# DOCUMENT RESUME


ED 466 279 JC 020 486

| | |
|---|---|
| AUTHOR | Hom, Williard |
| TITLE | Grouping Colleges by Changes in Enrollment Volume. |
| INSTITUTION | California Community Colleges, Sacramento. Office of the Chancellor. |
| PUB DATE | 2002-05-00 |
| NOTE | 21p. |
| AVAILABLE FROM | For full text: http://www.cccco.edu/. |
| PUB TYPE | Information Analyses (070) -- Reports - Research (143) |
| EDRS PRICE | MF01/PC01 Plus Postage. |
| DESCRIPTORS | College Role; *Community Colleges; Comparative Analysis; *Educational Assessment; *Enrollment; Enrollment Influences; Institutional Characteristics; Two Year Colleges |
| IDENTIFIERS | *California Community Colleges |



ABSTRACT

This document discusses the effort to find groupings for the enrollment change in California's community colleges. The new groupings can be utilized by community college strategic planners to improve programs and services based on exploring and analyzing various enrollment shifts. The information can also be used to make enrollment projections for the community colleges. The study analyzes longitudinal enrollment from 1991 to 1999 at 111 California public two-year institutions. Three dimensions were used to characterize the change in student enrollment: (1) slope of the change; (2) variability of the year-to-year change; and (3) consistency of the change across the state. Findings of the study show that enrollment stability is different at each campus. The variability in results reinforces the fact that policymakers cannot treat or consider every community college in the same manner. Some colleges have special needs and may be affected more by certain statewide regulations or standards. (Contains 14 references and several tables.) (MKF)

# Grouping Colleges by Changes in Enrollment Volume

Willard Hom
Director, Research & Planning
Chancellor's Office, California Community Colleges
e-mail: whom@cccco.edu
1102 Q Street
Sacramento, CA 95814-6511

Abstract

In various analyses of community colleges, a need can arise for the grouping of community colleges to help the analyst understand or interpret data for one or more colleges of interest. The question arises, "Is this college typical of the colleges in the state?"

This paper reports an effort to find groupings for the enrollment change in California's community colleges. Such groupings can help researchers and planners by exploring the various types of enrollment shifts that have occurred in the state's colleges since 1991. This information could aid planners who must search for explanations of their enrollment trends and/or who must do enrollment projections.

The analysis in this paper used longitudinal enrollment data in the Chancellor's Office MIS. Various statistical tools allowed us to investigate the (1) slope of the change; (2) the variability of the change; and (3) the association between change at each college with overall change in the state. Cluster analysis provided a method for exploring a potential group structure for the colleges according to these three factors.

## I. Introduction

This study tries to address the following question regarding variation in enrollment patterns over time. "Is this college typical of the colleges in the state?" In answering this question we explore the various types of enrollment shifts that have occurred in the state's community colleges since 1991. This information could aid planners who must search for explanations of their enrollment trends and/or who must do enrollment projections.

In terms of analytical approach, cluster analysis has been recommended as a tool for the empirical discovery of groupings among educational institutions (Brinkman & Teeter, 1987). A recent study by the National Center for Education Statistics used cluster analysis to categorize two-year colleges across the nation (Ronald A. Phipps, Jessica M. Shedd, and Jamie P. Merisotis, 2001). Even modern advocates of data mining techniques recognize the utility of cluster analysis as a tool for discovering natural groupings when analysts lack prior knowledge of group membership among the population of objects under examination. (Han & Kamber, 2001; and Witten & Frank, 2000). Hair & Black (2000) provide an accessible overview and explanation of the cluster analysis method.

## II. Methods

Data for this analysis came from the management information system (MIS) of the Chancellor's Office. Dr.Shuqin Guo compiled the data into one electronic file for this analysis. The years of data span the period of academic year 1991 through academic year 1999. Enrollment data for fall term, credit enrollment at 113 public two-year institutions in California were included. Two institutions were not among the 113 in the analysis because of incomplete enrollment data.

In this investigation, we used the following three dimensions to characterize enrollment change: (1) slope of the change; (2) variability of the year-to-year change; and (3) the consistency of the change across the state.

For our purposes, we defined slope of change as the trend or pattern that describes the pattern of enrollments over the study period. To operationalize this dimension, we attempted to fit a line, by college, to the time series formed by the nine years of enrollment counts for each college. The slope of the resulting trend line served as a simple measure of the overall angle of change for each college. The method of ordinary least squares regression was used to calculate the slope for each college. We assigned the values 1 through 9 serially to the periods 1991 through 1999, respectively, and used the enrollment count as the dependent variable and the serial numbers as the independent (or "predictor") variable in this simple regression equation. We used the standardized beta coefficient as our statistical measure of slope.

Next, we defined the variability of the change as the year-to-year percentage change in enrollment counts. In doing so, each college had a maximum of eight data points. (The first point in the time series had no prior data point with which to calculate a "change" in count.) We finished operationalizing this dimension for each college by computing the standard deviation of the percent year-to-year changes in enrollment counts across the nine years.

Finally, we defined the consistency of the change per college as the association of a college's year-to-year change with the year-to-year change in the statewide total enrollment.. The statewide total for this indicator is the sum of the fall term, credit enrollment counts of all of the colleges in this analysis for each academic year. Figure 1 shows the resulting data for the state totals. Figure 2 gives us a graph of the pattern of the enrollment counts across this study's time horizon of nine years. The chart clearly indicates a "trough" form of curve or pattern for the state enrollment totals.

| Period | Count of Students | Net Change from Prior Year | Net Change as a % of Prior Year |
|---|---|---|---|
| 1 | 1497333 | . | . |
| 2 | 1499570 | 2237 | 0.149 |
| 3 | 1376565 | -123005 | -8.203 |
| 4 | 1355509 | -21056 | -1.530 |
| 5 | 1336406 | -19103 | -1.409 |
| 6 | 1407335 | 70929 | 5.307 |
| 7 | 1442671 | 35336 | 2.511 |
| 8 | 1485851 | 43180 | 2.993 |
| 9 | 1535542 | 49691 | 3.344 |

