADS appearing as coded in the subject content area (non-electives) in the college admissions databases (e.g., course coded as English elective GB in CALPADS but submitted as an English course under requirement B in the college admissions databases). For this analysis, WestEd has allocated back the elective courses to their content area (but kept the GO courses as other electives). However, the electives categories also contain courses that should not be used to meet the subject content requirement, and allocating all subject content electives to the subject content requirement increases the number of students who meet the subject content requirements and decreases the number of students who meet the electives requirement. A finer examination of how the courses are allocated to the subject content area or subject content electives would be required to eliminate that bias.

To fix this challenge over the next several years, the categorization of courses in CALPADS to preparatory electives or A-F courses would need to be reviewed so that it matches, for each year and school, the categorization used by the UC/CSU system. If this crosswalk is embedded into CALPADS, the categorization of courses could be realized with increased accuracy.

2. Terms and marking periods conversion

Course records obtained from the UC and CSU admissions datasets were mostly structured into semester terms with one grade per semester. Although the UC records included some courses reported as full-year courses (for which only one grade is reported per course and year), trimester courses (three grades reported), or quarterly courses (four grades reported), courses with these alternative terms only represented about 1% of the courses in the matched UC sample. The CSU records were mostly translated into semester terms and summer courses.

In contrast, the CALPADS dataset describes courses through a combination of term (full year, semester, trimester, quarter, and a few other possibilities) and marking periods. The combination of terms and marking periods presented a different and more complex range of possibilities than what was presented in the UC or CSU admissions datasets:

- Full-year terms were much more frequent. While they accounted for less than 1% of the courses in the UC dataset, over 30% of the CALPADS courses were reported as full-year courses, most of them with two marking periods, one for each semester. The different structure could have consequences when computing the total passed units per year.
- The combination of marking periods was not always consistent with the term recorded. While in most cases the sum of the marking periods added up to the value of the full term (e.g., course with a full-year term reported with two semester marking periods), cases of incomplete or contradictory course structure were not infrequent.

For example, it is unclear if a “full-year” term with a combination of marking periods adding up to less than a full year (about 10% of the courses with a full-year term) corresponded to a course that was not

completed or an error in the reporting of the marking period for a completed course. WestEd chose to be conservative and allocated units for each course according to the minimum of the two values — term and sum of marking periods. This result could be related to the undercounting of units observed in the calculations using CALPADS.

To resolve this challenge, a closer look or series of checks could be implemented in CALPADS to check the integrity of the combinations of terms and marking periods submitted by the schools.

3. No validation rules applied

No validation rules were applied for this analysis. Especially for **mathematics** (C), applying the validation rules would have required a full examination of the course titles rather than just the A-G subject area. While feasible because both local course codes and state course codes were provided in the CALPADS extract, such a task would fall more within the scope of work for a full eligibility study than the scope of work for this data feasibility study. Because no validation rules were applied in the datasets being compared (CALPADS, CSU, UC, or RTI International extracts), the bias in the comparison of the records should be limited. However, the full set of validation rules should be applied to compute a full eligibility indicator or compare the total number of units passed across high school years.

4. CALPADS course records for grades 7 and 8 not available for the cohort of 2014–15 graduates

To estimate eligibility status, some math courses taken in grades 7 and 8 can be used to meet the **mathematics** (C) requirements. CALPADS course records for 2011–12 were the earliest available and still quite incomplete. Any eligibility computed using CALPADS today would underestimate the number of students meeting the requirement. However, course records for the following years were more complete, and a future analysis might be able to use course records from 2014–15 and later, for example, to track course completion in grades 7 and 8 of a later cohort of graduates.

5. ACT and SAT test requirements

While not used for the direct comparison of courses, data were provided by CDE on ACT and SAT test requirements including SAT scores for all students who took the test from 2012 to 2015 and ACT scores for the graduating cohorts of 2012 to 2015. Test scores are used to estimate eligibility. Because the ACT and SAT datasets did not include SSIDs, data had to be matched to the population of 2015 graduates using a similar algorithm to that used for matching the CSU and UC admissions datasets. WestEd matched the population of 428,410 graduates to 149,638 SAT test takers and 104,229 ACT test takers. The match rates were fairly high with 94% of the SAT test takers and 96% of the ACT test takers linked to their CALPADS student record. In the absence of SSIDs in the test score datasets, those linkages will have to be realized for any eligibility study, but the high matching rates suggest that the bias introduced by the matching into the estimation of the availability of test scores should be limited.

