IR Applications
Volume 18, February 8, 2009
Using Advanced Tools, Techniques, and Methodologies

Using a Data Mining Approach to Develop a Student Engagement-Based Institutional Typology

Jing Luan
Vice Chancellor, Educational Services and Planning
San Mateo County Community College District
luan@smccd.edu

Chun-Mei Zhao
Senior Scholar
The Carnegie Foundation for the Advancement of Teaching
zhao@carnegiefoundation.org

John C. Hayek
Associate Vice President
Kentucky Council on Postsecondary Education
john.hayek@ky.gov

Association for Institutional Research

Enhancing knowledge.
Expanding networks.

Abstract

Data mining provides both systematic and systemic ways to detect patterns of student engagement among students at hundreds of institutions. Using traditional statistical techniques alone, the task would be significantly difficult—if not impossible—considering the size and complexity in both data and analytical approaches necessary for this task. This study presents a step-by-step review on how the data mining technique is utilized to develop an institutional typology based on student behavioral data. The result provides a fresh angle to understand similarities and differences among four-year undergraduate colleges and universities, shifting away from previous institutional typologies, such as those based on institutional mission, resources, or reputation. The institutional engagement typology is derived through student behavioral data, and therefore, is advantageous in that it retains one of the most important components in understanding higher education—student behaviors. This data mining-based study broke new conceptual and methodological ground, and its resulting institutional learning engagement typology offers new perspectives on peer institution comparison, congruence between students and their institutions, as well as policy development regarding educational quality.

Acknowledgements

Many people provided invaluable assistance in carrying out this large scale study and contributed to its successful completion. The authors would like to thank Dr. George Kuh for making the national NSSE dataset available for the study; thanks should go to SPSS for its longstanding support of higher education and of the authors' research work over the years; a special thank-you goes to Dr. Richard (Dick) Borden at Palomar College for his knowledge in WinCross, factor analysis, and his humor and wisdom. Lastly, a heart-felt thank-you goes to Dr. Gerald (Gerry) McLaughlin for his patience with the authors and his magnificent editorial advice and to the AIR leadership and its staff for their support of data mining.

Background

The general public increasingly demands more accountability in

American higher education. Student engagement is perceived to be an integral part of the accountability by way of understanding how institutions engage students in educationally effective activities, therefore, fundamentally influencing the student learning outcomes. Understandably, the landscape of studying student learning outcomes is crowded with the unrelenting accumulation of data. What remains as a challenge to the research community is to find new tools that can both efficiently tame large datasets and uncover fresh conceptual and methodological models useful for understanding student learning engagement.

A relatively new tool in higher education research, data mining, provides a powerful way to detect patterns in data that would be significantly more difficult, if not impossible, to see using traditional statistical techniques alone (Dyche, 2000). Data mining is a collection of statistical and data management techniques previously unattainable due to limitations in computing power, data storage capacity, and statistical sophistication. However, as these technical barriers have fallen, data mining has become a mission-critical part of business research and a productivity tool in many industries such as healthcare, banking, and the retail sector. In higher education, data mining is mostly considered enigmatic due to a current lack of use and understanding.

Data mining is often referred to as "planned serendipity," that is, it searches for patterns or relations not confined by pre-established notions or hypotheses (Hair, Anderson, Tatham, & Black, 1998). Applying it in the higher education arena, data mining is an analytic approach that "capitalizes on the advances of technology and the extreme richness of data in higher education for improving research and decision making through uncovering hidden trends and patterns that lend them to predicative modeling using a combination of explicit knowledge base, sophisticated analytical skills and academic domain knowledge" (Luan, 2002, p. 3). Data mining does not intend to replace traditional statistics. Rather, data mining is an extension of statistics, and statistics is an integral component in data mining (Luan, 2003; Zhao & Luan, 2006).

Several key notions need to be stated for readers to have a contextual understanding of why data mining is chosen for this study. Data mining and traditional statistics have different intellectual traditions, and several intrinsic differences exist between them. First, traditional statistical approaches favor probabilistic models and tend to use sampled and experimental data. Alternatively, data mining has much newer origins, primarily due to the rapid expansion of computer capacity and the advancement of technology, such as the development of artificial intelligence, machine learning, management information systems and database storage and query methodology. Data mining usually works with large observational (often unstructured) datasets (Hand, Mannila, & Smyth, 2001).

Second, data mining is exploratory in nature, that is, a search for useful pattern in the data that is not restricted by pre-established notions of what patterns are expected. In this regard, data mining owes its heritage to John Tukey's exploratory data analysis (Mosteller & Tukey, 1977; Tukey, 1977). He emphasized the value of information already embedded in large amount of data. In contrast, traditional statistics try to understand relationships under a certain theoretical framework. The data mining process is not linear, rather, it is iterative. It is a process that uses a variety of data analysis tools to discover patterns and relations.

Further, traditional statistics emphasizes the confirmatory aspect that aims at identifying the "general cause" of a phenomenon or behavior to corroborate a theory that can be generalized into a wider population from a sample. Data mining, in contrast, has a strong pragmatic focus. Data mining, particularly predictive modeling, does not theorize behaviors. Rather, it just presents the patterns or relationships to inform, influence, and strategize practical applications. Within this perspective, the analyses that follow are presented as descriptive results, and the various statistical metrics in general are reported. Also in several cases, the results included are only a small part of those obtained. Additional statistics and results can be obtained by contacting the first author.

Purpose

The purpose of this paper is to provide a step-by-step look at the use of some data mining techniques to explore and identify a new institutional typology. This typology will be based on the pattern of student types. The student typology is derived from the patterns by which individual undergraduate students engage in educationally purposeful activities. An institutional typology is important in understanding the similarities and differences among colleges and universities. Especially in the case of the U.S. higher education system with nearly 4,400 institutions that enroll around 17.5 million students (close to 6% of the overall population), the vast diversity is one of its signature characteristics. There is a need to develop different frameworks to compare colleges and universities. The most widely used classification is the Carnegie Classification for Institutions of Higher Education (McCormick, 2001; the Carnegie Foundation for the Advancement of Teaching, 2008), which historically focuses on institutional-level characteristics such as level of federal support,[1] degree offerings, instructional program focus, etc.

What students do during their college career is a critical and fundamental aspect of undergraduate education (Astin, 1973, 1993; Pace, 1984; Pascarella & Terenzini, 1991). So far, no previous framework exists to understand similarities and differences among institutions of higher education with respect to student engagement patterns; none of the established institution classification frameworks focus on what students actually do. This study seeks to fill the void and to pilot a study to fully take advantage of the richness of student behavioral data. The data employed in the study is from the National Survey for Student Engagement (NSSE).

Classificatory activities are fundamentally critical to social sciences (de Ville 2001; Fenske, Keller, & Irwin, 1999) and are naturally seen as an intrinsic component of knowledge discovery in the data mining process (Berry & Linoff, 2000; Kantardzic, 2003). Thus, creating an institutional typology provides a unique opportunity to test and detail the extent to which data mining techniques may be applied to higher education research. This study delves into this opportunity. Specifically, this study explores the following questions:

1. What factor dimensions best capture the learning behaviors of the students?
2. What are the salient patterns of student behaviors and how are these student behavior patterns distributed within each institution?
3. Based on the percentage distributions of different types of students, is there an institutional engagement typology that captures similarities as well as variations of student engagement patterns among four-year colleges and universities?

Methodology

Data

Data from the National Survey of Student Engagement (NSSE) provided a perfect primary data source to utilize data mining techniques. NSSE[2] annually collects data from over 150,000 college students at hundreds of four-year colleges and universities across the nation. NSSE provides an alternative view of collegiate quality by focusing its attention on what students actually do during their college experience versus commercial ranking systems that rely almost exclusively on inputs such as SAT scores, class rank, and other institutional prestige or resource indictors (Boyer, 2003; Carini, Hayek, Kuh, et al. 2003; Kuh 2001; Twitchell, 2002; Zhao, Kuh, & Carini, 2005). The key data source for this study was NSSE's 2001 dataset consisting of 33,858 seniors[3] at 317 colleges and universities. Secondary data sources included information from IPEDS, Carnegie Classification, *US News and World Report*, and *Barron's* selectivity index.

