## Peer Groups for Comparing the Completion Rate in the 2016 Scorecard

### Background

In 2004, state legislation prompted the design and implementation of a performance measurement system for the California Community Colleges (CCC) known as the Accountability Reporting for the Community Colleges (ARCC). The California Community College system is the largest postsecondary educational system in the world, serving more than 2.1 million students in 2014-15, with 113 colleges campuses spread across 72 districts. The locally controlled colleges, each with multiple and complex missions, provide a variety of educational programs to a diverse student population in assorted communities throughout California.

California has recognized student and community diversity among its colleges and the importance of accounting for this diversity when comparing institutional performance. The diverse academic and economic environments of the students served by a college are important factors affecting individual student achievements and overall institutional performance. In evaluating performance, the Chancellor's Office has historically captured institutional differences through adjustment factors or selection variables. In 2007, ARCC used peer grouping to examine a college's performance for each of the seven college level indicators in the accountability report.

The development of the peer groups for each indicator included the selection of the most appropriate variables using bivariate correlations and hierarchical regression. This process assured that the environmental factors had an empirical, as well as a theoretical relationship with the performance indicator. To identify the members of each particular peer group, a classifications method known as cluster analysis was used. Using the same methodology for peer grouping as previous but with updated predictor variables, the Chancellor's Office has produced a new set of peer groups for the Completion Rate. The colleges can use the peer groups for comparing themselves on this indicator with similar colleges for evaluative purposes.

### Methodology

A preliminary step to finding the peer group for each college was to develop regression models to identify a parsimonious set of uncontrollable factors that predicted the Completion Rate. The potential uncontrollable factors, or predictor variables, were initially identified through an extensive literature review and have continued to be refined over the years. The factors that affect the outcome had to lie beyond the control of each college administration (uncontrollable factors often referred to as “environmental factors”) and be available through a feasible data source that the Chancellor’s Office (CCCCO) can use.

Using the parsimonious set of uncontrollable factors identified by regression modeling, cluster analysis (a standard multivariate statistical tool) was used to identify those colleges that most closely resemble the college of interest in terms of these uncontrollable factors on the specific performance metric.

Cluster analysis is a well-developed quantitative method of identifying groups of entities from a population of entities.  Major references for cluster analysis became available to researchers as early as 1963 (Sokal & Sneath, 1963).  This method can apply to any kind of entity, and past applications have clustered entities as diverse as colleges, states, cities, students, sports teams and players, patients, hospitals, and businesses, to mention a few.  In past years, researchers have used it for developing taxonomies, especially with respect to the biological studies (i.e., horticulture, zoology, and entomology).

Depending upon the objective of the researcher, the cluster analysis chooses one or more measurements (aka “variables”) of each entity in a population to produce a numerical indicator of “distance” between each entity in a given population.  The researcher’s objective is imperative in that this will drive the choice of measurements that more or less “determine” the eventual groupings or clusters.  If the researcher chooses measurements that poorly reflect the researcher’s objective, then the cluster analysis will probably produce a grouping that has marginal validity, if any.

Based upon the aforementioned inter-entity distances, cluster analysis then proceeds to identify sets of entities within a defined population by comparing sets of distances.  In the vernacular of cluster analysis, these distances are also called “proximities.”  If the population under study contains a very unique entity in it, then the cluster analysis may produce, among its groupings, a cluster of one (i.e., a group containing only one case) to preserve the uniqueness of this one entity with respect to the population under study and the researcher’s objective.

A procedure known as hierarchical clustering moves through a large number of iterations to progressively “join” one college to another college that the computer finds is its “closest neighbor.” The program will then join this resulting pair to the next most similar college (the next closest neighbor), and so on until no other colleges of sufficient similarity can be joined to this initial set.  The procedure then repeats this “joining” process for each of the remaining colleges that the program has not already joined with some other college.  The peer grouping used this well-established procedure.

Standard options for conducting a cluster analysis method were reviewed and the following steps for peer grouping were used:

- Define a practical number of clusters to be identified
- Select a proximity measure that effectively captures the difference or "distance" between colleges on the basis of their levels of analyst-specified variables.
- Select and use a cluster identification algorithm that applies a specific decision rule (i.e., a type of logic) to cluster the colleges into mutually exclusive groups.
- Prevent bias in the clustering that may result from using variables that use different scales of measurement (i.e., driving miles vs. student headcounts or percentage of students, and so forth).