Conclusion

This report provides the results of comparing the voluntarily submitted and incomplete 2011–2014 cohort CALPADS dataset to three other data sources: 1) UC admissions records, 2) CSU admissions records, and 3) transcript files collected directly from high schools. The comparisons examined, for each of the A-G subject areas, the rates of exact alignment in terms of the number of courses passed with a grade of at least C, the rate at which CALPADS showed a higher number of courses passed than the comparison source showed, the rate at which CALPADS showed a lower number of courses passed than the comparison source showed, the rate at which there were no courses passed in the CALPADS dataset (but for which there were courses passed in the comparison source), and the rate at which there were no courses passed in the comparison source (but for which there were courses passed in the CALPADS dataset).

In the CALPADS–UC comparison, the rates of exact alignment averaged approximately 75% (when weighted by the number of students) across the different A-G subject areas and across the years. In the CALPADS–CSU comparison, the rates of exact alignment averaged approximately 71% (when weighted by the number of students) across the different A-G subject areas and across the years. And in the CALPADS–transcript comparison, the rates of exact alignment averaged approximately 77% (when weighted by the number of students) across the different A-G subject areas and across the years. Because students with data indicating that they did not take a course in both data sources of each comparison were not included in the comparisons, the percentages of matches are a conservative estimate.

Across all three of the comparisons (CALPADS–UC, CALPADS–CSU, and CALPADS–transcripts), if there were differences between the CALPADS dataset and the comparison data source, then the comparison data source usually showed more courses being passed. With respect to the CALPADS–UC and the CALPADS–CSU comparisons, the **college preparatory elective** (G) courses always had the lowest rates of exact alignment across all of the A-G subject areas. In other words, accurate data may be available in CALPADS, but it is currently difficult to extract. The reason is that LEAs use their own SISs to populate CALPADS. CALPADS allows such electives to be categorized as both A-F and G with designations such as GA, GB, and so on. However, the SIS may not have that capability, so a course may be reported as an additional A-F course rather than a qualifying G course elective.

Finally, with respect to the GPA calculation, the CALPADS–UC and CALPADS–CSU comparisons were quite similar. More specifically, the CALPADS–UC comparison showed that 70% of the students had a calculated GPA that was less than one-tenth of one point different between the two data sources; the CALPADS–CSU comparison showed that 58% of the students had a calculated GPA that was less than one tenth of one point different between the two data sources.

Suggestions for potentially using CALPADS for the eligibility study in the future

CALPADS course record data for three of the school years analyzed in this study were voluntarily submitted and incomplete and, thus, may not be representative of the potential of using the CALPADS course record data mandated to be submitted since the 2014–15 school year. However, CALPADS data may be a viable substitute for transcript data if, as shown by the analyses above, a certain number of challenges in using 2011–2015 CALPADS course data for future eligibility studies is addressed in subsequent collections. Specifically, the challenges are:

- The categorization of courses into the electives category was problematic and led to a very low matching rate for that category. A review of the categorization of courses into A-G categories in CALPADS so that it matches, for each year and school, the categorization used by the UC/CSU system would solve that challenge. If this crosswalk is embedded into CALPADS, the categorization of courses could be realized with increased accuracy.
- The combinations of marking periods and terms were particularly complex and often did not add up to a clear description of the course length of instruction. Implementing a series of checks to verify the integrity of the combinations of terms and marking periods submitted by the schools would allow a better estimate of the number of units passed each year.
- Application of the validation rules requires looking beyond the A-G classifications at the specific course codes and labels. While CALPADS might be used in the future as an alternative for collecting school records through sampling upon resolution of the specific challenges highlighted above, transforming the different marking period systems and applying the set of validation rules require deep knowledge and information about the specific courses.

Appendix

Data sources

This study used as secondary sources extracts from administrative datasets from the California Longitudinal Pupil Achievement Data System (CALPADS), the University of California (UC) application data system, and the California State University (CSU) application data system.