Data Mining Technique

The core methodology for this study is data mining, specifically, unsupervised data mining techniques. Unsupervised data mining aims to classify and identify potentially meaningful patterns in data without a preconceived notion of an outcome (dependent) variable. To a large degree, this is to uncover the "natural" existence of clusters in the data. Unsupervised clustering techniques based on distance measures are used to generate

[1]The 2000 Carnegie Foundation dropped the level of federal support as a classification criterion.

[2]Four-year institutions chose to participate in the NSSE study on a voluntary basis; therefore, the data in this study are not a random sample representative of the national four-year institution universe. As a result, the institutional typology developed in this study is for heuristic purpose only. For detailed information on the psychometric properties of the NSSE survey and data, please refer to http://www.iun.edu/~oir/nsse/faq/2005/FAQ_2005_NSSE_Psychometric_Properties.pdf

[3]We intentionally chose seniors only in this study due to the following consideration: seniors survived the college career and are thus more identified with their respective institutions. That is, we believe seniors are a better representation of their institutions. This is based on Clark and Trow's (1966) work of student sub-cultures. We fully realize that an institutional typology derived from freshmen data may yield different results.

clusters that may lend themselves to becoming a typology. Unsupervised data mining techniques employed in this study include principal component analysis (PCA) and three clustering techniques (algorithms)—K-Means, TwoStep, and Kohonen. For a detailed explanation of the operation of the three algorithms, please refer to the SPSS manual (2004).

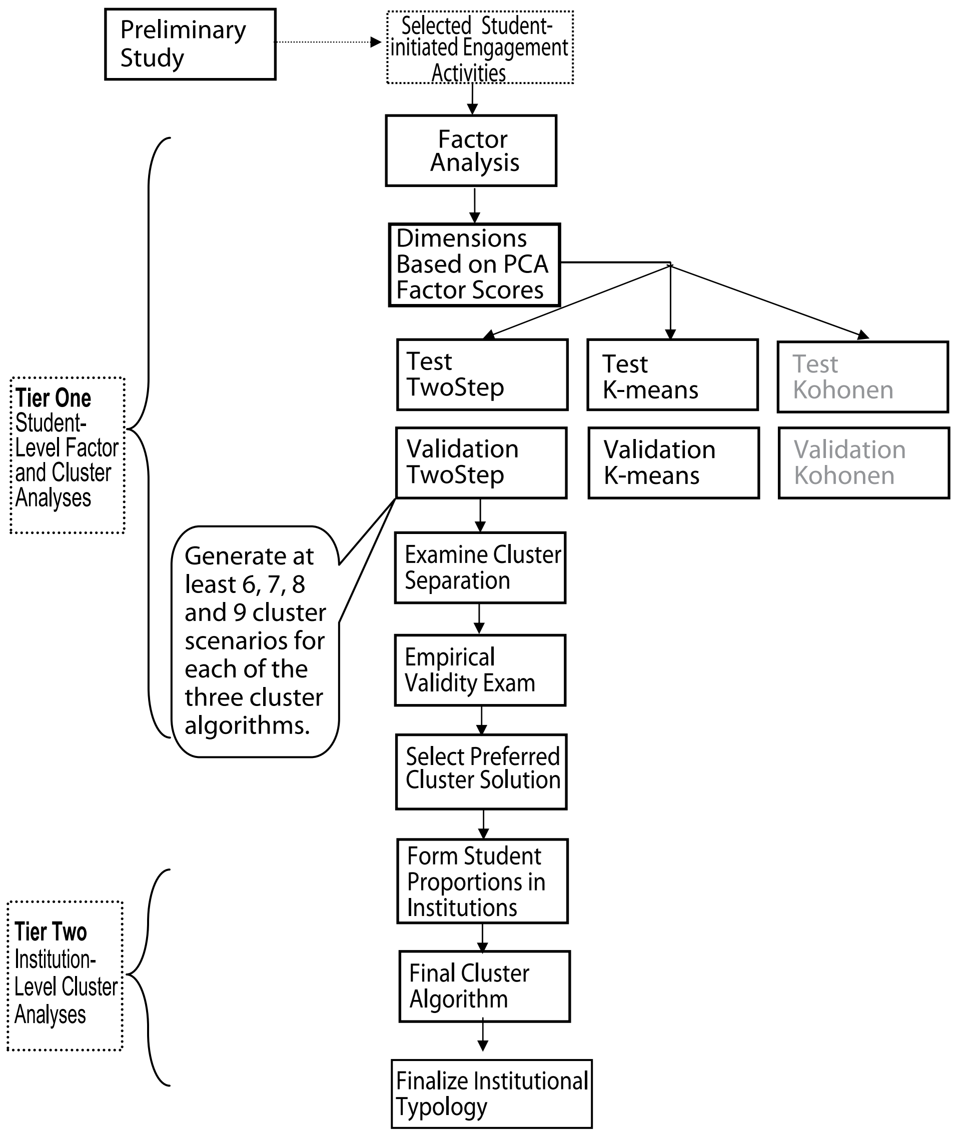
Tools

The major tools or software employed in the study were SPSS Clementine 8.0, WinCross 2.0, SPSS Base, and Microsoft Excel. Clementine is a comprehensive data mining application from SPSS that produced all the datasets and conducted all of the factor and cluster analyses for the study. WinCross is an efficient cross-tabulation software allowing multiple enumerations of cross tabulations and significance testing at one time. SPSS Base provided coding and aggregation ability for the data files. Excel, in conjunction with SPSS, was used to provide graphic rendition of a massive number of data tables for data visualization.

Analytic Strategy

Based on a preliminary exploration of the same data conducted earlier,[4] the study adopted a bottom-up and two-fold approach. That is, a factor analysis was conducted and then the derived factor scores were later used to conduct a two-tiered cluster analysis. The first-tier clustering is based on student-level data, which derived distinct student types. The second-tier clustering was conducted at the institutional level, which was based on the profiles of student engagement groups (percentages of various types of students with different engagement behaviors within each institution) derived from the first-tier clustering. The tiered clustering approach was designed to start with individual senior students' responses and rise up to an institutional-level typology.

The specific research steps are illustrated in Figure 1, including conducting factor analysis prior



Note: Grayed out elements denote discontinued process.

Figure 1. Data mining workflow.

A previous examination employed solely aggregated institutional-level NSSE data. The data mining clustering procedures produced clusters with poor face validity, and the result was difficult to interpret. One potential reason was that aggregation using central tendency measures, i.e., mean, disguised the rich complexity and difference among institutions residing in the student-level data.

to clustering analysis, performing empirical validation of the results, and finalizing institutional typology.

After identifying the research objectives, the original NSSE survey questions were closely examined to understand their empirical base. Only those questions that reflected student-initiated behaviors were selected.[5] Three dimensions of student information were used: undergraduates' academic and social activities, institutional and student characteristics (such as enrollment size, web use, parental education, location, major field of study, among others), and selected outcome measures (graduation, retention).

The study began with a data reduction procedure: Principal Component Analysis (PCA) with a variance maximizing (Varimax) rotation, in which the extraction of eigenvectors produces statistically independent factors. The Varimax rotation facilitates interpretation of the factors and maintains the orthogonal space. That is, PCA derives latent dimensions that are empirically uncorrelated with each other. This is a desirable feature in the ensuing cluster analysis as independent factor solutions can simplify cluster analysis results. In fact, PCA conducted prior to cluster analysis is a rather standard procedure in the line of classificatory studies (Bailey, 1994). It serves two important purposes: the attainment of dimensions and the individual factor scores for each of the dimensions. Upon identifying these dimensions, clustering algorithms can then obtain clusters using the factor scores[6] across these dimensions.

At the cluster analysis phase, the first-tier clustering generated student engagement typology—types of students based on their engagement behavioral activities. The second-tier cluster analysis produced institution-level engagement types through aggregating the percentage compositions of various student-level engagement types from the first-tier clustering. When a set of cluster(s) is identified and stabilized in the second-tier clustering, it is considered to be an "institutional typology". Under such a typology, institutions with similar profiles of student engagement types are grouped within the same category, and those who have divergent profiles are placed in different categories.