Figure 1: Enrollment Counts for the State

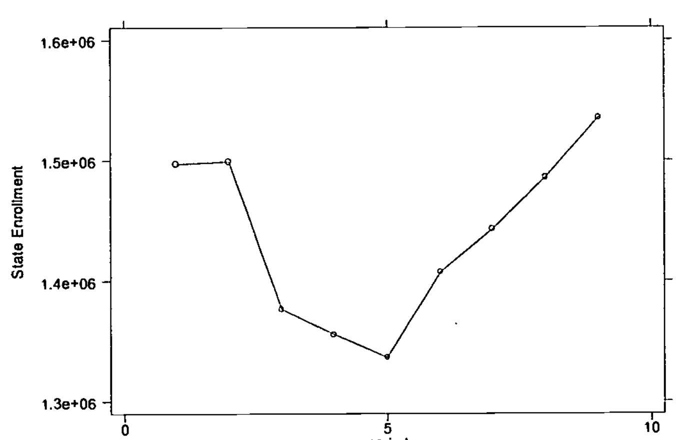


Figure 2: State Enrollment Trend, 1991-1999

We operationalized this third grouping dimension by computing the Spearman rank correlation coefficient of each college's year-to-year change with the state's year-to-year change. Readers who have a familiarity with the financial stock markets will see how a college's correlation to statewide change is analogous to the "beta" coefficient for an individual stock (as it relates to the total "market" pattern). People with a psychometric background may roughly analogize this concept to the use of item-to-total correlation in the development of attitude scales.

This third indicator of change deserves further explanation because it may seem to be a novel measure here. From a policy perspective, we would interpret a large positive correlation for a college as an indication that its change pattern follows that of the state as a whole (and many other colleges for that matter). Theoretically speaking, policies that try to address enrollment issues at the state level will generally apply to colleges with this large positive correlation because such institutions will tend to have similar needs. Of course, this also implies that colleges that have a low correlation or a negative correlation with the state total will tend to experience a different "effect," perhaps an undesired or unintended effect, from a policy designed to address a statewide trend.

In summary, the preceding steps gave us three numeric variables for each college. These variables were (1) the regression slope coefficient; (2) the standard deviation of the year-to-year percent change; and (3) the rank correlation coefficient between each college's year-to-year change in enrollment count and the state-wide year-to-year change in enrollment count. If we assume that these basic variables capture the primary dimensions of enrollment change, then a cluster analysis on these variables should provide us with a way to group colleges according to their similarity in enrollment change over the 1991-99 period. Figures 3, 4, and 5 display the histogram and summary statistics for these three variables.

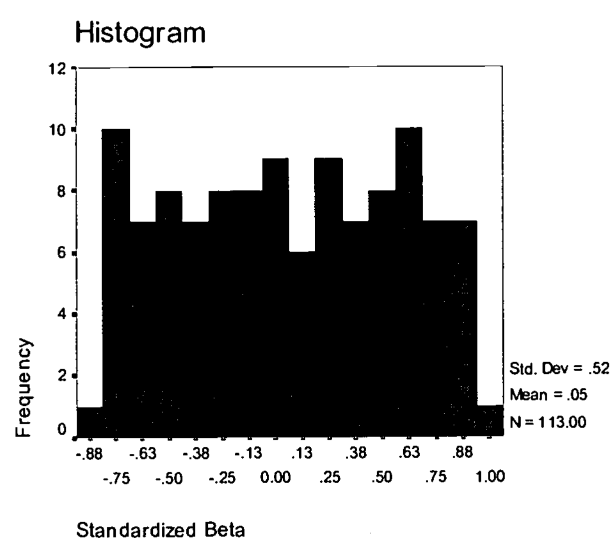


Figure 3: Graph and Summary Statistics for Slope of Change

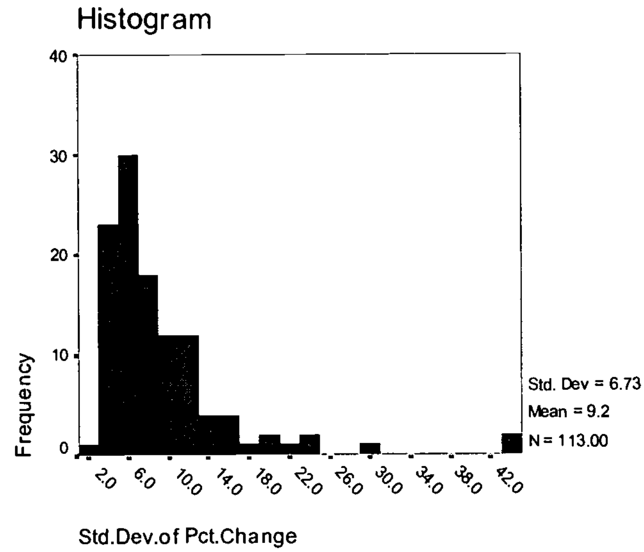


Figure 4: Graph and Summary Statistics for Annual Percent Change

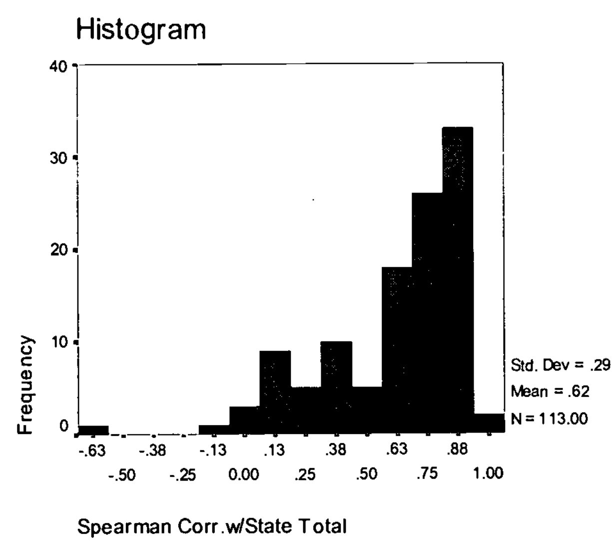


Figure 5: Graph and Summary Statistics for Association with State Total

Before we executed any clustering algorithms, we checked for *multicollinearity* among the three variables. As pointed out by Hair, et al. (1998) and by Everitt & Rabe-Hesketh (1997), multicollinearity among the clustering variables would motivate the use of the *Mahalanobis* distance measure in order to reach an appropriate cluster solution (or structure). Figure 6 displays the bivariate correlation table for the three variables.

