Comparison of Cluster Analysis Methodologies for Characterization of Classroom Observation Protocol for Undergraduate STEM (COPUS) Data

Kameryn Denaro
Brian Sato
Ashley Harlow
Andrea Aebersold
Mayank Verma

Working Paper #23-17

November 2023

Comparison of Cluster Analysis Methodologies for Characterization of Classroom Observation Protocol for Undergraduate STEM (COPUS) Data

Submission type: article

Characters: 51,870

Running title: COPUS Clustering

Corresponding Authors:

Kameryn Denaro, Teaching and Learning Research Center, University of California, Irvine, 92697

Mailing address: 3000 Anteater Instruction and Research, Irvine, CA 92697-4150

Email: kdenaro@uci.edu Phone: 949-824-1398 Fax:

Brian Sato, Department of Molecular Biology and Biochemistry & Division of Teaching Excellence and Innovation, University of California, Irvine, 92697

Mailing address: 2238 McGaugh Hall MC3900, Irvine, CA 92697

Email: bsato@uci.edu Phone: 949-824-0661 Fax: 949-824-8551

Other authors:

Ashley Harlow, School of Education, University of California, Irvine, 92697

Andrea Aebersold, Division of Teaching Excellence and Innovation, University of California, Irvine, 92697

Mayank Verma, Division of Teaching Excellence and Innovation, University of California, Irvine, 92697

Keywords: undergraduate, COPUS, teaching practices, active learning

Abstract

The Classroom Observation Protocol for Undergraduate STEM (COPUS) provides descriptive feedback to instructors by capturing student and instructor behaviors occurring in the classroom. Due to the increasing prevalence of COPUS data collection, it is important to recognize the means by which researchers determine whether groups of courses or instructors have unique classroom characteristics. One approach utilizes cluster analysis, highlighted by a recently developed tool, the COPUS Analyzer (Stains et al., 2018), that enables the characterization of COPUS data into one of seven clusters representing three groups of instructional styles (didactic, interactive, and student-centered). Here, we present evidence that a predictive cluster analysis tool may not be appropriate as a means to analyze COPUS data through the examination of a novel 250 course dataset. We perform a *de novo* cluster analysis and compare results to the COPUS Analyzer output, and in doing so identify a number of contrasting outcomes regarding course characterizations. Additionally, we present two ensemble clustering algorithms: (1) k-means and (2) partitioning around medoids (PAM). Both ensemble algorithms categorize our classroom observation data into one of two clusters, traditional lecture or active learning. Finally, we discuss implications of these findings for education research studies that leverage COPUS data.

Introduction

A national focus on implementing evidence-based teaching practices to improve the quality of Science, Technology, Engineering, and Mathematics (STEM) education has been promoted by, among others, the National Research Council (NRC, 2012), the President's Council of Advisors on Science and Technology (PCAST, 2012), and the Association of American Universities (AAU, 2019). These organizations highlight the benefits of active learning pedagogies (Chickering & Gamson, 1987; Freeman et al., 2014; Hake, 1998; Ong et al., 2011; Prince, 2004; Ruiz-Primo et al., 2001; Smith et al., 2005; Tomkin et al., 2019) as practices that improve learning for all students, particularly those from diverse backgrounds (Eddy & Hogan, 2014; Handelsman et al., 2004; Ong et al., 2011; Theobald et al., 2020).

Despite these findings, the implementation of evidence-based teaching practices is generally not occurring widespread in STEM classrooms (Smith et al., 2014; Stains et al., 2018). While professional development opportunities to train instructors in the use of these practices are numerous, there is often a disconnect between the instructor perception of their implementation of active learning pedagogies and what is actually occurring in the classroom (Derting et al., 2016; Ebert-May et al., 2011). Thus, there is value in classroom observation data that provide an objective way to identify what both the student and instructor are doing within a classroom (Smith et al., 2013; Smith et al., 2014). These observations give a more standardized assessment of the class compared to surveys, responses to which may be influenced by student and instructor

interpretation or bias. These data can then be used in the assessment of the effectiveness of instruction strategies.