CALPADS data sources

Data extracts, including the population of students who graduated in 2015, their names and demographic information, and all course information for those students from 2011–12 to 2014–15 school years were obtained from CALPADS. All extracts could be linked by the California Department of Education (CDE) Statewide Student Identifier (SSID).

In addition, school-level information files were downloaded from the CDE website for school year 2014–15. Files included free or reduced-price meal (FRPM) data, including the unduplicated counts and percentages of students eligible to receive FRPM under the National School Lunch Program (NSLP) and enrollment as well as a list of public schools and districts that included school type. School files could be linked to student-level data files using the County District School (CDS) code.

Finally, ACT and SAT scores for all students who took the test from 2012 to 2015 were provided by CALPADS. Those files did not include an SSID and had to be matched to the CALPADS population of 2015 graduates using a matching algorithm similar to the one described in the appendix section *Linked analysis datasets*.

UC data sources

Data extracts, including students' identification information (names, date of birth, SSID where available), students' demographic information, courses, and test assessments were obtained from the UC. The UC maintains a unique identification system across the campuses. Applicants' identification information, race/ethnicity, course records, and school information as well as test assessment records were received as a set of extracts that could be linked by a unique UC applicant identifier.

CSU data sources

Data extracts, including students' demographics information, courses, and test assessments were obtained from the CSU. Because the application to CSU campuses was disaggregated to each campus, WestEd received applicants' data separately for each campus.[7] There was no unique identification number for applicants across CSU campuses for the years of the study, and campuses identified a unique

[7] Data for the San Diego campus were not available.

student using an application number specific to each campus. WestEd received 22 data extracts, one for each campus, that included students' identification information (name, date of birth, SSID when available), demographic information, courses submitted for the application to each CSU campus with school CEEB (College Board), grades, and test information (SAT and ACT scores).

CALPADS, UC, and CSU populations of analysis

CALPADS 2014–15 population of graduates

The CDE sent a data extract for a population of 432,705 students who graduated in 2015; 428,435 of those students had graduated from grade 12. All students had an SSID, but a few duplicate SSIDs were included in the file, corresponding mainly to double school completion codes. WestEd gave priority to regular diplomas in order to unduplicate the population of graduates, obtaining a population analysis of 428,410 graduates.

UC applicants

A consolidated list of 90,533 graduates from California high schools who applied to the UC system was provided by the University of California Office of the President (UCOP). A school of record with a CDS code was provided for the population of California UC applicants and an SSID when available. Among the population of UC applicants, 76,324 (84%) students had an SSID and all had a CDS code.

CSU applicants

CSU campuses identify a unique student using an application number for each campus. WestEd observed that, in some cases, the same student was registered as an applicant under several application numbers for the same campus. To obtain a unique list of applicants and their records by campus and across campuses, the first step was to de-duplicate students within each campus. Next, because the same students could have applied to several campuses, a process to identify students across campuses was developed. Those two steps are detailed below.

De-duplication of CSU applicants within each campus

To identify students who might have submitted several applications to the same campus, combinations of a name, date of birth, and school of record were examined to identify unique students.

School of record: CSU data included a CEEB code (College Board) and a CSU local code. Using a crosswalk between a CEEB code, a CSU local code, and the school name, a CDS code was allocated for each course record, as available. A school of record was defined for each student by selecting the highest non-missing CDS code for the highest school grade level of the application. In the case of two application numbers for the same student, the CDS code selected was the one for the application with the highest number of course records. Schools corresponding to planned courses were not included in the process of selecting a school of record. If no CDS code was successfully linked to the CSU course data, school

information was selected in the following order: non-missing CEEB code, non-missing CSU local code, and local code.

Campus student-level list of students: To identify students who might have submitted several applications to the same campus, combinations of a SOUNDEX[8] transformation of the first part[9] of the first name, a SOUNDEX transformation of the first part of the last name, a SOUNDEX transformation of the middle name, date of birth, and CDS code were examined to identify unique students.

- If the combination appeared only one time in the campus data, records were identified as unique students.
- If the combination appeared several times in the campus data, students were considered the same if they had the same gender (Soundex tends to erase gender differentiation in names [e.g., Alberto and Alberta would be coded the same]).