Because the correlation table shows no sign of multicollinearity, we concluded that use of the *Mahalanobis* distance measure was unnecessary.

| | | Std.Dev.of Pct.Change | Standardized Beta | Spearman Corr.w/State Total |
|---|---|---|---|---|
| Std.Dev.of Pct.Change | Pearson Correlation | 1 | .090 | -.126 |
| | Sig. (2-tailed) | . | .343 | .185 |
| | N | 113 | 113 | 113 |
| Standardized Beta | Pearson Correlation | .090 | 1 | .098 |
| | Sig. (2-tailed) | .343 | . | .301 |
| | N | 113 | 113 | 113 |
| Spearman Corr.w/State Total | Pearson Correlation | -.126 | .098 | 1 |
| | Sig. (2-tailed) | .185 | .301 | . |
| | N | 113 | 113 | 113 |

Figure 6: Bivariate Correlations for Clustering Variables

We then executed a hierarchical cluster analysis, applying the average linkage algorithm on squared Euclidean distances for standardized values (Z-values) of the three variables. Because cluster analysis can produce very divergent groupings with the use of different algorithms and options, we repeated the cluster analysis with the Ward clustering algorithm.

Some practitioners of cluster analysis advocate yet another refinement of a cluster analysis project. Gore (2000) and Johnson & Wichern (1998) recommend the use of both distance (or "dissimilarity") measures (such as the squared Euclidean metric) as well as a similarity measure (such as the Pearson correlation). Consequently, we executed a third clustering approach that applied the average linkage algorithm to the Pearson similarity measure although there are criticisms of this similarity measure as well (Lorr, 1987; and Dunn & Everitt, 1982). All of the clustering algorithms used in this analysis applied standardization to the cluster variables as a prudent practice for this kind of project (Lorr, 1987; Hair, et al., 1998; and Everitt & Rabe-Hesketh, 1997).

## III. Results

A specialized graph known as a *dendrogram* gives the clearest presentation of the groupings found by a clustering algorithm. Unfortunately, the *dendrogram* is also hard for the layperson to understand, and its size often makes it awkward to present within a document. Cluster analysts can alternatively describe their results by tabulating the mean and standard deviation for each group found by the cluster algorithm. We take this approach below by presenting the means and standard deviations of the cluster variables for each group in Figure 7. This figure uses the output from the Ward algorithm on Euclidean squared distances.

| Group Label | Cases | Mean of Group | SD of Group |
|---|---|---|---|
| 2 | 26 | 0.103 | 0.198 |
| 8 | 15 | 0.633 | 0.117 |
| 9 | 15 | -0.577 | 0.177 |
| 6 | 13 | -0.202 | 0.142 |
| 1 | 10 | 0.412 | 0.237 |
| 4 | 10 | -0.508 | 0.164 |
| 7 | 8 | -0.530 | 0.314 |
| 5 | 7 | 0.833 | 0.101 |
| 11 | 4 | 0.685 | 0.161 |
| 3 | 3 | 0.453 | 0.068 |
| 10 | 1 | -0.570 | NA |
| 12 | 1 | 0.840 | NA |

Note: Uses Standardized Beta (mean = .052; sd = .516)

| Group Label | Cases | Mean of Group | SD of Group |
|---|---|---|---|
| 2 | 26 | 5.7 | 1.4 |
| 8 | 15 | 6.4 | 1.6 |
| 9 | 15 | 8.4 | 2.6 |
| 6 | 13 | 12.1 | 3.2 |
| 1 | 10 | 14.2 | 4.3 |
| 4 | 10 | 4.7 | 1.2 |
| 7 | 8 | 11.0 | 2.1 |
| 5 | 7 | 5.7 | 1.7 |
| 11 | 4 | 23.8 | 4.1 |
| 3 | 3 | 6.1 | 2.3 |
| 10 | 1 | 44.0 | NA |
| 12 | 1 | 44.6 | NA |

Note: Uses Std.Dev. Of Pct.Change (mean = 9.2; sd = 6.7)

| Group Label | Cases | Mean of Group | SD of Group |
|---|---|---|---|
| 2 | 26 | 0.816 | 0.114 |
| 8 | 15 | 0.809 | 0.064 |
| 9 | 15 | 0.458 | 0.126 |
| 6 | 13 | 0.816 | 0.061 |
| 1 | 10 | 0.758 | 0.106 |
| 4 | 10 | 0.672 | 0.115 |
| 7 | 8 | 0.106 | 0.050 |
| 5 | 7 | 0.369 | 0.153 |
| 11 | 4 | 0.154 | 0.080 |
| 3 | 3 | -0.256 | 0.331 |
| 10 | 1 | 0.617 | NA |
| 12 | 1 | 0.738 | NA |

Note: Uses Spearman Correlation (mean = .617; sd = .293)

Figure 7: Tabulation of Means and Standard Deviations by Cluster Group

In Figure 7, "Group Label" refers to the arbitrary name that the cluster program assigns to a cluster group so that the analyst can distinguish group memberships. This number has no other significance or meaning; a high group label like 10 does not denote more of any variable than a lower group label. The column for "Cases" denotes the number of colleges that are in a particular cluster group, as denoted by a group label. We note that clusters containing very few cases tend to identify "outliers" in a study population.

The "Mean of Group" column tells us the central tendency of the set of cases within a group or cluster. By examining this column, we can see what variable at which level distinguishes one group from the other groups. We note that group 10 and group 12 contain only one college in each of them. The high value for standard deviation of percent change distinguishes these two cases as so unique as to deserve their own single-case clusters.

Groups 2, 8, and 9 are the three largest clusters in the results in terms of cases. With Figure 7, we would interpret Group 2 to be those colleges with almost no change (the standardized beta coefficient is near zero). Group 8 resembles Group 2 in terms annual percent change and in correlation with the state total, but Group 8 has a much more positive growth pattern (mean beta slope of .633). Group 9, although containing 15 colleges like Group 8, differs on average markedly from Group 8 on all three variables. Colleges in Group 9, compared to those in Group 8, would tend to have a large decline in enrollment, greater annual percentage change, and a less "cyclical" pattern (low association with the state pattern). We could proceed with this analysis to quite some depth, but time and space compel us to reserve that work for another time.

The three tables in Figure 7 partially demonstrate the effectiveness of the cluster result. The standard deviation for each of the three cluster variables is smaller than the overall standard deviation of the ungrouped population (the ungrouped statistics appear in the note below each table). In cluster analysis, our goal is the formation of homogeneous groupings, and the small within-group standard deviations indicate our success on this objective. Space does not permit us to reiterate the above tabulation for the other two cluster methods we tested, but the results generally resemble those in Figure 7.