The following section reports on how the four steps listed above were implemented.

- The peer grouping identifies seven distinct peer groups for all the community colleges in the system. This "target" of seven groups addressed administrative concerns over the identification of too many peer groups and a plethora of single-college peer groups (that is, the finding of some colleges that lacked any statistical peers for comparison).
- The chosen measure of distance between each community college in the system is the so-called squared Euclidean distance. This is the most common measure of proximity in cluster analysis.
- For the peer grouping Ward's method for clustering was used because this method was found to work well with the data. According to Bailey (1994), Ward's method "begins with each object treated as a cluster of one. Then objects are successively combined. The criterion for combination is that the within-cluster variation as measured by the sum of within-cluster deviation from cluster means (error sum of squares) is minimized. Thus, average distances among all members of the cluster are minimized." Ward's method has a tendency to produce clusters of approximately similar size, such as the number of members in each cluster (Everitt, Landau, & Leese, 2011).
- Each measure was converted so that different units of measurement would have no effect upon the clustering solutions. These measures were converted by standardizing the variables to unit variance (also known as converting measurements to z-scores). This can be performed using the following formula (Snedecor & Cochran, 1980):

  *z = (raw score for a case – mean of the sample) / (standard deviation of the sample)*

# Peer Groups for Comparing Performance on the Completion Rate in the 2016 Scorecard

| Group 1: | Group 2: | Group 3: | Group 4: | Group 5: | Group 6: | Group7: |
|---|---|---|---|---|---|---|
| ANTELOPE VALLEY | CUESTA | BUTTE | COMPTON | ALLAN HANCOCK | ALAMEDA | BERKELEY CITY |
| BAKERSFIELD | CYPRESS | CHAFFEY | EAST L.A. | AMERICAN RIVER | CABRILLO | CANADA |
| BARSTOW | DE ANZA | CITRUS | L.A. TRADE-TECH | CANYONS | CHABOT | FOOTHILL |
| CERRITOS | DIABLO VALLEY | COSUMNES RIVER | LASSEN | CERRO COSO | CONTRA COSTA | IRVINE VALLEY |
| COALINGA | FOLSOM LAKE | CRAFTON HILLS | PALO VERDE | COASTLINE | L.A. CITY | MARIN |
| COPPER MOUNTAIN | FULLERTON | EL CAMINO | SAN BERNARDINO | COLUMBIA | L.A. VALLEY | SADDLEBACK |
| DESERT | GLENDALE | EVERGREEN VALLEY | SANTA ANA | CUYAMACA | LANEY | SAN DIEGO MIRAMAR |
| FRESNO CITY | GOLDEN WEST | GROSSMONT | SOUTHWEST L.A. | FEATHER RIVER | LONG BEACH CITY | SAN FRANCISCO CITY |
| HARTNELL | L.A. PIERCE | MT. SAN ANTONIO | TAFT | GAVILAN | MERRITT | WEST VALLEY |
| IMPERIAL VALLEY | LAS POSITAS | MT. SAN JACINTO | | LAKE TAHOE | SAN DIEGO CITY | |
| L.A. HARBOR | MOORPARK | REEDLEY | | MIRA COSTA | SAN JOSE CITY | |
| L.A. MISSION | OHLONE | SACRAMENTO CITY | | MISSION | WEST L.A. | |
| LEMOORE | ORANGE COAST | SHASTA | | MONTEREY | | |
| LOS MEDANOS | PALOMAR | SOLANO | | NAPA VALLEY | | |
| MENDOCINO | PASADENA CITY | VENTURA | | REDWOODS | | |
| MERCED | SAN DIEGO MESA | WOODLAND | | SANTA ROSA | | |
| MODESTO | SAN MATEO | | | SANTIAGO CANYON | | |
| OXNARD | SANTA BARBARA CITY | | | SISKIYOUS | | |
| PORTERVILLE | SANTA MONICA | | | | | |
| RIO HONDO | SIERRA | | | | | |
| RIVERSIDE | SKYLINE | | | | | |
| SAN JOAQUIN DELTA | | | | | | |
| SEQUOIAS | | | | | | |
| SOUTHWESTERN | | | | | | |
| VICTOR VALLEY | | | | | | |
| YUBA | | | | | | |

**Results of Revised Peering Grouping**

The development of college-level services area indices that represent the economic and education characteristics or environments of the student served have been useful as predictor variables in the initial accountability framework (van Ommeren, Liddicoat & Hom, 2008). The Chancellor's Office has updated these indices with current Census data, as well as explored additional indices such as the Academic Performance Index.