Data were manually checked to ensure correspondence to an adequate aggregation level. The manual check consisted of examining full first name, last name, middle name, date of birth, school of enrollment, and California SSID for randomly selected duplicates. Those checks allowed WestEd to identify that the rule identified unique students in all cases examined given the provided information. Note that further duplicates were identified (e.g., one application with a middle name and another application for potentially the same student from the same campus without a middle name). However, the cases were rare enough and could not lead to a de-duplication rule without identifying records that were legitimately different students. Therefore, campus data contained some potential additional duplicate applications that could not be further de-identified using the information provided. Table A1 presents the number of applications by CSU campus defined by this process.

[8] SOUNDEX is an algorithm that codes a name as a short sequence of characters and numerals based on the way a name sounds rather than the way it is spelled. It was originally developed by Robert C. Russell and Margaret K. Odell in 1918. An updated version, the American SOUNDEX, was used in the 1930s for a retrospective analysis of United States censuses from 1890 through 1920. The National Archives and Records Administration (NARA) maintains the current set of rules that defines the algorithm for the official implementation of SOUNDEX used by the U.S. Government. The SAS built-in function SOUNDEX is based on the American SOUNDEX algorithm without the restriction to four characters.

[9] In cases of compound/hyphenated names, the first part of a name was defined by the presence of a blank or special character in the name.

Table A1: Number of applications by CSU campus before consolidation across campuses

Campus name	Original number of applicants (N=512,528)	De-duplicated number of applicants (N=507,416)	Percentage of applicants with SSID	Percentage of applicants with CDS code	Percentage of duplicates within campus
Bakersfield	8,827	8,808	44%	96%	0%
Channel Islands	22,891	22,886	41%	97%	0%
Chico	9,880	9,818	40%	97%	1%
Dominguez Hills	16,524	16,357	42%	94%	1%
East Bay	14,693	14,677	43%	97%	0%
Fresno	20,226	19,761	45%	97%	2%
Fullerton	42,057	41,812	42%	98%	1%
Humboldt	13,276	13,276	40%	93%	0%
Los Angeles	31,545	31,527	42%	96%	0%
Long Beach	58,012	57,980	43%	96%	0%
Maritime	1,320	1,320	31%	84%	0%
Monterey Bay	16,765	15,626	41%	97%	7%
Northridge	33,875	33,858	38%	97%	0%
Pomona	36,713	34,083	46%	97%	8%
Sacramento	23,304	23,212	42%	97%	0%
San Bernardino	13,821	13,810	42%	97%	0%
San Francisco	34,965	34,940	41%	96%	0%
San José	30,218	30,207	44%	96%	0%
San Luis Obispo	47,125	47,122	45%	92%	0%
San Marcos	13,887	13,860	41%	97%	0%
Sonoma	15,478	15,352	38%	98%	1%
Stanislaus	7,126	7,124	43%	98%	0%

De-duplication of CSU applicants across campuses

Data from all campuses were next consolidated into one dataset. A unique student was defined as a combination of a SOUNDEX transformation of the first part of the first name, a SOUNDEX transformation of the first part of the last name, a SOUNDEX transformation of the middle name, gender, and date of birth.

Specific checks were run to identify students' applications across campuses.

Unique combinations of names, gender, and date of birth should correspond to a unique SSID in the consolidated CSU data. Because the SSID was self-reported, WestEd checked that SSIDs indeed corresponded to unique students. Relatively rare cases of the same SSIDs for different students were examined and, in many cases, corresponded to slightly different versions of the names and particularly the presence of the middle name in one campus application but not another.

In cases of the same SSID matching several variations of names:

- If the students had the same first name and last name or the same SOUNDEX transformation of the first part of the first name and SOUNDEX transformation of the first part of the last name, they were considered the same students.
- If none of those conditions was true, students were considered different and one of the SSIDs was deleted.

Multiple combinations of names, gender, and DOB should correspond to a unique student applying to several campuses. WestEd checked that the multiple combinations were indeed the same students by identifying different SSIDs appearing for one combination of names, gender, and date of birth. In cases of students with same names having different SSIDs across the different campuses, the SSID of the name combination appearing in the files with the highest frequency was allocated, and the SSID was deleted from the records of the name combination appearing in the files with the lowest frequency.

From this process, the pool of 507,416 applications was reduced to 185,232 unique students applying to any CSU campus. Table A2 presents the number of applications across CSU campuses defined by this process.