## IV. Discussion

As indicated by Hair, et al. (1998), the different clustering algorithms tend to accentuate a particular cluster outcome. "*Average linkage approaches tend to be biased toward the production of clusters with approximately the same variance.... Ward's method...tends to combine clusters with a small number of observations. It is also biased toward the production of clusters with approximately the same number of observations...*"

We will need to do much more work in order to reach an interpretation of these cluster structures before the groupings can help in the analysis of enrollment planning. To use these results, we would need to distinguish a "true" structure in enrollment patterns from the "method bias" that often results from applying different statistical approaches to a single set of data. Ideally, further analysis will integrate this important interpretation of the cluster results with steps to check the validity of the clustering results. In terms of incremental modifications or enhancements to the development of a structure already done here, we consider in the next paragraphs some other steps that may warrant future effort.

The cluster analysis performed here could be expanded to include other indicators of enrollment change, and a future study could explore these alternatives. For example, some basic indicators to test could be the number of runs within a time series; the time interval in which either a peak or a trough occurred in the time series for each college (very useful with the curve evidenced during 1991-1999); the leverage and influence of the most recent year upon the fitted regression line (perhaps using Cook's D); and the level of fit to the straight line (perhaps using the R-Square measure). Naturally, the more data points that we can analyze in the time series, the more indicators of pattern we may have to explore in a meaningful way. In addition, an analyst could test the use of the *Mahalanobis* distance measure as an alternative to the Euclidean squared distance and to the Pearson similarity measure.

Another alternative to test would be the use of the Pearson correlation, in lieu of the Spearman rank correlation, to measure the association of each college's enrollment pattern to the overall state pattern. We chose the Spearman correlation because it is a more robust measure of association than the Pearson correlation. However, we may have

traded off some sensitivity to patterns of association in order to obtain that limitation of effect from extreme values.

Even if the aforementioned modifications do not get tested, analysts should still consider replicating this study's basic analysis some time in the future. In time series work, additional observations often enable analysts to try other statistical tools, and these tools may ferret out patterns that we cannot easily observe. Furthermore, the passage of years will tend to make this clustering study somewhat obsolete, given that new patterns of enrollment change can easily develop.

## V. Conclusion

At a minimum, this study has explored some basic measures of enrollment variation. The three dimensions used in our cluster analysis have value not only at the multivariate level (that is, via the cluster analysis) but also at the univariate level. We can see how the colleges vary according to slope of change; annual percentage change; and consistency with the state total. Each of these measures may enhance planning for the colleges as "stand-alone" indicators of enrollment stability and direction.

In order to apply the multivariate quality of these four measures to policy, we should do further work on the cluster results. An extension of the work presented here should probably undertake additional evaluation of cluster validity, and methods to do this are available (Jain & Dubes, 1988; Anderberg, 1973; Whitten & Frank, 2000; and Johnson & Wichern, 1998).

If additional analysis validates a particular cluster structure in this study, then analysts and planners will have more useful information here. In terms of planning enrollment projections, the groupings represented by a valid cluster structure roughly indicate the variety, or breadth, of enrollment patterns that a projection system would need to accommodate. The clustering also indicates which colleges may be most suitable for a particular type of projection model.

Aside from the aid to planning projection methodology, the resulting groupings may inform two policy issues facing community colleges in California. As noted by Sneath & Sokal (1973), "numerical taxonomy" provides heuristic information in that analysts can advance, or begin to formulate, some theories for further development. In our case, we want to advance our knowledge of factors behind the enrollment trends of different colleges. By identifying basic categories of enrollment variation, we can begin to see what common threads (or causal factors) exist that, at least in part, determine a particular enrollment pattern. Understanding the causal factors behind enrollment patterns would help colleges to develop ways to manage their enrollments as well as to forecast them.

At an administrative level, groupings also help us understand that some colleges have inherently different qualities about them (with enrollment stability being one major quality) that should factor into how we treat or consider them when policy-making occurs. Real groupings reinforce the argument that the policy makers really cannot treat or consider every college in the same way. State wide regulations will not affect every college equally, and many colleges will have special needs.

## VI. References


Anderberg, M.R. (1973). *Cluster Analysis for Applications*. New York: Academic Press.

Brinkman, P.T. & Teeter, D.J. (1987). Methods for Selecting Comparison Groups. *New Directions for Institutional Research*, no.55, 5-23.

Dunn, G. & Everitt, B.S. (1982). *An Introduction to Mathematical Taxonomy*. New York: Cambridge University Press.

Everitt, B.S. & Rabe-Hesketh, S. (1997). *The Analysis of Proximity Data*. London: Arnold.

Gore, P.A. (2000). Cluster Analysis. In Howard D.A.Tinsley & Steven D. Brown. (Eds.) *Handbook of Multivariate Statistics and Mathematical Modeling*. (pp.297-321). San Diego: Academic Press.

Hair, J.F., et al. (1998). *Multivariate Data Analysis, Fifth Edition*. Upper Saddle River, New Jersey: Prentice-Hall.

Hair, J.F. & Black, W.C. (2000). Cluster Analysis. In Laurence G.Grimm & Paul R. Yarnold (Eds.). *Reading and Understanding More Multivariate Statistics*. Washington, D.C.: American Psychological Association.

Han,J. & Kamber, M. (2001). *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann.

Jain, A.K. & Dubes, R.C. (1988). *Algorithms for Clustering Data*. Englewood Cliffs, New Jersey: Prentice-Hall.

Johnson, R.A. & Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis*. Upper Saddle River, New Jersey: Prentice-Hall.

Lorr, M. (1987). *Cluster Analysis for the Social Sciences*. San Francisco: Jossey-Bass.

Phipps, R.A., Shedd, J.M., & Merisotis, J.P. (2001). *A Classification System for 2-Year Postsecondary Institutions*. Washington, D.C.: National Center for Education Statistics.

Sneath, P.H.A. & Sokal, R.R. (1973). *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. San Francsco: W.H.Freeman.

Witten, I.A. & Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations*. San Francisco: Morgan Kaufmann.

# Grouping Colleges by Changes in Enrollment Volume

Willard Hom
Director, Research & Planning
Chancellor's Office, California Community Colleges
e-mail: whom@cccco.edu
1102 Q Street
Sacramento, CA 95814-6511

Abstract

In various analyses of community colleges, a need can arise for the grouping of community colleges to help the analyst understand or interpret data for one or more colleges of interest. The question arises, "Is this college typical of the colleges in the state?"