The predictors for the Completion Rate (2009-10 to 2014-15) are:

- API: The Academic Performance Index is an index calculated by the California Department of Education for each high school in the state based on standardized test scores in a number of subjects. A variable of this index was developed by the CCCCO that assigns a weighted API (based on 2008 API) to each college based on the proportion of enrolled students from a given high school (Fall 2009).

- BA Index: The Bachelor of Arts/Sciences Index represents the bachelor degree attainment of the population, 25 years or older in a college's service area. This index, created by CCCCO, combines the enrollment patterns (Fall 2009) of students by ZIP code of residence with educational data for ZCTA (ZIP Code Tabulation Area) codes obtained from the American Community Survey.

- Pct Age 25+: The percentage of students at a community college in the Fall of 2009 that are age 25 years or older, obtained from the CCCCO MIS.

To assist users evaluate the data completeness of each predictor, the percent of students with missing information by college is shown in Appendix.

The table below shows the regression coefficients of predictors at each step of the hierarchical model predicting the Completion Rate. The complete model has an adjusted $R^2$ = .69 with the regression coefficients for all predictors significant at the .05 level. Based on the standardized beta coefficients, the BA+ provides the largest relative contribution to the model. Multicollinearity is negligible in the final regression and the residuals appeared to be normally distributed.

*Hierarchical Regression Analysis Summary for the Completion Rate (2009-10 to 2014-15)*

| Step | Variables | B | Std. Error | Standardized Coefficients | Model $R^2$ |
|---|---|---|---|---|---|
| **1** | (Constant ) | -60.7 | 9.2 | | |
| | API | .15 | .01 | .74 | .55 |
| **2** | (Constant) | -34.0 | 9.9 | | |
| | API | .10 | .02 | .49 | |
| | BA+ | .31 | .06 | .38 | .63 |
| **3** | (Constant) | - 16.8 | 9.8 | | |
| | API | .08 | .01 | .43 | |
| | BA+ | .36 | .06 | .45 | |
| | 25+ | -.21 | .05 | -.25 | .69 |

**Discussion**

The first variable entered was a composite Academic Performance Index (API) score for each college. This weighted API was calculated by the Chancellor's Office based on the proportion of students from a given high school at each college. This weighted API acts as a proxy of K-12 academic preparation which literature has shown to be a significant predictor of college success.

Entered next was a community based predictor variable, the Bachelor Plus Index. This college level variable, also developed by the Chancellor's Office, reflects the educational attainment of the population 25 years old and over for the service area of the college. Research indicates that a major predictor of college success is the level of parent education. In addition, studies indicate that the socioeconomic background of an area has a link to educational outcomes of those who grow up in a neighborhood (the so-called "neighborhood effect"). The BA Index might be considered a proxy for these other variables or a combination of such variables in the broader context of a community's socioeconomics.

The last variable entered was percent of students 25 years old and over is negatively associated with the student progress and achievement rate. Possibly, colleges with greater percentages of "older" students focus on education that does not include a certificate, degree or outcomes related to transfer. For example, older students might already be in the workforce but continue to take courses to enhance their job skills or other interests without degree or transfer as their goal.

**References**


Bailey, K. (1994). Methods of Social Research. Fourth Edition. New York. The Free Press.

Everitt, B.S., Landau, S., Leese, M. and Stahl, D. (2011). Cluster Analysis. Fifth edition. Wiley.

Hom, W. (2008). Peer Grouping: The Refinement of Performance Indicators. *Journal of Applied Research in the Community College, 15*(2), 49-55.

Snedecor, G.W. and Cochran, W.G. (1980). Statistical Methods. Seventh Edition. Ames Iowa: The Iowa State University Press.

Sokal, R. R., Sneath, P. H. (1961). Principles of Numerical Taxonomy. San Francisco. W.H. Freeman & Co.