One of several key features of data mining is its emphasis on conducting multiple analyses using several outcomes from the same dataset for the purpose of comparing and contrasting to obtain the optimal solution(s). This study considered this approach as an "algorithmic bias" test. Recent articles have started to recognize this important capacity, among them, Angus (2003), who pointed out that with data mining, "you get every possible report evaluated, and then it delivers you the most relevant ones" (p. 48). The data mining approach used in this study typifies this procedure (i.e., to subject the data to multiple clustering algorithms of TwoStep, K-Means, and Kohonen). Each algorithm generated various cluster scenarios. For example, TwoStep algorithm produced scenarios of seven, eight, and nine, meaning there was a scenario with seven clusters, another with eight clusters, and another with nine clusters. This practice of looking at multiple alternatives should be part of traditional statistical analysis but is very rarely practiced (Lei & Koehly, 2003).

The cluster results were validated in several ways. First, the study split the data file into two equal parts (called *test dataset* and *validation dataset*). All three algorithms were run against each dataset to produce like scenarios, and various scenarios were compared for general consistency. Second, the cluster membership within each scenario was compared to look for large differences. To evaluate the distribution of membership (cluster-size validation) within a cluster solution, a general rule was used—the size of the smallest cluster membership should be more than 20% of the largest cluster membership.[7] Although cluster size and number of clusters are influenced by the nature of data (Hans & Kamber, 2001), cluster size is a priori and determining acceptable cluster size is subjective (Lazarevic,

[5] An extensive discussion of the reasons for using student behavioral data as opposed to attitudinal or demographical data can be found in Luan (2006).

[6] The factor scores are regression-based standardized scores. That is, they are calculated by taking the standardized score on each variable, multiplying by the corresponding factor loading of the variable for a given factor, and then summing them up.

[7] An extensive discussion of the approaches to understanding cluster membership, cluster size, and cluster validations can be found in Luan (2006).

[8] There are two ways to interpret Figure 3 and Table 5. One is to focus on the comparison between a group and the overall sample, and the other is to focus on the individual group comparisons with each other. Since the emphasis of the study is to identify distinctive group memberships, our discussions, therefore, are centered on the latter comparisons.

et al., 1999; Sun, 2002). This study opted to allow the smallest cluster to be at least 20% of the largest cluster, so that the smallest cluster is not dwarfed into oblivion by the largest cluster.

It is important to point out that there are no perfect clusters. Even if a validated cluster exists mathematically, it may not be the appropriate typology for practical purposes. Contextual knowledge is essential to finalize any type of typology.

Findings

Three distinct phases were used to extract and analyze the findings. Phase One extracted factors capturing student learning behaviors, the first-tier clustering in Phase Two explored student level learning engagement types followed by the second-tier clustering to uncover institution engagement typologies. Each phase produced a set of findings.

During the Phase One, after several rounds of generating and evaluating factor dimensions, the study reached an optimal group of factor dimensions. There were three sets of factor dimensions: the first set had 7 dimensions, the second had 8 dimensions, and the third had 9 dimensions. Of these three sets of factor dimensions, the 9-factor dimension was selected because it extracted most meaningful components that covered extensive aspects of student engagement behaviors. Table 1 contains the factor loadings and names of the nine student engagement dimensions.

Table 1

Nine Factor Dimensions Extracted in Clementine

	Factor Dimensions (Total Variance Explained 50.2%)								
	1	2	3	4	5	6	7	8	9
Supportive Environment									
Emphasize: Helping you cope with your non-academic responsibilities (work, family, etc.)	.744								
Emphasize: Providing the support you need to thrive socially	.744								
Emphasize: Providing the support you need to help you succeed academically	.723								
Quality: Relationships with administrative personnel and offices	.635								
Emphasize: Encouraging contact among students from different economic, social, and racial or ethnic backgrounds	.628					.305			
Quality: Relationships with faculty members	.628	.361							
Quality: Relationships with other students	.447								
Interaction with Faculty									
Discussed grades or assignments with an instructor		.685							
Talked about career plans with a faculty member or advisor		.591					.337		
Used e-mail to communicate with an instructor		.584							
Discussed ideas from your readings or classes with faculty members outside of class		.563							
Received prompt feedback from faculty on your academic performance (written or oral)	.333	.537							
Used an electronic medium (listserv, chat group, Internet, etc.) to discuss or complete an assignment		.401		.301					
Asked questions in class or contributed to class discussions		.398							-.349
Course-emphasis on Higher Order Thinking Abilities									
Synthesizing and organizing ideas, information, or experiences into new, more complex interpretations and relationships			.775						

Table 1 *(continued)*

Analyzing the basic elements of an idea, experience, or theory			.748						
Making judgments about the value of information, arguments, or methods			.733						
Applying theories or concepts to practical problems or in new situations			.725						
Collaborative and Integrative Learning									
Made a class presentation				.658					
Worked with other students on projects during class				.644					
Worked with classmates outside of class to prepare class assignments				.564					
Participated in a community-based project as part of a regular course				.448			.307		
Worked on a paper or project that required integrating ideas or information from various sources				.394					
Class Preparation									
Number of assigned textbooks, books, or book-length packs of course readings					.712				
Number of written papers or reports of 5 or more pages					.689				
Number of written papers or reports of fewer than 5 pages					.646				
Preparing for class (studying, reading, writing, rehearsing, and other activities related to your academic program)					.420			.362	
Diverse Perspectives									
Had serious conversations with students of a different race or ethnicity than your own						.832			
Had serious conversations with students who differ from you in terms of their religious beliefs, or political opinions						.791			
Discussed ideas from your readings or classes with others outside of class (students, family members, coworkers, etc.)		.341				.464			
Co-curricular									
Worked with faculty members on activities other than coursework (committees, orientation, student life activities, etc.)							.631		
Participating in co-curricular activities (organizations, campus publications, student government, or social fraternity							.597		
Working for pay on campus							.572		
Tutored or taught other students (paid or voluntary)							.532		
Working Hard									
Came to class without completing readings or assignments								-.593	
Relaxing and socializing (watching TV, partying, exercising, playing computer and other games, etc.)								-.592	
Prepared two or more drafts of a paper or assignment before turning it in				.318				.532	
Worked harder than you thought you could to meet an instructor's standards or expectations								.430	
Passive Learning									
Memorizing facts, ideas, or methods from your courses and readings so you can repeat them in pretty much the same form									.533
Emphasize: Spending significant amounts of time studying and on academic work	.306								.508
Number of books read on your own (not assigned) for personal enjoyment or academic enrichment					.313	.305			-.344
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.									
A rotation converged in 11 iterations.									

Phase Two of the analysis was intended to generate student engagement types. This phase relied upon unsupervised data mining technique that was achieved through using K-Means and TwoStep algorithms. Both techniques produced multiple cluster scenarios. After a close examination of consistency and difference among the scenarios, the 8-cluster scenario generated by the TwoStep approach was selected. The decision was based on four pieces of information: (a) the graphical rendition of the cluster separation; (b) cluster membership distribution; (c) empirical cluster validation; and (d) cluster membership demographics.

Figure 2 presents the graphical rendition of the cluster separation using the eight clusters produced by the TwoStep clustering algorithm. In clustering, the further apart the clusters are, the more different the clusters. The standardized factor mean scores within each cluster were used to produce the charts, as shown in Table 2. Comparing to all the other clusters, the eight clusters graphed in Figure 2 showed the largest amount of difference among themselves.