Classroom Observation Data Collection and Analysis

A number of protocols have been developed over the past two decades to better describe what is occurring within a higher education classroom (Frey et. al., 2016; Owens et al., 2017; Reimer et al., 2016; Sawada et al., 2002). One of the most commonly utilized is the Classroom Observation Protocol for Undergraduate STEM (COPUS) (Akiha et al., 2017; Daher et al., 2018; Deligkaris & Chan, 2020; Jiang & Li, 2018; Liu et al., 2018; Ludwig & Prins, 2019; Lund et al., 2015, McVey et al., 2017; Reisner et al., 2020; Riddle et al., 2020; Smith et al., 2013; Stains et al., 2018; Tomkin et al., 2019; Weaver et al., 2015; Wieman & Gilbert, 2015; Wolyniak & Wick, 2019). COPUS consists of twenty-five distinct codes that classify instructor and student behaviors (see Table 1, taken from Smith et al., 2013) which are recorded in two-minute intervals by observers. COPUS does not require observers to make judgments regarding teaching quality, but rather categorizes classroom activities by “What the instructor is doing” and “What the students are doing”.

Due to the increasing prevalence of COPUS data collection and presentation in education research, it is important to consider how researchers analyze these data. The most common tactic is to present COPUS data in a descriptive form, highlighting particular codes of interest and often comparing the relative presence of these codes between two scenarios (Akiha et al., 2017; Jiang & Li, 2018; Kranzfelder et al., 2019;

Lewin et al., 2016; Liu et al., 2018; McVey et al., 2017; Reisner et al., 2020; Riddle, 2019; Smith et al., 2013; Solomon et al., 2018; Weaver et al., 2015). For example, Lewin et al. (2016) highlighted the frequency of the “Instructor Lecturing” code for classes that utilized clickers and those that did not. Akiha et al. (2017) examined the frequency of various codes across middle school, high school, and undergraduate courses and determined whether there were differences among classes at various education levels using the Kruskal-Wallis test. It is also possible to take this analysis a step further and incorporate multiple regression models to identify the impact of various course or instructor characteristics on the presence of specific classroom practices. For example, to assess the effectiveness of their professional development program, Tomkin et al. (2019) identified differences in the utilization of various COPUS codes between faculty who did and did not participate in the program using multiple linear regression models, poisson regression models, and zero-inflated poisson regression models with the individual codes serving as the outcome variables. A third technique used to analyze COPUS data is cluster analysis. Cluster analysis is a data mining technique which allows researchers to cluster a set of observations into similar (homogenous) groupings based on a set of features. This has been utilized by the Stains group (Lund et al., 2015; Stains et al., 2018), and enables researchers to characterize a particular course based on the entirety of the collected COPUS data and identify distinct patterns of classroom behaviors present across a data set. Additionally, cluster analysis is used when researchers are in the exploratory phase of their analysis (Kaufmann & Rousseeuw, 1990; Ng & Han, 1994) and allows for identifying groups of

observations when you do not have a particular response variable of interest (Fisher, 1958; Hartigan & Wong, 1979; Hastie et al., 2001; Kaufmann & Rousseeuw, 1987; MacQueen, 1967; Pollard, 1981).

As a product of their cluster analysis, Stains et al. (2018) generated a COPUS Analyzer tool based on an original dataset of 2,008 individual class periods collected from over 500 STEM instructors across at least 25 institutions in the United States. They note that the COPUS Analyzer (http://www.copusprofiles.org/) "automatically classifies classroom observations into specific instructional styles, called COPUS Profiles". Despite the ease of use of the COPUS Analyzer, we argue that this tool, or similar clustering systems developed locally by education researchers based on prior collected data sets, is not an appropriate means to evaluate and classify new COPUS data. Since cluster analysis is a statistical learning algorithm that uses an unsupervised learning technique (i.e. there is no outcome variable used in the analysis), clustering algorithms are meant to be descriptive, not predictive. In general, clustering algorithms are able to find locally optimal partitions and split the data into *k* clusters; new data incorporated into an existing data set often result in different clusters being identified, and thus clustering should not be used as a predictive tool (Ben-David et al., 2006; Fisher, 1958; Gareth et al., 2013; Hartigan, 1975; Hartigan & Wong, 1979; Hastie et al., 2001; Wong, 1979). Due to this nature of cluster analysis, utilizing an existing cluster analysis to predict the cluster that new COPUS data would fall into, could then potentially incorrectly cluster that data. Mischaracterization of COPUS data could then lead to a research team drawing flawed conclusions from that analysis.