Table A2: Number of CSU campuses applied to by CSU applicants

Number of applications to CSU campuses	Frequency	Percent
1	51,904	28
2	33,109	18
3	39,179	21
4	46,404	25
5	7,990	4
6	3,562	2
7	1,588	1
8+	1,496	1
Total	185,232	100

Among the resulting population of CSU applicants, 73,921 (40%) students had an SSID and 173,825 (94%) had a CDS code.

CALPADS, UC, and CSU course records

CALPADS course records

WestEd received a data extract of 18,513,846 course records including SSID, academic school year, school code and name, local course code and description, state course code and description, A-G indicator and admissions requirement code, instructional-level code (including UC-certified Honors and college credit courses), academic term, marking period, final grade, credits attempted, and credits earned.

Upon merging that extract with the population of 428,410 students who graduated from grade 12, WestEd obtained 18,392,085 course records for the 2015 graduates; 3,892 graduates had no course records available.

A-G indicator and admissions requirement code

A course was considered an A-G course based on the A-G indicator or if an A-G category was provided for the course. Sixty-seven percent of the course records were classified as A-G courses, but that percentage varied by year as presented in Table A3.

Table A3: Percentage of A-G courses per academic year for 2014–15 graduates

	2011–2012	2012–2013	2013–2014	2014–2015	Total
Percentage of A-G courses per academic year	59.4	69.6	72.0	66.2	67.2

Grades

Grades were coded as a character variable and were allocated a grade value as follows: A=4, B=3, C=2, D=1, F=0. Other grades were not allocated a grade value, and grades pass/fail were not included in the analysis. Among A-G courses, less than a half percent of the records (4,919) could not be allocated a grade value.

Term code

This element represents the term in which a given course section occurred. Term codes available are presented in Table A4.

Table A4: Academic terms

Abbreviation	Full Name	Description
FY	Full year	A session that lasts the full academic year
H1	First hexmester	The first of six hexmesters in an academic year
H2	Second hexmester	The second of six hexmesters in an academic year
H3	Third hexmester	The third of six hexmesters in an academic year
H4	Fourth hexmester	The fourth of six hexmesters in an academic year
H5	Fifth hexmester	The fifth of six hexmesters in an academic year
H6	Sixth hexmester	The sixth of six hexmesters in an academic year
IS	Intersession	An academic session that occurs during a short break during the academic year (not necessarily a longer, summer break), typical of year-round schools
Q1	First quarter	The first of four quarters of an academic year
Q2	Second quarter	The second of four quarters of an academic year
Q3	Third quarter	The third of four quarters of an academic year
Q4	Fourth quarter	The fourth and final quarter of an academic year
S1	First semester	The first of two semesters in an academic year
S2	Second semester	The second of two semesters in an academic year
SP	Supplemental session	A session that occurs on evenings, after school, or on weekends
SS	Summer session	An academic session that occurs during the summer break
T1	First trimester	The first of three trimesters in an academic year
T2	Second trimester	The second of three trimesters in an academic year
T3	Third trimester	The third of three trimesters in an academic year
Z1	Other first term	The first term in a set of terms not otherwise defined in this code set
Z2	Other second term	The second term in a set of terms not otherwise defined in this code set
Z3	Other third term	The third term in a set of terms not otherwise defined in this code set
Z4	Other fourth term	The fourth term in a set of terms not otherwise defined in this code set
Z5	Other fifth term	The fifth term in a set of terms not otherwise defined in this code set
Z6	Other sixth term	The sixth term in a set of terms not otherwise defined in this code set
Z7	Other seventh term	The seventh term in a set of terms not otherwise defined in this code set
Z8	Other eighth term	The eighth term in a set of terms not otherwise defined in this code set
Z9	Other ninth term	The ninth term in a set of terms not otherwise defined in this code set

There was not enough information in the dataset to be able to interpret terms classified as "other term lengths," and those were not included in the estimate of the number of A-G courses for admission. All other terms were allocated an annual prorated term value. For example, completing an annual course (term=FY) would be associated with a term value of 1, a semester course (term=S1 or S2) would be associated with a term value of 0.5, a quarter course (term=Q1, Q2, Q3, or Q4) would be associated with a term value of 0.25, and a trimester course (term=T1, T2, or T3) would be associated with a term value of 0.34.