To illustrate the generic process of interpreting Figure 2, it is apparent that Cluster 1 has a much higher score on Co-curricular Activities, while remaining slightly above average in other dimensions. Cluster 2 is somewhat similar to Cluster 1, except for high scores on Interaction with Faculty and low

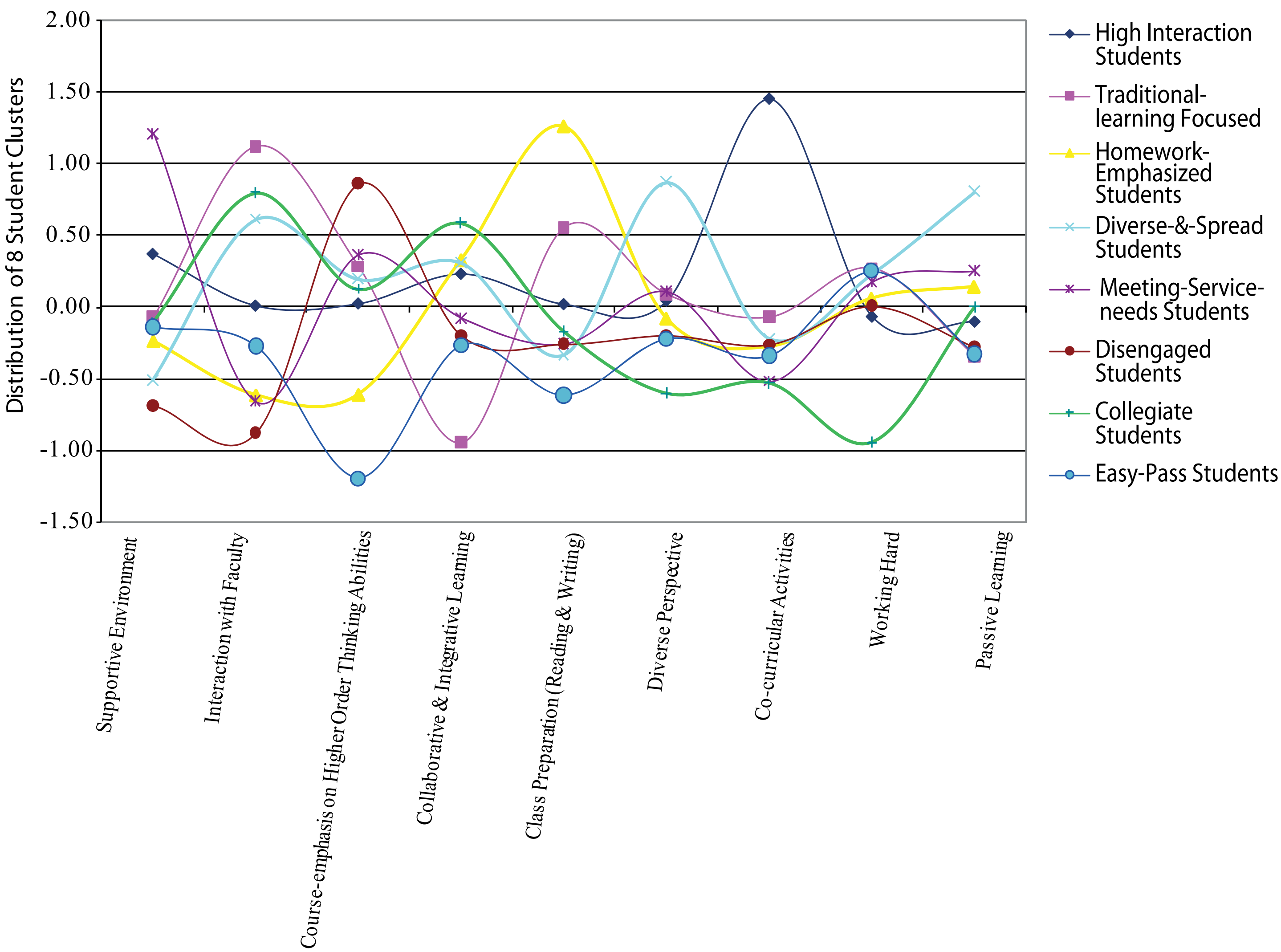


Figure 2. **TwoStep** generated eight student cluster scenarios using standardized factor mean scores from the nine dimensions

Table 2

Standardized Factor Mean Scores from the Nine Dimensions for the Eight Student Clusters

	Supportive Environment	Interaction with Faculty	Course-emphasis on Higher order Thinking Abilities	Collaborative and Integrative Learning	Class Preparation (Reading & Writing)	Diverse Perspective	Co-curricular Activites	Working Hard	Passive Learning	Number of Students
Cluster 1 (High-Interaction Students)	0.37	0.01	0.02	0.23	0.02	0.04	**1.45**	-0.07	-0.10	6140
Cluster 2 (Traditional-Learning-Focused students)	-0.07	**1.12**	0.28	**-0.94**	0.56	0.09	-0.07	0.27	-0.34	3911
Cluster 3 (Homework-Emphasized Students)	-0.24	-0.61	-0.61	0.33	**1.26**	-0.08	-0.28	0.06	0.14	3638
Cluster 4 (Diverse-and-Spread Students)	-0.51	0.61	0.20	0.31	-0.33	**0.87**	-0.22	0.23	**0.80**	3909
Cluster 5 (Meeting-Service-Needs Students)	**1.21**	-0.66	0.36	-0.08	-0.26	0.10	-0.52	0.17	0.25	4004
Cluster 6 (Disengaged Students)	**-0.68**	**-0.87**	**0.86**	-0.20	-0.26	-0.20	-0.26	0.01	-0.28	4303
Cluster 7(Collegiate Students)	-0.11	**0.80**	0.12	**0.59**	-0.17	**-0.60**	**-0.53**	**-0.95**	0.00	3826
Cluster 8 (Easy-Pass Students)	-0.14	-0.27	**-1.19**	-0.27	**-0.62**	-0.22	-0.34	0.25	-0.33	4548

Note. Bold-faced numbers indicate either relative highest or lowest factor mean scores in the rows.

on Collaborative and Integrative Learning. Cluster 3 is similar to Cluster 2 only half the time, while remaining quite different the rest of the time. Graphics like these are an important and necessary data mining technique to assist with interpreting quantitative data and conducting, among other things, face validity verification. The sheer number of data tables produced by Clementine would be extremely difficult to interpret without data visualization.

The next step, upon identifying the student engagement cluster scenarios, was to conduct empirical validation using individual students' records. One of the effective validation approaches was to use student characteristics as well as institutional characteristics to examine the characteristics of the students in these clusters. The student characteristics include race, gender, enrollment status, on-off campus status, specific major field of study, aggregated major field of study, transfer status, fraternity or sorority membership, age, and parental education. The institutional characteristics include Carnegie Classification, sector (private versus public), and Barron's selectivity. Z-tests (Morton, 2000) of proportions were employed to determine if there are significant differences across various student engagement types by these student and institutional characteristics.

The results of the numerous Z-tests of proportions were generated by WinCross, a powerful cross-tabulation software that allows individual comparisons of each subgroup of a demographic field against all other subgroups of that field at the same time (see Table 3). The conventional way is to compute each pair individually, which results in massive number of individual tables. Adding an alternative significance level (e.g., $p < .05$ vs. $p < .01$) would double the number of tables.

Selected results from WinCross are presented in Table 3. Whenever a cell (sub-group) was significantly different, a capital letter was placed underneath the cell with the larger proportion. For example, in Table 3, using the Cluster 1 in the first table, letter "J" appeared underneath the cell of "Male." In this case, the letter "J" refers to Females, which is in column "J." Therefore, the table indicates that there is a higher proportion of males than females in this cluster and the difference is significant. Additional tables are available from the first author.

Using the eight clusters of student engagement types derived through the TwoStep clustering algorithm, WinCross found a number of clusters to be significantly different from each other at the .01 level on various demographics. Combining