Van Ommeren, A., Liddicoat, C. and Hom, W. (2008). Developing Service Area Indices for Community Colleges: California's Method and Experience. *Community College Journal of Research and Practice, 32*, 463-479.

| | Means of Predictors | | | Completion Rate for the 2009-10 cohort (Overall) | | | Peer Group Colleges | |
|---|---|---|---|---|---|---|---|---|
| Peer Group | API | Bachelor Plus Index | Pct Students Age 25+ | Lowest Peer | Highest Peer | Average | # Peers | Colleges in the Peer Group |
| 1 | 690 | 17.7% | 41.0% | 25.3% | 47.8% | 39.3% | 26 | ANTELOPE VALLEY, BAKERSFIELD, BARSTOW, CERRITOS, COALINGA, COPPER MOUNTAIN, DESERT, FRESNO CITY, HARTNELL, IMPERIAL VALLEY, L.A. HARBOR, L.A. MISSION, LEMOORE, LOS MEDANOS, MENDOCINO, MERCED, MODESTO, OXNARD, PORTERVILLE, RIO HONDO, RIVERSIDE, SAN JOAQUIN DELTA, SEQUOIAS, SOUTHWESTERN, VICTOR VALLEY, YUBA |
| 2 | 763 | 37.7 | 36.8 | 47.2 | 65.6 | 54.3 | 21 | CUESTA, CYPRESS, DE ANZA, DIABLO VALLEY, FOLSOM LAKE, FULLERTON, GLENDALE, GOLDEN WEST, L.A. PIERCE, LAS POSITAS, MOORPARK, OHLONE, ORANGE COAST, PALOMAR, PASADENA CITY, SAN DIEGO MESA, SAN MATEO, SANTA BARBARA CITY, SANTA MONICA, SIERRA, SKYLINE |
| 3 | 725 | 25.4 | 37.3 | 39.0 | 53.0 | 45.6 | 16 | BUTTE, CHAFFEY, CITRUS, COSUMNES RIVER, CRAFTON HILLS, EL CAMINO, EVERGREEN VALLEY, GROSSMONT, MT. SAN ANTONIO, MT. SAN JACINTO, REEDLEY, SACRAMENTO CITY, SHASTA, SOLANO, VENTURA, WOODLAND |
| 4 | 655 | 17.2 | 56.9 | 23.6 | 45.2 | 34.0 | 9 | COMPTON, EAST L.A., L.A. TRADE-TECH, LASSEN, PALO VERDE, SAN BERNARDINO, SANTA ANA, SOUTHWEST L.A., TAFT |
| 5 | 742 | 28.6 | 56.3 | 33.5 | 57.0 | 46.4 | 18 | ALLAN HANCOCK, AMERICAN RIVER, CANYONS, CERRO COSO, COASTLINE, COLUMBIA, CUYAMACA, FEATHER RIVER, GAVILAN, LAKE TAHOE, MIRA COSTA, MISSION, MONTEREY, NAPA VALLEY, REDWOODS, SANTA ROSA, SANTIAGO CANYON, SISKIYOUS |
| 6 | 688 | 33.0 | 49.0 | 33.6 | 51.8 | 42.1 | 12 | ALAMEDA, CABRILLO, CHABOT, CONTRA COSTA, L.A. CITY, L.A. VALLEY, LANEY, LONG BEACH CITY, MERRITT, SAN DIEGO CITY, SAN JOSE CITY, WEST L.A. |
| 7 | 775 | 48.8 | 53.4 | 45.4 | 64.9 | 54.2 | 9 | BERKELEY CITY, CANADA, FOOTHILL, IRVINE VALLEY, MARIN, SADDLEBACK, SAN DIEGO MIRAMAR, SAN FRANCISCO CITY, WEST VALLEY |
| State-wide* | **721** | **28.5** | **45.3** | — | — | **45.3** | **111** | |

*: These are the averages of all community colleges (n=111).

Appendix

Percent of students with missing age, zip code, or High School are shown below. The peer groups are created using only non-missing data on students.

* For High School information, students who; 1) were 22 or older or special admit, 2) did not go to a high school, or 3) went to one out-of-state, are excluded from the calculation.

** Four colleges did not report High School information on students enrolled in Fall 2008 and 2009, therefore, information from Fall 2007 was used.