Marking period

The marking period code is used to report the period within a course section in which a course mark is given to a student for a particular grade. The term and the marking period of the course are recorded into two different variables. For example, a course can be an annual course, and grades are available for two semesters or four quarters. Marking periods have the same possible values as the academic terms presented above and were also allocated an annual prorated value. There was not enough information in the data to be able to interpret marking periods classified as "other period lengths," and those were not included in the estimate of the number of A-G courses for admission.

Course aggregation

Using the local course code included in the data, term values were aggregated for each course/term by adding the annual prorated values of marking periods corresponding to the same course. For each student each year, WestEd computed the sum of A-G courses taken by A-G category by adding the annual prorated values of the combinations of terms and marking periods, as well as the sum of A-G courses passed with a grade of C or better by A-G category.

UC course records

A K–12 course extract was also provided for the population of UC applicants. Table A5 presents the distribution of number of courses for UC applicants.

Table A5: Distribution of number of courses for UC applicants

	Mean	Median	Number of students	Minimum	Maximum
Number of courses	22.8	22.0	90,533	1	71

The UC course record included a term variable. For courses reported as full-year courses only one grade was reported per course and year, for trimester courses three grades were reported, and for quarterly courses four grades were reported. Grade values and prorated term values were computed as described in the CALPADS course section. For each student each year, WestEd computed the sum of A-G courses taken by A-G category, as well as the sum of A-G courses passed with a grade of C or better by A-G category.

CSU course records

Course records, including course label, A-G classification, Honors indicator, and grades were included in each campus data extract for CSU applicants. From the set of courses submitted at each campus by CSU applicants, the final set of courses was defined as a de-duplicated set of courses; subject areas; grade level; grades in fall, spring ,and summer; as well as the presence of any Honors classes. WestEd checked the distribution of number of courses by application in the original data (de-duplicated within campus) after they were de-duplicated within campuses and across campuses. A large increase in the number of courses per student would signal the matching of students with different sets of courses and was used as a quality check to define the appropriate criteria for aggregation.

Table A6: Distribution of number of courses for CSU applicants

	Mean	Median	Number of students	Minimum	Maximum
Number of courses by original application	22.9	23.0	512,528	1	78
Number of courses after consolidation within campus	22.9	23.0	507,416	1	78
Number of courses after aggregation across campuses	23.4	23.0	185,232	1	82

Courses were already organized in semesters (fall or spring) or summer courses (summer 1 or summer 2). Term values were aggregated for each course by adding the annual prorated values of periods corresponding to the same course. Grade values were computed as described in the CALPADS course section.

For each student each year, WestEd computed the sum of A-G courses taken by A-G category, as well as the sum of A-G courses passed with a grade of C or better by A-G category.

Linked analysis datasets

Although each system — CALPADS, UC, and CSU — has its own unique student identifier, no systematic common identifier links a student among those systems. As noted before, the UC and CSU application databases include a self-reported SSID that can be used to link to the CALPADS extracts, but it was not included for every student and the quality of the indicator was unknown. Therefore, a process to match the records for each individual student across the systems was developed.

This study used a deterministic and fuzzy sequential matching process in which the names of individuals, as well as date of birth and school, were used to link across the databases.

The general methodology for constructing the linked analysis dataset is described below. The matching results for the CALPADS–UC and the CALPADS–CSU linkages are presented in the next section.

Preparation for making the match

Before starting the matching process, students' first name, last name, and date of birth were thoroughly examined to evaluate their discriminating power and the presence of compound/hyphenated names. Additional variables available in both datasets (i.e., middle name, gender, ethnicity, school) were also examined and researchers set up a process of quality control for the matching process.

Discriminating power of the matching fields

Since CALPADS data represent the population of students to which WestEd was matching, WestEd examined the specificity of the planned matching variables on this dataset: out of 428,410 students who graduated in 2014–15, about 500 combinations of first names, last names, and dates of birth appeared more than one time, representing a percentage of duplicate values on the matching variables of about a tenth of a percent (0.12%). When adding middle name and/or school of graduation to sort out the duplicates, WestEd was able to de-duplicate virtually all records that had this information available.