Table 3

WinCross of the Eight Clusters and Select Demographics

	RACE								SEX		FULL-TIME		ON/OFF CAMPUS	
	TOTAL	American Indian	Asian	Black	Latino	White	Otrace	Multiple race	Male	Female	On <F/T	Off F/T	campus	campus
	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)	(M)	(N)
TOTAL	34279 100.0	216 100.0	1837 100.0	1896 100.0	1412 100.0	27187 100.0	49 100.0	1591 100.0	12321 100.0	21746 100.0	5558 100.0	28496 100.0	7635 100.0	26402 100.0
High-Interaction Students	6140 17.9	33 15.3	309 16.8	366 19.3	233 16.5	4905 18.0	10 20.4	270 17.0	2359 19.1 J	3741 17.2	423 7.6	5671 19.9 K	2482 32.5 N	3615 13.7
Traditional-Learning Focused Students	3911 11.4	16 7.4	102 5.6	97 5.1	88 6.2	3418 12.6 BCDEH	14 28.6 BCDEH	164 10.3 CDE	1359 11.0	2532 11.6	401 7.2	3492 12.3 K	1209 15.8 N	2679 10.1
Homework-Emphasized Students	3638 10.6	22 10.2	258 14.0 DEFGH	131 6.9	138 9.8 D	2910 10.7 D	2 4.1	166 10.4 D	1116 9.1	2493 11.5 I	407 7.3	3209 11.3 K	710 9.3	2901 11.0 M
Diverse-and-Spread Students	3909 11.4	34 15.7	242 13.2 F	352 18.6 CEFH	192 13.6 F	2821 10.4	10 20.4	243 15.3 F	1269 10.3	2610 12.0 I	607 10.9	3272 11.5	694 9.1	3186 12.1 M
Meeting-Service-Needs Students	4004 11.7	36 16.7	276 15.0 F	300 15.8 FG	276 19.5 CDFGH	2879 10.6	3 6.1	224 14.1 F	1273 10.3	2709 12.5 I	861 15.5 L	3110 10.9	593 7.8	3375 12.8 M
Disengaged Students	4303 12.6	27 12.5	259 14.1	243 12.8	188 13.3	3372 12.4	4 8.2	199 12.5	1580 12.8	2698 12.4	953 17.1 L	3320 11.7	636 8.3	3640 13.8 M
Collegiate Students	3826 11.2	15 6.9	152 8.3	142 7.5	109 7.7	3273 12.0 BCDEGH	2 4.1	123 7.7	1630 13.2 J	2174 10.0	483 8.7	3320 11.7 K	715 9.4	3085 11.7 M
Easy-Pass Students	4548 13.3	33 15.3	239 13.0	265 14.0	188 13.3	3609 13.3	4 8.2	202 12.7	1735 14.1 J	2789 12.8	1423 25.6 L	3102 10.9	596 7.8	3921 14.9 M

the knowledge gained so far, the following student engagement types were defined:

1. High-Interaction
2. Traditional-Learning-Focused
3. Homework-Emphasized
4. Diverse-and-Spread
5. Meeting-Service-Needs
6. Disengaged
7. Collegiate
8. Easy-Pass

Generally speaking, High-Interaction, Traditional-Learning-Focused, Homework-Emphasized, and Collegiate students tend to be traditional students that enroll full-time and do not have much off-campus responsibilities. In contrast, Diverse-and-Spread, Meeting-Service-Needs, Disengaged, and Easy-Pass students tend to be part-time, non-traditional students, working professionals, and minorities in majors such as engineering. Summarized in Table 4, these student engagement types served as data in the final phase—defining the institutional typology.

For the final phase, Phase Three, upon obtaining the student engagement types, an institutional-level file was created that contained the *percentage distributions* of various

Table 4

Student Engagement Types

Student Types	Engagement Description	Background description
High-Interaction Students	(Doer, Busy Bee): Students in this group are highly engaged in a variety of co-curricular activities (such as organizations, campus publications, student government, social fraternity or sorority…, working with faculty members on activities other than course work, working for pay on campus, tutoring or teaching other students), perceive their campus environment to be supportive and friendly, and are positively engaged in collaborative and integrative learning together with their peers.	This type is over-represented by male, full-time students, living on campus, tend to be in math and science majors, or in Bac-LA, Bac-Gen, and Master's institutions, non-transfer students, traditional age, higher parental education level, and studying in more competitive institutions.
Traditional-Learning-Focused Students	(Learner, Teacher's Pet, Academic): Students in this group are highly engaged in frequently interacting with faculty members. They work hard academically and think the course work provide enough emphasis on higher order thinking abilities. They are less engaged in collaborative and integrative learning.	Over-represented by whites, full-time students, living on campus, humanity and social science majors, non-transfer students, non-Greek, traditional aged, high parental education level, and tend to study in public and more competitive institutions. In addition, students in Bac-LA institutions are over-represented in this group.
Homework-Emphasized Students	(Reader, Avid Reader): Students clustered under this category do a lot of reading and writing and spend more time preparing for classes. They also do more collaborative and integrative learning with their peers. They interact less frequently with faculty members and are less likely to engage in co-curricular activities. They perceive their campus environment to be less supportive.	White or Asian students are over-represented, and there is a lower percentage of Blacks in this group. They tend to be traditional-aged, full-time and living off-campus. They are less likely to major in math and sciences. They tend to study in a large-size university with mid-ranged competitiveness.
Diverse-and-Spread Students	(Driller): These students frequently encounter people from diverse backgrounds, interact frequently with faculty members, and think course work adequately emphasized higher-order thinking abilities. They work hard and perceive their institutions emphasize a great deal on academic learning. Their course learning also highly involves memorizing the facts. They are engaged in collaborative and integrative learning to a great degree but, nonetheless, spend less time reading and writing or engaging in co-curricular activities.	This group of students is over-represented by Black, female, off-campus, math or sciences major, transfer, adult, and lower parental education level. They also tend to study in large size institutions such as Doc-ext, Doc-int, or Master's institutions.
Meeting-Service-Needs Students	(Contented): Students in this group are very satisfied with the supportiveness of their campus environment despite the fact that they do not interact frequently with their faculty members or peers and do not engage actively in co-curricular activities. They engage in diversity activities to a moderate degree but do not spend a good deal of time in academic work.	This type of students includes a higher percentage of Latino/Black students, tend to be part-time, off-campus, and major in social science and pre-professional fields. There is a higher percentage of transfer, non-Greek, adult students, and students from families with a lower parental education level in this group. These students tend to be from Master's, Bac-Gen, or less competitive institutions.
Disengaged Students	(Over-Challenged): Students in this group are less satisfied with their campus environment due to a lack of support in meeting their academic and social needs. In addition, they feel their course work challenges them to a great degree in terms of higher-order thinking abilities. These students also do not have extensive bonds with their institutions because of a lack of engagement in both social and academic activities.	This group is over-represented by part-time, off-campus, transfer, non-Greek, adult students or students from families with a lower parental education level. They tend to major in social sciences and pre-professional fields and from non-Bac-LA institutions. There are also a relatively high percentage of students in this group studying in private and more competitive institutions.
Collegiate Students	(Conventional): This group of students actively engages in interaction with faculty members and in collaborative and integrative learning with their peers. They perceive their course work moderately emphasizes higher-order thinking abilities. On the other hand, they are less engaged in diversity-related and co-curricular activities and do not spend a lot of time preparing for class and doing homework.	This group of students tends to be White, male, full-time, living off-campus, majoring in math/science and pre-professional, from Doc-ext, Doc-Int, and Master's institutions. They also tend to be non-transfer, Greek, traditional aged, with higher parental education level, and from private and competitive institutions.
Easy-Pass Students	(Insufficiently Challenged): The most apparent feature of this group of students is that they do not perceive their course-work to emphasize higher-order thinking abilities. In other words, they are not sufficiently challenged in course learning. They work hard but, generally speaking, they are less engaged in the all the other social and academic aspects of college life.	This group of students is over-represented by male, part-time, off-campus living, and humanities and pre-professional majors. They tend to be transfer students, non-Greek, adult students, and with lower parental education level. They are less likely to study in a Bac-LA institution; however, they tend to be in private and less competitive institutions.

student engagement types within each participating institution. This file was produced by assigning each student an engagement type membership based on student engagement types obtained in the previous step and then aggregating the data from the individual student level to the institutional level using the IPEDS identification numbers as the key. The TwoStep clustering algorithm was used again for the cluster analysis to produce institutional cluster scenarios ranging from three to ten clusters. Again, multiple cluster solutions were obtained. A four-cluster model and the cluster separations are presented in Figure 3 and Table 5[8] for the purpose of illustrating one of the better institutional cluster scenarios.