This paper reports an effort to find groupings for the enrollment change in California's community colleges. Such groupings can help researchers and planners by exploring the various types of enrollment shifts that have occurred in the state's colleges since 1991. This information could aid planners who must search for explanations of their enrollment trends and/or who must do enrollment projections.

The analysis in this paper used longitudinal enrollment data in the Chancellor's Office MIS. Various statistical tools allowed us to investigate the (1) slope of the change; (2) the variability of the change; and (3) the association between change at each college with overall change in the state. Cluster analysis provided a method for exploring a potential group structure for the colleges according to these three factors.

## I. Introduction

This study tries to address the following question regarding variation in enrollment patterns over time. "Is this college typical of the colleges in the state?" In answering this question we explore the various types of enrollment shifts that have occurred in the state's community colleges since 1991. This information could aid planners who must search for explanations of their enrollment trends and/or who must do enrollment projections.

In terms of analytical approach, cluster analysis has been recommended as a tool for the empirical discovery of groupings among educational institutions (Brinkman & Teeter, 1987). A recent study by the National Center for Education Statistics used cluster analysis to categorize two-year colleges across the nation (Ronald A. Phipps, Jessica M. Shedd, and Jamie P. Merisotis, 2001). Even modern advocates of data mining techniques recognize the utility of cluster analysis as a tool for discovering natural groupings when analysts lack prior knowledge of group membership among the population of objects under examination. (Han & Kamber, 2001; and Witten & Frank, 2000). Hair & Black (2000) provide an accessible overview and explanation of the cluster analysis method.

## II. Methods

Data for this analysis came from the management information system (MIS) of the Chancellor's Office. Dr.Shuqin Guo compiled the data into one electronic file for this analysis. The years of data span the period of academic year 1991 through academic year 1999. Enrollment data for fall term, credit enrollment at 113 public two-year institutions in California were included. Two institutions were not among the 113 in the analysis because of incomplete enrollment data.

In this investigation, we used the following three dimensions to characterize enrollment change: (1) slope of the change; (2) variability of the year-to-year change; and (3) the consistency of the change across the state.

For our purposes, we defined slope of change as the trend or pattern that describes the pattern of enrollments over the study period. To operationalize this dimension, we attempted to fit a line, by college, to the time series formed by the nine years of enrollment counts for each college. The slope of the resulting trend line served as a simple measure of the overall angle of change for each college. The method of ordinary least squares regression was used to calculate the slope for each college. We assigned the values 1 through 9 serially to the periods 1991 through 1999, respectively, and used the enrollment count as the dependent variable and the serial numbers as the independent (or "predictor") variable in this simple regression equation. We used the standardized beta coefficient as our statistical measure of slope.

Next, we defined the variability of the change as the year-to-year percentage change in enrollment counts. In doing so, each college had a maximum of eight data points. (The first point in the time series had no prior data point with which to calculate a "change" in count.) We finished operationalizing this dimension for each college by computing the standard deviation of the percent year-to-year changes in enrollment counts across the nine years.

Finally, we defined the consistency of the change per college as the association of a college's year-to-year change with the year-to-year change in the statewide total enrollment. The statewide total for this indicator is the sum of the fall term, credit enrollment counts of all of the colleges in this analysis for each academic year. Figure 1 shows the resulting data for the state totals. Figure 2 gives us a graph of the pattern of the enrollment counts across this study's time horizon of nine years. The chart clearly indicates a "trough" form of curve or pattern for the state enrollment totals.

| Period | Count of Students | Net Change from Prior Year | Net Change as a % of Prior Year |
|---|---|---|---|
| 1 | 1497333 | . | . |
| 2 | 1499570 | 2237 | 0.149 |
| 3 | 1376565 | -123005 | -8.203 |
| 4 | 1355509 | -21056 | -1.530 |
| 5 | 1336406 | -19103 | -1.409 |
| 6 | 1407335 | 70929 | 5.307 |
| 7 | 1442671 | 35336 | 2.511 |
| 8 | 1485851 | 43180 | 2.993 |
| 9 | 1535542 | 49691 | 3.344 |

Figure 1: Enrollment Counts for the State

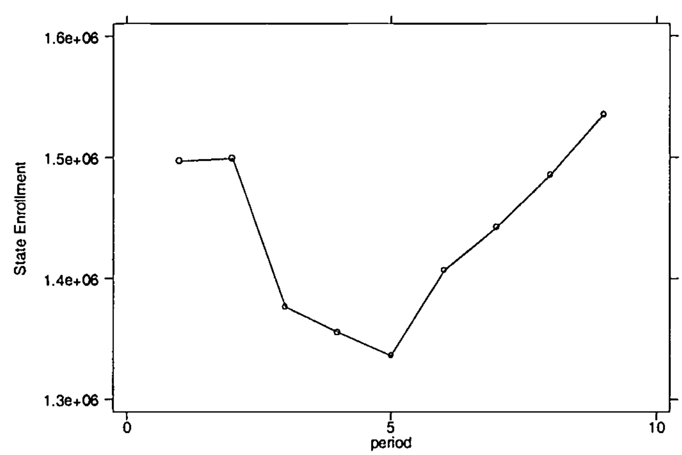


Figure 2: State Enrollment Trend, 1991-1999

We operationalized this third grouping dimension by computing the Spearman rank correlation coefficient of each college's year-to-year change with the state's year-to-year change. Readers who have a familiarity with the financial stock markets will see how a college's correlation to statewide change is analogous to the "beta" coefficient for an individual stock (as it relates to the total "market" pattern). People with a psychometric background may roughly analogize this concept to the use of item-to-total correlation in the development of attitude scales.

This third indicator of change deserves further explanation because it may seem to be a novel measure here. From a policy perspective, we would interpret a large positive correlation for a college as an indication that its change pattern follows that of the state as a whole (and many other colleges for that matter). Theoretically speaking, policies that try to address enrollment issues at the state level will generally apply to colleges with this large positive correlation because such institutions will tend to have similar needs. Of course, this also implies that colleges that have a low correlation or a negative correlation with the state total will tend to experience a different "effect," perhaps an undesired or unintended effect, from a policy designed to address a statewide trend.