Compound/hyphenated names

The name fields were evaluated for the presence of compound/hyphenated names (names with two or more words separated by a blank or a special character in the same data field) because the presence of several names in a field can create difficulties in accurately matching individuals across datasets. The percentage of compound/hyphenated first and last names in the CALPADS dataset was 2% and 9%, respectively. Accordingly, strategies that use only the first name of a compound name were included in the matching process.

In cases of compound/hyphenated names, two versions of each name were kept in two separate fields: one corresponding to the name as it was provided with no blank or separator, and one storing only the first part (as defined by the presence of a blank or special character) of the compound/hyphenated name. Fields were used sequentially in the matching process.

Process for making the match

The matching process was developed as six successive steps written in SAS software.[10] The process used a sequence of deterministic and fuzzy matches using the SAS software SOUNDEX function.[11] From one step to the next, only the residual records — those not matched in a previous step — were kept in the pool to be matched in a subsequent step.

Step 1 of the process linked two datasets using the self-reported SSID. Step 2 of the process matched only those students who were unmatched in the first step and linked the exact text strings[12] recorded for the first name and last name, the initial of the middle name, the date of birth, and the school of record to match across two datasets. Step 3 matched only those students who were unmatched in the

[10] Version 9.3 of the SAS System for Windows. Copyright © 2002–2003 SAS Institute Inc.

[11] An algorithm that codes a name as a short sequence of characters and numerals based on the way it sounds.

[12] All text fields were cleaned up and set to lower case; symbols and other special characters and blanks were deleted.

previous steps by first name, last name, date of birth, and school of record. Step 4 matched by first name, last name, date of birth, and gender.

Because of the prevalence of compound/hyphenated names and potential difference in spelling, steps 5 and 6 were structured to capture different combinations for recording compound/hyphenated names along with the birth date. Step 5 of the match used the first word (as separated by a blank or special character) from the first name and the first word in the last name transformed using the SOUNDEX function along with gender, middle initial, and school of record; step 6 also used the transformed names along with gender and school of record.

At each step, the set of students from the UC or CSU admissions datasets who matched exactly only one student in the CALPADS dataset were kept as final matches, while the set of students for whom there were duplicate matches in the CALPADS dataset were not included as matches.

CALPADS–UC linkage results

The population of 410,518 graduates was matched to the population of 90,533 UC applicants. The matching rates by step are reported in Table A7.

Table A7: Matching rate to CALPADS population for UC applicants

Matching variables	Unique matches	Matching rate
SSID	70,486	78%
First and last names, date of birth, middle initial, CDS code	13,099	14%
First and last names, date of birth, CDS code	3,918	4%
First and last names, date of birth, gender	590	1%
Soundex of first part of names, date of birth, CDS code	337	0%
Soundex of first part of names, date of birth, middle initial, CDS code	383	0%
Total matching (out of 90,533 UC applicants)	**88,813**	**98%**

Note: Sums may not total due to rounding.

Due to a much higher inclusion of SSIDs in the UC records, nearly all applicants could be matched with a record in CALPADS, 78% using the SSID alone.

CALPADS–CSU linkage results

Using the process described above, the population of 410,518 graduates was matched to the population of 185,232 CSU applicants. The matching rates by step are reported in Table A8.

Table A8: Matching rate to CALPADS population for CSU applicants

Matching variables	Unique matches	Matching rate
SSID	67,532	36%
First and last names, date of birth, middle initial, CDS code	55,976	30%
First and last names, date of birth, CDS code	16,461	9%
First and last names, date of birth, gender	4,776	3%
Soundex of first part of names, date of birth, CDS code	2,250	1%
Soundex of first part of names, date of birth, middle initial, CDS code	1,630	1%
Total matching (out of 185,232 CSU applicants)	**148,625**	**80%**

The inclusion of SSIDs in the CSU application records allowed a straight match of over one third of the students to their CALPADS records. Among those without an SSID (or an incorrectly reported SSID that could not be linked to the CALPADS records), an additional 30% could be matched, with a one-to-one match, to their CALPADS records using names, date of birth, middle initial, and CDS code of the school of record. All additional steps added about 15% additional matches. The total match rate of 80% is considered a reasonable match rate of populations across two different systems.

Reference

Silver, D., Hensley, E., Hong, Y., Siegel, P., & Bradby, D. (2017). *University eligibility study for the public high school class of 2015.* Research Triangle Park, NC: RTI International.