Type One (Diverse Curriculum and Student Life) institutions feature a large percentage of non-traditional students. These institutions have the largest percentage (22%) of students who do not feel course work emphasizes higher-order thinking abilities (Easy-Pass students), but it also has the largest percentage of students who feel course-work challenges them to a great extent on higher-order thinking abilities (19% Disengaged students). Neither group sees the environment as supportive or nor do they interact much with faculty. They are somewhat above average on the percentage that sees the institution as meeting their

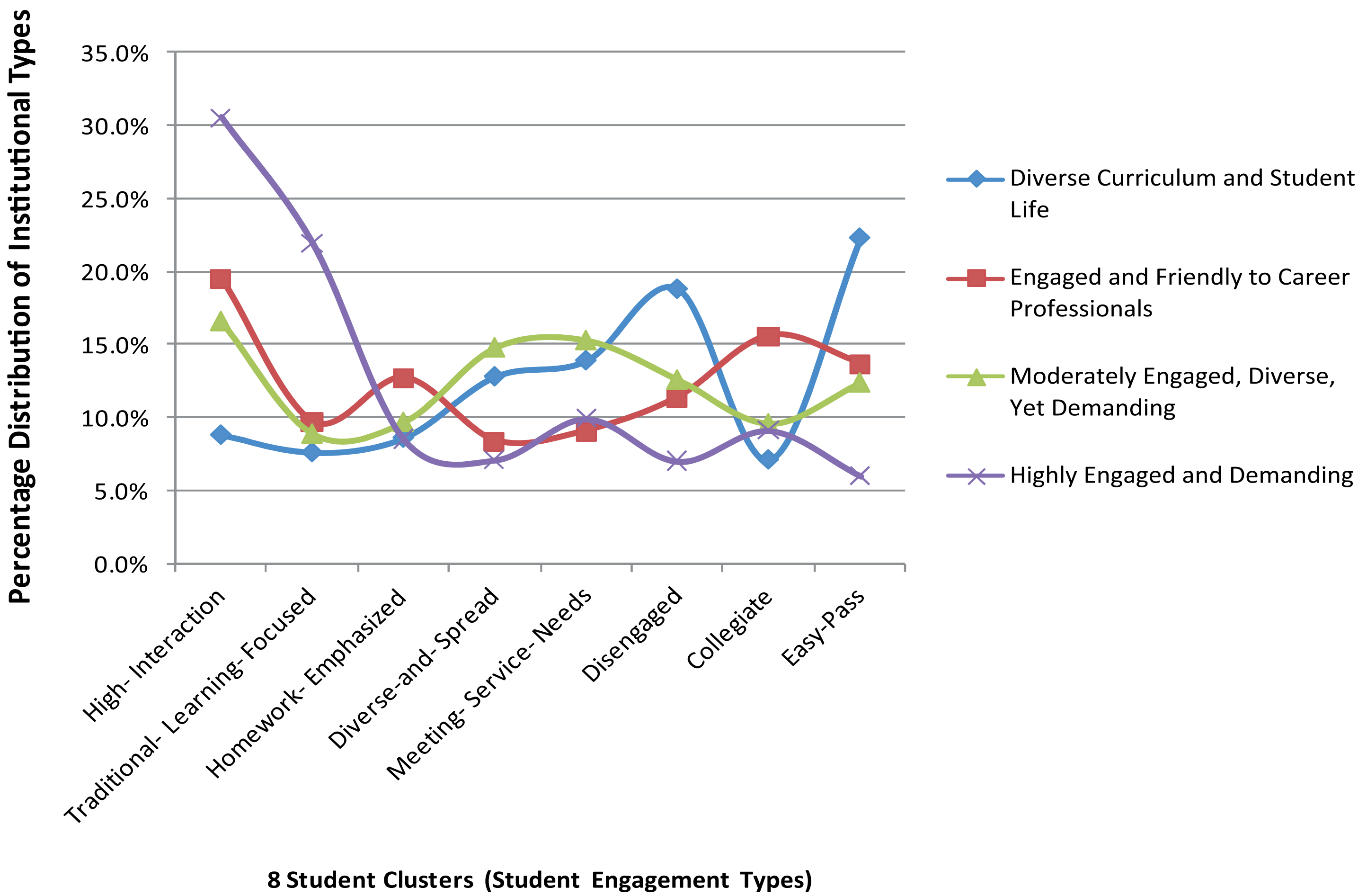


Figure 3. An example of a 4-cluster NSSE institution typology using percentage distributions of student engagement types.

[8] There are two ways to interpret Figure 3 and Table 5. One is to focus on the comparison between a group and the overall sample, and the other is to focus on the individual group comparisons with each other. Since the emphasis of the study is to identify distinctive group memberships, our discussions, therefore, are centered on the latter comparisons.

Table 5

An Example of a 4-Cluster NSSE Institution Typology Using Percentage Distributions of Student Engagement Type

Institution Type	High-Interaction	Traditional-Learning-Focused	Homework-Emphasized	Diverse-and-Spread	Meeting-Service-Needs	Disengaged	Collegiate	Easy-Pass	Number of Institutions
Type One: Diverse Curriculum and Student Life	8.8%	7.6%	8.6%	**12.8%**	**13.9%**	**18.8%**	**7.1%**	**22.3%**	57
Type Two: Engaged and Friendly to Career Professionals	**19.5%**	9.7%	**12.7%**	8.4%	9.1%	11.4%	**15.6%**	**13.7%**	90
Type Three: Moderately Engaged, Diverse, Yet Demanding	**16.6%**	8.9%	9.7%	**14.8%**	**15.3%**	**12.6%**	9.6%	12.4%	97
Type Four: Highly Engaged and Demanding	**30.5%**	**21.9%**	8.5%	**7.1%**	9.9%	**7.0%**	9.1%	**6.0%**	73
Total percentage of student types	17.9%	11.4%	10.6%	11.4%	11.7%	12.6%	11.2%	13.3%	

Note. Bold-faced numbers indicate either relative highest or lowest percentage distributions in the rows.

service needs. Fourteen percent (14%) are very satisfied with the supportiveness of their campus environment, and 13% engage in diversity-related activities and spend more time on academic learning. Fifty-seven institutions belong to this group which accounts for 18% of all institutions in the study.

Type Two (Engaged and Friendly to Career Professionals) institutions feature about 20% of the students who are highly engaged and immerse themselves in a wide array of co-curricular activities. About 16% are students in a conventional sense in that they actively engaged in campus activities; however, they report less interaction with people from diverse backgrounds. Another 13% of students are reading- and writing-intensive, interact less frequently with faculty members, and do not actively participate in co-curricular activities. About 14% of students tend to be more lax with their academic work and are not actively involved in campus life due to their other responsibilities. This type of institution only has about 10% of students who are highly engaged in interactions with faculty members. Ninety institutions belong to the Type Two cluster, which equates to 28.4% of all colleges and universities in the study.

Type Three (Moderately Engaged, Diverse, Yet Demanding) institutions generate a slightly lower percentage of high-interaction students (17%). Approximately 15% of students, who tend to be non-traditional students, are satisfied with campus services and support; 15% of the students engage in diversity related activities, yet feel academically challenged and work rigorously. Ninety-seven institutions are in this type or 30.6% of the institutions in the study.

Type Four (Highly Engaged and Demanding) institutions have over 50% of students who are highly engaged in traditional learning-focused academic and social activities on campus (21.9%) and frequently interact with faculty members (30.5%). Type Four institutions have a high percentage of traditional, residential students. This type of institution has the lowest percentage of students who see it as being a low academic challenge (6% Easy-Pass students) and also the lowest percentage of Disengaged students (7%). There are 73 institutions in this type or about 23% of all institutions in this study. Only slightly over 30% of the students are highly engaged in various campus social and academic activities.

To further understand the institutional engagement typology, the study compared the 4-cluster type with the widely accepted 2000 Carnegie Classification. The analysis examined the overlapping as well as divergence of the two different classification approaches and, therefore, provided an anchor to further understand diverse colleges and universities from multiple lenses. Table 6 is a cross tabulation of the four types of institutions and the 2000 Carnegie Classification. Except for liberal arts colleges that, on average, generate more intensive student engagement and, therefore, are concentrated in the fourth cluster type, distributions of other colleges and universities in the study did not necessarily appear to converge with the Carnegie categories.

Discussion and Implications

This study used a "bottom-up" strategy that looked at student-level behavioral data to initially arrive at student types of learning engagement. It then examined the percentage distribution of different student types within each institution, and finally conducted a second round of cluster grouping to establish an institutional typology. Using student-level data to derive the institutional typology demonstrated a clear advantage in that it naturally retained the important component of students, and their unique behavioral experiences. Therefore, it produced much richer information to better understand higher education institutions.