In summary, the preceding steps gave us three numeric variables for each college. These variables were (1) the regression slope coefficient; (2) the standard deviation of the year-to-year percent change; and (3) the rank correlation coefficient between each college's year-to-year change in enrollment count and the state-wide year-to-year change in enrollment count. If we assume that these basic variables capture the primary dimensions of enrollment change, then a cluster analysis on these variables should provide us with a way to group colleges according to their similarity in enrollment change over the 1991-99 period. Figures 3, 4, and 5 display the histogram and summary statistics for these three variables.

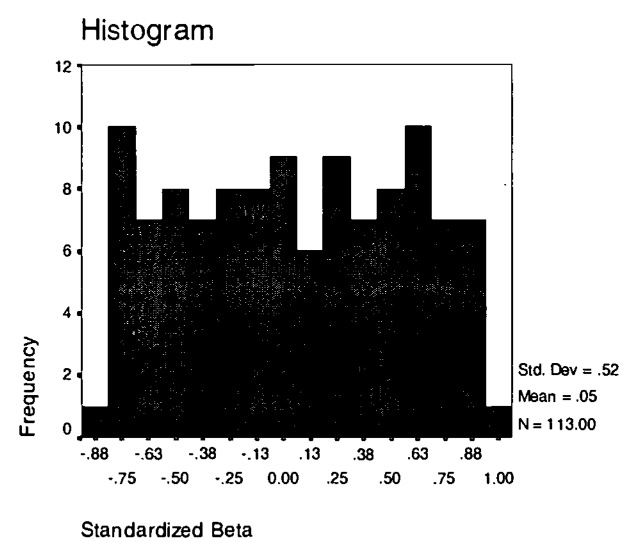


Figure 3: Graph and Summary Statistics for Slope of Change

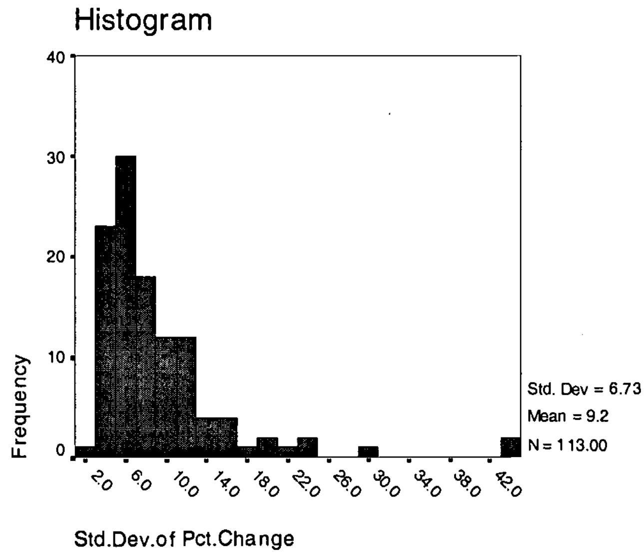


Figure 4: Graph and Summary Statistics for Annual Percent Change

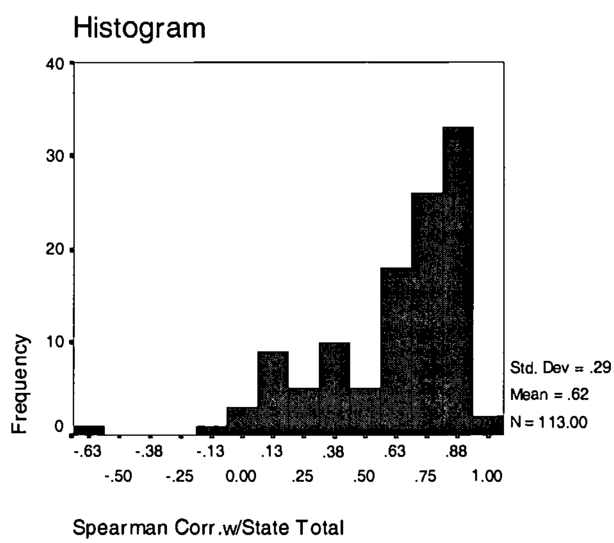


Figure 5: Graph and Summary Statistics for Association with State Total

Before we executed any clustering algorithms, we checked for *multicollinearity* among the three variables. As pointed out by Hair, et al. (1998) and by Everitt & Rabe-Hesketh (1997), multicollinearity among the clustering variables would motivate the use of the *Mahalanobis* distance measure in order to reach an appropriate cluster solution (or structure). Figure 6 displays the bivariate correlation table for the three variables.

Because the correlation table shows no sign of multicollinearity, we concluded that use of the *Mahalanobis* distance measure was unnecessary.

| | | Std.Dev.of Pct.Change | Standardized Beta | Spearman Corr.w/State Total |
|---|---|---|---|---|
| Std.Dev.of Pct.Change | Pearson Correlation | 1 | .090 | -.126 |
| | Sig. (2-tailed) | . | .343 | .185 |
| | N | 113 | 113 | 113 |
| Standardized Beta | Pearson Correlation | .090 | 1 | .098 |
| | Sig. (2-tailed) | .343 | . | .301 |
| | N | 113 | 113 | 113 |
| Spearman Corr.w/State Total | Pearson Correlation | -.126 | .098 | 1 |
| | Sig. (2-tailed) | .185 | .301 | . |
| | N | 113 | 113 | 113 |

Figure 6: Bivariate Correlations for Clustering Variables

We then executed a hierarchical cluster analysis, applying the average linkage algorithm on squared Euclidean distances for standardized values (Z-values) of the three variables. Because cluster analysis can produce very divergent groupings with the use of different algorithms and options, we repeated the cluster analysis with the Ward clustering algorithm.

Some practitioners of cluster analysis advocate yet another refinement of a cluster analysis project. Gore (2000) and Johnson & Wichern (1998) recommend the use of both distance (or "dissimilarity") measures (such as the squared Euclidean metric) as well as a similarity measure (such as the Pearson correlation). Consequently, we executed a third clustering approach that applied the average linkage algorithm to the Pearson similarity measure although there are criticisms of this similarity measure as well (Lorr, 1987; and Dunn & Everitt, 1982). All of the clustering algorithms used in this analysis applied standardization to the cluster variables as a prudent practice for this kind of project (Lorr, 1987; Hair, et al., 1998; and Everitt & Rabe-Hesketh, 1997).