*** Two colleges did not report High School information on students enrolled in Fall 2009, therefore, information from Fall 2008 was used.

| College | Percent of missing age (Fall 2009) | Percent of missing zip code (Fall 2009) | Percent of missing High School* (Fall 2009) |
|---|---:|---:|---:|
| ALAMEDA | 0.0 | 3.0 | 3.7** |
| ALLAN HANCOCK | 0.0 | 0.0 | 16.6 |
| AMERICAN RIVER | 0.0 | 0.2 | 11.9 |
| ANTELOPE VALLEY | 0.0 | 0.1 | 27.2 |
| BAKERSFIELD | 0.0 | 0.2 | 5.7 |
| BARSTOW | 0.2 | 4.4 | 1.2 |
| BERKELEY CITY | 0.0 | 3.4 | 4.7** |
| BUTTE | 0.1 | 2.3 | 3.4 |
| CABRILLO | 0.0 | 0.0 | 22.2 |
| CANADA | 0.0 | 9.7 | 3.6 |
| CANYONS | 0.0 | 0.1 | 29.6 |
| CERRITOS | 1.2 | 0.3 | 23.9 |
| CERRO COSO | 0.0 | 0.0 | 23.3 |
| CHABOT HAYWARD | 0.0 | 0.0 | 1.2 |
| CHAFFEY | 0.0 | 0.0 | 17.5 |
| CITRUS | 0.0 | 9.3 | 14.9 |
| COASTLINE | 0.1 | 7.1 | 19.6 |
| COLUMBIA | 0.0 | 0.2 | 16.1 |
| COMPTON | 0.0 | 0.0 | 5.1 |
| CONTRA COSTA | 0.5 | 0.8 | 26.6 |
| COPPER MOUNTAIN | 0.4 | 0.0 | 77.0*** |
| COSUMNES RIVER | 0.0 | 0.1 | 8.3 |
| CRAFTON HILLS | 0.0 | 0.1 | 35.5 |
| CUESTA | 0.1 | 16.1 | 6.8 |
| CUYAMACA | 6.0 | 27.7 | 13.8 |
| CYPRESS | 0.0 | 1.7 | 0.1 |
| DEANZA | 0.0 | 0.0 | 3.7 |
| DESERT | 0.1 | 0.2 | 2.6 |
| DIABLO VALLEY | 0.1 | 1.0 | 22.7 |
| EAST LA | 0.0 | 0.0 | 3.9 |

| College | Percent of missing age (Fall 2009) | Percent of missing zip code (Fall 2009) | Percent of missing High School* (Fall 2009) |
|---|---|---|---|
| EL CAMINO | 0.0 | 0.0 | 1.3 |
| EVERYGREEN VALLEY | 0.0 | 0.1 | 55.5 |
| FEATHER RIVER | 0.6 | 0.8 | 25.3 |
| FOLSOM LAKE | 0.0 | 0.1 | 6.6 |
| FOOTHILL | 0.1 | 0.1 | 6.3 |
| FRESNO CITY | 0.0 | 1.9 | 71.5 |
| FULLERTON | 0.0 | 1.7 | 0.0 |
| GAVILAN | 0.1 | 0.6 | 5.8 |
| GLENDALE | 28.3 | 0.2 | 0.9 |
| GOLDEN WEST | 0.0 | 14.5 | 7.1 |
| GROSSMONT | 0.0 | 1.6 | 8.8 |
| HARTNELL | 0.3 | 0.0 | 18.0 |
| IMPERIAL | 0.0 | 1.4 | 30.5 |
| IRVINE | 0.0 | 0.0 | 6.3 |
| LA CITY | 0.0 | 0.0 | 1.4 |
| LA HARBOR | 0.0 | 0.0 | 1.1 |
| LA MISSION | 0.0 | 0.0 | 1.8 |
| LA PIERCE | 0.0 | 0.0 | 0.1 |
| LA SWEST | 0.0 | 0.0 | 2.2 |
| LA TRADE | 0.0 | 0.0 | 2.6 |
| LA VALLEY | 0.0 | 0.0 | 0.8 |
| LAKE TAHOE | 0.1 | 0.0 | 18.0 |
| LANEY | 0.0 | 4.2 | 9.7** |
| LAS POSITAS | 0.0 | 0.0 | 0.4 |
| LASSEN | 0.1 | 0.1 | 59.2 |
| LONG BEACH | 0.2 | 0.1 | 22.1 |
| LOS MEDANOS | 0.1 | 1.0 | 15.7 |
| MARIN | 0.1 | 0.0 | 8.3 |
| MENDOCINO | 0.0 | 0.8 | 7.8 |
| MERCED | 0.7 | 1.7 | 4.3 |
| MERRITT | 0.0 | 4.6 | 3.6** |
| MIRA COSTA | 0.0 | 0.1 | 19.9 |
| MISSION | 0.2 | 0.9 | 38.2 |
| MODESTO | 0.0 | 0.1 | 8.7 |
| MONTEREY | 0.0 | 0.4 | 0.5 |
| MOORPARK | 0.0 | 3.7 | 3.3 |
| MT SAN ANTONIO | 0.1 | 6.6 | 5.3 |
| MT. SAN JACINTO | 0.1 | 0.0 | 20.9 |
| NAPA | 0.2 | 0.2 | 8.9 |
| OHLONE | 0.0 | 0.1 | 7.2 |
| ORANGE COAST | 0.0 | 11.1 | 9.1 |