Although only the four-cluster solution was presented in the findings as an example of how to apply this study, the study actually generated eight different solutions for the final institutional typology, ranging from three clusters to nine clusters. Classification, by nature, is not stand-alone or fixed. As pointed out by the science philosopher, Abraham Wolf (1930), "classification is not only of individuals into classes, but also of classes into wider or higher classes, and of those into higher classes" (p. 32). Therefore, classification can also be viewed as a hierarchical structure with each layer representing different level of granularity, and the more specific institution types can be

Table 6

Cross-tabulation of the Four-Typology Scenario and 2000 Carnegie Classification

Institution Engagement Type		2000 Carnegie Classification (McCormick, 2001)						
		DRU-EXT	DRU-INT	MASTERS	BAC-LA	BAC-GEN	OTHER	Total
Diverse Curriculum and Student Life	Count	11	8	30	2	4	2	57
	% Within Institution Engagement Type	19.3%	14.0%	52.6%	3.5%	7.0%	3.5%	100.0%
Engaged and Friendly to Career Professionals	Count	18	8	45	7	11	1	90
	% Within Institution Engagement Type	20.0%	8.9%	50.0%	7.8%	12.2%	1.1%	100.0%
Moderately Engaged, Diverse, Yet Demanding	Count	19	15	43	10	9	1	97
	% Within Institution Engagement Type	19.6%	15.5%	44.3%	10.3%	9.3%	1.0%	100.0%
Highly Engaged and Demanding	Count	1	2	16	45	6	3	73
	% Within Institution Engagement Type	1.4%	2.7%	21.9%	61.6%	8.2%	4.1%	100.0%
Total	Count	49	33	134	64	30	7	317
	% Within Institution Engagement Type	15.5%	10.4%	42.3%	20.2%	9.5%	2.2%	100.0%

grouped under a broader category, like a tree diagram. In this study, a fewer-category cluster scenario represented a bird's-eye view of the institution engagement patterns. Looking deeper, the more elaborate and refined cluster solutions would emerge under this general framework and, therefore, present a more detailed picture of student engagement among colleges and universities. There is a traceable linkage across the various institution cluster scenarios.

In choosing a final solution type, it was also important to assess whether or not there might be an "optimal" number of clusters for the given context. Despite the fact that a more diversified higher education system may better serve the needs of an increasingly diversified student body, the American higher education system is actually gravitating towards more homogeneity, as a result of institutions mimicking a similar set of values and components presented in the top research universities (DiMaggio & Powell, 1983). In addition, for an institutional typology to retain practicability and simplicity, we deliberately limited the number of categories under ten.

"How good is good engagement?" This is a question often-asked, yet difficult to answer. In this study, data mining provided a new framework based on empirical thresholds to examine the range and degree of student engagement among different types of institutions. These empirical thresholds were based on the percentage distributions of the types of student engagement. Theoretically, students learn more with more engagement. However, when taking into consideration the difference of student body among different types of institutions, it is neither fair nor appropriate to consider a single one-size-fit-all engagement type as the universal and ideal engagement type.

Student engagement data have proven to be a powerful alternative to understanding collegiate quality. An institutional typology based on student engagement in effective educational practices represents a paradigm shift from the dualistic view of popular college ranking practices. Ranking assumes that colleges are stacked up as a pyramid with the best atop. In contrast, institution engagement typology categorizes institutions into coherent groups strictly based on student learning engagement activities, and, therefore, presents an objective description of colleges and universities. Instead of assuming one type of institution to be superior to others, institution engagement typology recognizes the diversity, difference, and uniqueness of colleges and universities. In this sense, it is non-rankable.

Naturally, students are a critical factor to consider in describing colleges and universities. Numerous studies have shown that both the educational quality and the overall fit between students and colleges are critical in student learning and development (Clark & Trow, 1966; Kuh, Hu, & Vesper, 2000; Pace, 1963, 1969). However, the common ranking approaches focus on passive and subjective criteria such as institutional resources and reputation, and student difference is largely ignored. The institutional engagement typology developed based on student engagement patterns, acknowledges that a "best" college or university is a rich learning environment providing an optimal balance of challenge and support depending on the unique situation of each student. By incorporating the active ingredient of student behaviors, the institution engagement typology is an important breakthrough in understanding higher education institutions.

The data mining methodology and resulting typology appear to have a number of implications for higher education researchers and practitioners. First, the national distribution of institution engagement types presents a fresh view of American higher education system from the perspective of student engagement, divergent from previous typological efforts. Second, NSSE's institutional engagement typology provides an empirical reference to address an often-asked question "how good is good engagement?" It provides a comparative threshold to examine the range and degree of student engagement among different types of institutions. Third, the typology crystallizes the uniqueness of institutions. Guided by the typology, institutions can identify peer institutions with a similar niche. Finally, the institutional engagement typology can assist to identify effective institutional policy and practices that foster improvement in engagement aspects important to institutional mission.

In addition, several other strands of information derived through this effort are valuable. For example, the intermediate step identified *student* engagement typology, in turn, percentage composition of various types of students within each institution. Using it as a mirror, an institution can see its student population with more clarity. Institutions can check the alignment of the composition of different student types against institution's mission and take timely actions if divergence occurs. In addition, using this information, institutions can purposely target resources to further desired change in more efficient ways.

The institution engagement typology incorporates the largely ignored component, students, in understanding colleges and universities. It provides a fresh outlook in the dialogue of collegiate quality and has very practical implications. It can be used to assist higher education researchers as well as the general public to understand American higher education system from the lens of student engagement, to help institutions achieve self-understanding and improvement, and to guide parents and students in making informed decisions on college choice.

Future research tasks may be to continue exploring and validating the typology using NSSE data from later years or from a first-year student sample. The use of data from senior students is informative since the student-institution fit may have implications on the types of students likely to be retained until they are seniors. In contrast, the freshman data will have a much larger number of students who did not transfer but who matriculated directly to the institution. The model structure, or "model stream" in data mining terms, developed for this study, could be modified fairly quickly to apply to new datasets and look at possible patterns among non-surveyed schools. Further drill-down studies within a particular cluster of institutions could also be helpful to uncover institutional programs and services that may better fulfill the needs of certain students.

Conclusion

Several conclusions appear warranted from the study. First, although somewhat complicated to document and understand, the step-by-step process of this study provides evidence that data mining techniques affirm themselves as a new and powerful approaches to exploring higher education data and research. The more data mining is used and understood, the greater the opportunities for uncovering new patterns in student or institutional behavior that up till now could easily have gone unnoticed using traditional statistical techniques. The challenge will be to increasingly demonstrate applications that are easily transferable to assist with better decision making and interpretation of data at all levels of analyses.

Second, for the very first time, data mining was utilized to create a meaningful institutional typology based on student engagement data. The combination of this new statistical approach with relatively new data from NSSE provides a significant shift away from previous institution classification systems. The Carnegie Classification of Institutions was specifically not used as a guide for this study since it was not created for the purpose of understanding learning, especially students' engagement in learning. Naturally, the final institutional typology yielded different results from the Carnegie Classification.

Finally, the existence of an institutional typology based on student engagement patterns provides a framework for more meaningful dialogue on undergraduate quality. The results could allow participating NSSE colleges and universities within a particular type to compare among themselves for the purposes of identifying benchmarks, leveraging their strengths, and making improvements, where necessary. The process may open up an entirely new set of peer or aspiration institutions for a given college or university. In addition, as an alternative to existing rankings and classifications of colleges and universities, the student engagement-based typology might provide the general public the possibility of better matching students' learning styles with what institutions can uniquely offer.

Final Thoughts

Employing a data mining-based approach to develop an institutional typology based on student engagement data is a pioneering endeavor in higher education research—both from a conceptual and a methodological perspective. Using student-level behavioral data to derive institution typologies is advantageous over prior typological studies in that it provides much richer information and naturally retains the important component of students in understanding higher education institutions. Research informs us that there is much more variation at the student level than at the institutional level when looking at process indicators such as engagement. The sophistication and technological know-how of data mining has just begun to shed new light on higher education and institutional research.

Editor's Note:

The calls for accountability and accessibility in the political and the public rhetoric have reinforced the concern about student success.

The calls for student engagement in active learning have reinforced the concern for effective academic quality. The calls for a sustainable society have reinforced the concern for social, economic, and environmental issues. All of these movements, and the multitude of variations they take, have moved our institutions to a self-reflection using various analytics. The challenge is that there is no absolute standard by which institutions can determine how well they are doing in the pursuit of their goals. The traditional wisdom is comparison with other appropriately similar institutions. This becomes problematic since most groupings are not based on student aspects of learning.