## III. Results

A specialized graph known as a *dendrogram* gives the clearest presentation of the groupings found by a clustering algorithm. Unfortunately, the *dendrogram* is also hard for the layperson to understand, and its size often makes it awkward to present within a document. Cluster analysts can alternatively describe their results by tabulating the mean and standard deviation for each group found by the cluster algorithm. We take this approach below by presenting the means and standard deviations of the cluster variables for each group in Figure 7. This figure uses the output from the Ward algorithm on Euclidean squared distances.

| Group Label | Cases | Mean of Group | SD of Group |
|---|---|---|---|
| 2 | 26 | 0.103 | 0.198 |
| 8 | 15 | 0.633 | 0.117 |
| 9 | 15 | -0.577 | 0.177 |
| 6 | 13 | -0.202 | 0.142 |
| 1 | 10 | 0.412 | 0.237 |
| 4 | 10 | -0.508 | 0.164 |
| 7 | 8 | -0.530 | 0.314 |
| 5 | 7 | 0.833 | 0.101 |
| 11 | 4 | 0.685 | 0.161 |
| 3 | 3 | 0.453 | 0.068 |
| 10 | 1 | -0.570 | NA |
| 12 | 1 | 0.840 | NA |

Note: Uses Standardized Beta (mean = .052; sd = .516)

| Group Label | Cases | Mean of Group | SD of Group |
|---|---|---|---|
| 2 | 26 | 5.7 | 1.4 |
| 8 | 15 | 6.4 | 1.6 |
| 9 | 15 | 8.4 | 2.6 |
| 6 | 13 | 12.1 | 3.2 |
| 1 | 10 | 14.2 | 4.3 |
| 4 | 10 | 4.7 | 1.2 |
| 7 | 8 | 11.0 | 2.1 |
| 5 | 7 | 5.7 | 1.7 |
| 11 | 4 | 23.8 | 4.1 |
| 3 | 3 | 6.1 | 2.3 |
| 10 | 1 | 44.0 | NA |
| 12 | 1 | 44.6 | NA |

Note: Uses Std.Dev. Of Pct.Change (mean = 9.2; sd = 6.7)

| Group Label | Cases | Mean of Group | SD of Group |
|---|---|---|---|
| 2 | 26 | 0.816 | 0.114 |
| 8 | 15 | 0.809 | 0.064 |
| 9 | 15 | 0.458 | 0.126 |
| 6 | 13 | 0.816 | 0.061 |
| 1 | 10 | 0.758 | 0.106 |
| 4 | 10 | 0.672 | 0.115 |
| 7 | 8 | 0.106 | 0.050 |
| 5 | 7 | 0.369 | 0.153 |
| 11 | 4 | 0.154 | 0.080 |
| 3 | 3 | -0.256 | 0.331 |
| 10 | 1 | 0.617 | NA |
| 12 | 1 | 0.738 | NA |

Note: Uses Spearman Correlation (mean = .617; sd = .293)

Figure 7: Tabulation of Means and Standard Deviations by Cluster Group

In Figure 7, "Group Label" refers to the arbitrary name that the cluster program assigns to a cluster group so that the analyst can distinguish group memberships. This number has no other significance or meaning; a high group label like 10 does not denote more of any variable than a lower group label. The column for "Cases" denotes the number of colleges that are in a particular cluster group, as denoted by a group label. We note that clusters containing very few cases tend to identify "outliers" in a study population.

The "Mean of Group" column tells us the central tendency of the set of cases within a group or cluster. By examining this column, we can see what variable at which level distinguishes one group from the other groups. We note that group 10 and group 12 contain only one college in each of them. The high value for standard deviation of percent change distinguishes these two cases as so unique as to deserve their own single-case clusters.

Groups 2, 8, and 9 are the three largest clusters in the results in terms of cases. With Figure 7, we would interpret Group 2 to be those colleges with almost no change (the standardized beta coefficient is near zero). Group 8 resembles Group 2 in terms annual percent change and in correlation with the state total, but Group 8 has a much more positive growth pattern (mean beta slope of .633). Group 9, although containing 15 colleges like Group 8, differs on average markedly from Group 8 on all three variables. Colleges in Group 9, compared to those in Group 8, would tend to have a large decline in enrollment, greater annual percentage change, and a less "cyclical" pattern (low association with the state pattern). We could proceed with this analysis to quite some depth, but time and space compel us to reserve that work for another time.

The three tables in Figure 7 partially demonstrate the effectiveness of the cluster result. The standard deviation for each of the three cluster variables is smaller than the overall standard deviation of the ungrouped population (the ungrouped statistics appear in the note below each table). In cluster analysis, our goal is the formation of homogeneous groupings, and the small within-group standard deviations indicate our success on this objective. Space does not permit us to reiterate the above tabulation for the other two cluster methods we tested, but the results generally resemble those in Figure 7.

## IV. Discussion

As indicated by Hair, et al. (1998), the different clustering algorithms tend to accentuate a particular cluster outcome. "*Average linkage approaches tend to be biased toward the production of clusters with approximately the same variance.... Ward's method...tends to combine clusters with a small number of observations. It is also biased toward the production of clusters with approximately the same number of observations...*"

We will need to do much more work in order to reach an interpretation of these cluster structures before the groupings can help in the analysis of enrollment planning. To use these results, we would need to distinguish a "true" structure in enrollment patterns from the "method bias" that often results from applying different statistical approaches to a single set of data. Ideally, further analysis will integrate this important interpretation of the cluster results with steps to check the validity of the clustering results. In terms of incremental modifications or enhancements to the development of a structure already done here, we consider in the next paragraphs some other steps that may warrant future effort.

The cluster analysis performed here could be expanded to include other indicators of enrollment change, and a future study could explore these alternatives. For example, some basic indicators to test could be the number of runs within a time series; the time interval in which either a peak or a trough occurred in the time series for each college (very useful with the curve evidenced during 1991-1999); the leverage and influence of the most recent year upon the fitted regression line (perhaps using Cook's D); and the level of fit to the straight line (perhaps using the R-Square measure). Naturally, the more data points that we can analyze in the time series, the more indicators of pattern we may have to explore in a meaningful way. In addition, an analyst could test the use of the *Mahalanobis* distance measure as an alternative to the Euclidean squared distance and to the Pearson similarity measure.

Another alternative to test would be the use of the Pearson correlation, in lieu of the Spearman rank correlation, to measure the association of each college's enrollment pattern to the overall state pattern. We chose the Spearman correlation because it is a more robust measure of association than the Pearson correlation. However, we may have

traded off some sensitivity to patterns of association in order to obtain that limitation of effect from extreme values.