| College | Percent of missing age (Fall 2009) | Percent of missing zip code (Fall 2009) | Percent of missing High School* (Fall 2009) |
|---|---|---|---|
| OXNARD | 0.0 | 2.4 | 2.0 |
| PALO VERDE | 1.0 | 0.0 | 45.4*** |
| PALOMAR | 0.0 | 0.1 | 7.0 |
| PASADENA | 0.0 | 0.3 | 6.7 |
| PORTERVILLE | 0.0 | 0.3 | 6.2 |
| REDWOODS | 0.0 | 0.6 | 3.3 |
| REEDLEY COLLEGE | 0.0 | 1.0 | 62.6 |
| RIO HONDO | 0.0 | 0.0 | 17.0 |
| RIVERSIDE | 0.0 | 2.5 | 7.6 |
| SACRAMENTO CITY | 0.0 | 5.9 | 24.2 |
| SADDLEBACK | 0.0 | 0.1 | 5.1 |
| SAN BERNARDINO | 0.0 | 0.0 | 34.7 |
| SAN DIEGO CITY | 0.0 | 0.0 | 6.3 |
| SAN DIEGO MESA | 0.0 | 0.0 | 6.1 |
| SAN DIEGO MIRAMAR | 0.0 | 0.0 | 4.4 |
| SAN FRANCISCO | 0.0 | 4.2 | 5.7 |
| SAN JOAQUIN DELTA | 0.0 | 0.0 | 3.4 |
| SAN JOSE CITY | 0.0 | 0.2 | 63.3 |
| SAN MATEO | 0.0 | 4.0 | 4.0 |
| SANTA ANA | 0.2 | 1.4 | 7.4 |
| SANTA BARBARA | 0.0 | 1.7 | 4.8 |
| SANTA MONICA | 0.0 | 0.1 | 0.5 |
| SANTA ROSA | 0.0 | 0.0 | 1.6 |
| SANTIAGO CANYON | 0.1 | 0.7 | 3.9 |
| SEQUOIAS | 0.0 | 0.4 | 1.1 |
| SHASTA | 0.1 | 0.0 | 7.4 |
| SIERRA | 0.0 | 70.5 | 2.2 |
| SISKIYOUS | 0.0 | 0.1 | 7.5 |
| SKYLINE | 0.0 | 2.6 | 4.8 |
| SOLANO | 0.0 | 0.2 | 5.8 |
| SOUTHWESTERN | 0.0 | 0.3 | 29.3 |
| TAFT | 0.0 | 0.3 | 29.8 |
| VENTURA | 0.0 | 2.4 | 1.7 |
| VICTOR VALLEY | 0.1 | 0.0 | 68.3 |
| WEST HILLS COALINGA | 0.1 | 1.2 | 56.3 |
| WEST HILLS LEMOORE | 0.0 | 1.0 | 59.4 |
| WEST LA | 0.0 | 0.0 | 0.9 |
| WEST VALLEY | 0.2 | 0.8 | 29.2 |
| WOODLAND | 0.3 | 0.1 | 3.3 |
| YUBA | 0.1 | 0.1 | 1.9 |