Selecting comparative institutions for benchmarking is where this IR Applications by Luan, Zhao and Hayek is going to be a major first step in thinking about forming groups of institutions based on student experiences. Theirs is an excellent example of using data about student learning to research the types of institutions students attend. They start with a huge amount of NSSE data and focus almost exclusively on the data reflecting student actions and/or behaviors as collected in the survey. The objective is to reduce the data into understandable chunks and then to use those chunks to describe the types of institutions. It is valuable to note that their transparency of methodology is essential in evaluating their results from the first steps to the last conclusions, and it is a rather complex methodology requiring professional judgments as well as analytical skill.

For example, is it desirable to select only a subset of the NSSE data—say exclude those from special purpose institutions? This would most likely result in a different set of factors, which would result in a different set of student groups, which would result in a different set of institutions. Should more student characteristics such as FTE/HC or percent Pell Grants or average debt or curriculum have been included in the second stage of clustering where institutions are grouped into major categories? An institution concerned about affordable education might want to include average student economic attributes at this second stage of the analysis. If different variables are included, the resulting institutional categories would most likely change. Figure 1 provides an excellent place to consider how the methodology might be modified in these and a multitude of other ways to fit unique needs for various purposes.

The range of options for variables at both stages, the range of different ways students and institutions might be selected for those analyses, and the range of alternatives in both the factor and the cluster analyses are among the decisions that need to be made in a purposeful and informed manner. All of these will influence the results and the interpretation of the results.

I really enjoyed the way Luan, Zhao, and Hayek laid out their reasoning and discussed some of their options. They also described some of the challenges they faced in terms of the magnitude of the data and the magnitude of the results from their analyses. Both of these discussions provide valuable insight for those who would help us take the next steps in looking at our outcomes and in understanding our student cultures.

References

Angus, J. (2003). BA predicts the future. *InfoWorld, 25(*34), p. 46.

Astin, A. (1973). Measurement and determinants of the outputs of higher education. In L. Solomon, & P. Taubman (Eds.), *Does college matter? Some evidence on the impacts of higher education.* New York: Academic Press.

Astin, A. (1993). An empirical typology of college students. *Journal of College Student Development, 34,* 36–46.

Bailey, K. (1994). *Typologies and taxonomies: An introduction to classification techniques.* (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-102). Thousand Oaks, CA: Sage.

Berry, M., & Linoff, G. (2000). *Mastering data mining: The art and science of customer relationship management* (2nd ed.). New York: John Wiley & Sons.

Boyer, P. (2003). *College ranking exposed: The art of getting a quality education in the 21st century.* Australia: Thomson Peterson's.

Carini, R. M., Hayek, J. C., Kuh, G. K., Kennedy, J. M., & Ouimet, J. A. (2003). College student responses to web and paper surveys: Does mode matter? *Research in Higher Education, 44*(1), 1–19.

Carnegie Foundation for the Advancement of Teaching, The (2008). *The Carnegie classification of institutions of higher education.* Retrieved January 10, 2009, from www.carnegiefoundation.org/classifications

Clark, B. R., & Trow, M. (1966). The organizational context. In T. M. Newcomb, & E. K. Wilson (Eds.), *College peer groups: Problems and prospects for research* (pp. 17–70). Chicago: Aldine.

de Ville, B. (2001). *Microsoft data mining: Integrated business intelligence for e-commerce and knowledge management.* Boston: Digital Press.

DiMaggio, P., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociology Review, 48(*2), 147–160.

Dyche, J. (2000). *e-Data: Turning data into information with data warehousing*. Reading, MA: Addison Wesley Longman, Inc.

Fenske, R., Keller, J., & Irwin, G. (1999). Toward a typology of early intervention programs. *Advances in Education Research,* Vol. 4.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Han, J. W., & Kamber, M. (2001) *Data mining: Concept and techniques.* San Francisco: Academic Press.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining.* Cambridge, MA: The MIT Press.

Kantardzic, M. (2003). *Data mining explained: Concepts, models, methods, and algorithms*. Boston: Digital Press.

Kuh, G. (2001). *The National Survey of Student Engagement: Conceptual framework and overview of psychometric properties*. Bloomington, IN: Indiana University Center for Postsecondary Research and Planning.

Kuh, G. D., Hu, S., & Vesper, N. (2000). They shall be known by what they do: An activities-based typology of college students. *Journal of College Student Development*, *41(*2), 228–244.

Lazarevic, A., Xiaowei Xu, Fiez, T., & Obradovic, Z. (1999, July). Clustering–regression–ordering steps for knowledge discovery in spatial databases. Presentation at the *International Joint Conference on Neural Networks* (IJCNN '99), Washington, DC.

Lei, P. W., & Koehly, L. M. (2003). Linear discriminant analysis versus logistic regression: A comparison of classification errors in the two-group case. *The Journal of Experimental Education, 72(*1), 25–49.

Luan, J. (2002, June). *Data mining: Predictive modeling & clustering essentials*. Presentation at the 44th AIR Forum, Toronto, Canada.

Luan, J. (2003, May). *Developing learner concentric learning outcome typologies using clustering and decision trees of data mining (The OIndex report).* Presentation at 43rd AIR Forum, Tampa, FL.

Luan, J. (2006, July 19). Using Academic Behavior Index (AB-Index) to develop a learner typology for managing enrollment and course offerings—A data mining approach. *IR Applications.* Vol.10.

McCormick, A. C. (Ed.) (2001). *The Carnegie Classification of institutions of higher education, 2000 Edition*. Menlo Park, CA: The Carnegie Foundation for the Advancement of Teaching.

Morton, C. (Ed.) (2000) *WinCross user guide.* Chicago: The Analytical Group.

Mosteller, F., & Tukey, J. (1977). *Data analysis and regression.* Reading, MA: Addison-Wesley.

Pace, C. R. (1963). Differences in campus atmosphere. In W. W. Charters, & N. Gage (Eds.), *Readings in the social psychology of education.* Boston: Allyn & Bacon.

Pace, C. R. (1969). *College and university environment scales* (3rd ed.). Princeton, NJ: Educational Testing Service.

Pace, C. R. (1984). *Student effort: A new key to assessing quality (Report No. 1).* Los Angeles: University of California, Higher Education Research Institute.

Pascarella, E. T., & Terenzini, P. T. (1991). *How college affects students: Findings and insights from twenty years of research.* San Francisco: Jossey-Bass.

Sun, Y. (2002) Determining the size of spatial clusters in focused tests: Comparing two methods by means of simulation in a GIS. *Journal of Geographical Systems, 4,* 359–370.

Tukey, J. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Twitchell, J. (2002, November 25). The branding of higher education: Universities say they hate magazine's best and worst rankings. Oh, really? *Forbes*.

Wolf, A. (1930). *Essentials of scientific method.* New York: The Macmillan Company.

Zhao, C., Kuh, G. D., & Carini, R. M. (2005). A comparison of international student and American student engagement in effective educational practices. *Journal of Higher Education*, Vol. 76.

Zhao, C., & Luan, J. (2006). Data mining: Going beyond traditional statistics. In J. Luan, & C. M. Zhao, (Eds), Chapter 1 of Data mining in action: Case studies of enrollment management, *New Directions for Institutional Research, No. 131.* San Francisco: Jossey-Bass

IR Applications is an AIR refereed publication that publishes articles focused on the application of advanced and specialized methodologies. The articles address applying qualitative and quantitative techniques to the processes used to support higher education management.

Authors can submit contributions from various sources such as a Forum presentation or an individual article. The articles should be 10-15 double-spaced pages, and include an abstract and references. Reviewers will rate the quality of an article as well as indicate the appropriateness for the alternatives. For articles accepted for *IR Applications*, the author and reviewers may be asked for comments and considerations on the application of the methodologies the articles discuss.

Articles accepted for *IR Applications* will be published on the AIR Web site and will be available for download by AIR members as a PDF document. Because of the characteristics of Web-publishing, articles will be published upon availability providing members timely access to the material.

Please send manuscripts and/or inquiries regarding *IR Applications* to Dr. Gerald McLaughlin.