Even if the aforementioned modifications do not get tested, analysts should still consider replicating this study's basic analysis some time in the future. In time series work, additional observations often enable analysts to try other statistical tools, and these tools may ferret out patterns that we cannot easily observe. Furthermore, the passage of years will tend to make this clustering study somewhat obsolete, given that new patterns of enrollment change can easily develop.

## V. Conclusion

At a minimum, this study has explored some basic measures of enrollment variation. The three dimensions used in our cluster analysis have value not only at the multivariate level (that is, via the cluster analysis) but also at the univariate level. We can see how the colleges vary according to slope of change; annual percentage change; and consistency with the state total. Each of these measures may enhance planning for the colleges as "stand-alone" indicators of enrollment stability and direction.

In order to apply the multivariate quality of these four measures to policy, we should do further work on the cluster results. An extension of the work presented here should probably undertake additional evaluation of cluster validity, and methods to do this are available (Jain & Dubes, 1988; Anderberg, 1973; Whitten & Frank, 2000; and Johnson & Wichern, 1998).

If additional analysis validates a particular cluster structure in this study, then analysts and planners will have more useful information here. In terms of planning enrollment projections, the groupings represented by a valid cluster structure roughly indicate the variety, or breadth, of enrollment patterns that a projection system would need to accommodate. The clustering also indicates which colleges may be most suitable for a particular type of projection model.

Aside from the aid to planning projection methodology, the resulting groupings may inform two policy issues facing community colleges in California. As noted by Sneath & Sokal (1973), "numerical taxonomy" provides heuristic information in that analysts can advance, or begin to formulate, some theories for further development. In our case, we want to advance our knowledge of factors behind the enrollment trends of different colleges. By identifying basic categories of enrollment variation, we can begin to see what common threads (or causal factors) exist that, at least in part, determine a particular enrollment pattern. Understanding the causal factors behind enrollment patterns would help colleges to develop ways to manage their enrollments as well as to forecast them.

At an administrative level, groupings also help us understand that some colleges have inherently different qualities about them (with enrollment stability being one major quality) that should factor into how we treat or consider them when policy-making occurs. Real groupings reinforce the argument that the policy makers really cannot treat or consider every college in the same way. State wide regulations will not affect every college equally, and many colleges will have special needs.

## VI. References


Anderberg, M.R. (1973). *Cluster Analysis for Applications*. New York: Academic Press.

Brinkman, P.T. & Teeter, D.J. (1987). Methods for Selecting Comparison Groups. *New Directions for Institutional Research*, no.55, 5-23.

Dunn, G. & Everitt, B.S. (1982). *An Introduction to Mathematical Taxonomy*. New York: Cambridge University Press.

Everitt, B.S. & Rabe-Hesketh, S. (1997). *The Analysis of Proximity Data*. London: Arnold.

Gore, P.A. (2000). Cluster Analysis. In Howard D.A.Tinsley & Steven D. Brown. (Eds.) *Handbook of Multivariate Statistics and Mathematical Modeling*. (pp.297-321). San Diego: Academic Press.

Hair, J.F., et al. (1998). *Multivariate Data Analysis, Fifth Edition*. Upper Saddle River, New Jersey: Prentice-Hall.

Hair, J.F. & Black, W.C. (2000). Cluster Analysis. In Laurence G.Grimm & Paul R. Yarnold (Eds.). *Reading and Understanding More Multivariate Statistics*. Washington, D.C.: American Psychological Association.

Han,J. & Kamber, M. (2001). *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann.

Jain, A.K. & Dubes, R.C. (1988). *Algorithms for Clustering Data*. Englewood Cliffs, New Jersey: Prentice-Hall.

Johnson, R.A. & Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis*. Upper Saddle River, New Jersey: Prentice-Hall.

Lorr, M. (1987). *Cluster Analysis for the Social Sciences*. San Francisco: Jossey-Bass.

Phipps, R.A., Shedd, J.M., & Merisotis, J.P. (2001). *A Classification System for 2-Year Postsecondary Institutions.* Washington, D.C.: National Center for Education Statistics.

Sneath, P.H.A. & Sokal, R.R. (1973). *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. San Francsco: W.H.Freeman.

Witten, I.A. & Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations*. San Francisco: Morgan Kaufmann.

**U.S. Department of Education**
*Office of Educational Research and Improvement (OERI)*
*National Library of Education (NLE)*
*Educational Resources Information Center (ERIC)*

ERIC®

# REPRODUCTION RELEASE

(Specific Document)

## I. DOCUMENT IDENTIFICATION:

Title: Grouping Colleges by Changes in Enrollment Volume

Author(s): Willard C. Hom

| Corporate Source: | Publication Date: May 2002 |
|---|---|

## II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

| The sample sticker shown below will be affixed to all Level 1 documents | The sample sticker shown below will be affixed to all Level 2A documents | The sample sticker shown below will be affixed to all Level 2B documents |
|---|---|---|
| PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1 | PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A | PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B |
| Level 1 ☑ | Level 2A ☐ | Level 2B ☐ |
| Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy. | Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only | Check here for Level 2B release, permitting reproduction and dissemination in microfiche only |

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

*I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.*

**Sign here,→ please**

| Signature: Willard C. Hom | | Printed Name/Position/Title: Willard C. Hom, Administrator | |
|---|---|---|---|
| Organization/Address: Chancellor's Office, Calif. Community Colleges. 1102 Q St. Sacramento 95814-6511 | | Telephone: 916-327-5887 | FAX: |
| | | E-Mail Address: whom@cccco.edu | Date: June 3, 2002 |



(over)

## III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

## IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

## V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

**ERIC Clearinghouse for Community Colleges**
**University of California, Los Angeles**
**3051 Moore Hall, Box 951521**
**Los Angeles, CA 90095-1521**
**Telephone: (800) 832-8256**
**Fax: (310) 206-8095**

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

**ERIC Processing and Reference Facility**
**4483-A [illegible]bes B[illegible]levard**
**Lanham, [illegible]ryl[illegible]d 20706**

**Telephone: [illegible]-552-4200**
**Toll Free: [illegible]99-3742**
**FAX: [illegible]01-55[illegible]700**
**e-mail: [illegible]icfac@in[illegible]ed.gov**
**WWW: http://ericfac.piccard.csc.com**