Study Aims

In this paper, we aim to explore whether different methods of clustering COPUS data produce contrasting outcomes using a novel dataset from 250 unique courses. Specifically, we will address the following questions:

1. Do clustering results for our dataset vary when utilizing the COPUS Analyzer versus *de novo* cluster analysis guided by the parameters established by the Analyzer?
2. How do *de novo* clustering results differ when the COPUS data are transformed (i.e. combining the codes into a condensed set or using a subset of the COPUS codes) in the various ways presented in the literature prior to clustering?
3. How do *de novo* clustering results differ when utilizing k-means algorithms versus partitioning around medoids (PAM) algorithms?

Methods

Participants and Procedures

The COPUS data were collected across 250 courses during the fall (n = 70), winter (n = 85), and spring (n = 95) quarters during the 2018-19 academic year at a research-intensive university in the Western US. Observed courses were selected if they were the following: lecture courses (excluding lab sections, discussions, and seminar courses), undergraduate courses (graduate courses excluded), and courses held in rooms with capacity for 60 students or greater. Courses were spread across STEM and non-STEM disciplines (in this work, the traditional definition of STEM

excluding social sciences is utilized), and were taught by a variety of faculty position types (tenured and non-tenured, including research track and teaching track) who were or were not Active Learning Certified, which means the instructor completed an 8-week active learning professional development series offered by the study's institution. Descriptive information regarding the courses included in the study and the faculty instructing them can be found in Table 2. Summary statistics for the individual COPUS codes are in the supplemental materials (Table S1).

We documented classroom behaviors in 2-minute intervals throughout the duration of the class session using the 25 COPUS codes. For each class session, we created three different datasets as previously described; (1) we used the subset of codes as described in Stains et al. (2018), (2) we collapsed the 25 codes into 8 codes as described in Smith et al. (2014), and (3) we utilized all 25 COPUS codes (Smith et al., 2013). Descriptions of each can be found in Table 1.

We also identified the COPUS profiles for each classroom session as reported by the COPUS analyzer (http://www.copusprofiles.org/). The COPUS Analyzer provides COPUS profiles that fall into one of seven clusters representing three groups of instructional styles, which are characterized as didactic, interactive, and student-centered. The didactic instructional style represents classes where more than 80% of the class period included the "Instructor Lecturing" code. The interactive instructional style was characterized by course periods where instructors supplemented lecturing with other group activities or clicker questions with group work. The student-centered instructional style represents classes where even larger portions of the

course period were dedicated to group activities relative to the interactive instructional style.

Despite the fact the COPUS protocol was designed based on the observation of STEM courses, we felt that it was appropriate to include non-STEM observation data for a variety of reasons. One, because our dataset was restricted to large enrollment lecture courses, this eliminated the presence of course types (for example, lab courses) that are unique to STEM fields. Second, if a STEM lecture was inherently different from a non-STEM lecture, we would expect to see unique distributions of STEM-specific codes in our data set. We performed a 2-sample t-test for each of the 25 codes to test for a difference in the amount of time spent on a certain code for STEM and non-STEM classes and applied a Bonferroni correction to account for multiple testing setting $\alpha^* = \frac{0.05}{25} = 0.002$. We found that of the 25 codes, COPUS code usage for STEM and non-STEM courses differed for only 2 codes ("Student Individual Thinking/Problem Solving" and "Instructor Real-time Writing on the Board"). These data are presented in Table S2. Additionally, as it is not our goal to make pedagogical conclusions or recommendations regarding the specific courses present in our data set, but instead to utilize these data to make conclusions about methodologies for COPUS data analysis, we felt it was appropriate to include both STEM and non-STEM courses.

Data Collection Procedures

Each course included in the study was observed twice within a quarter. A team of 10 COPUS observers were trained by a single individual. This training involved the

description of the COPUS codes, hands on time with the Generalized Observation and Reflection Platform (GORP – UC Davis, gorp.ucdavis.edu) which was used to collect COPUS data, and presentation of lecture videos that observers used to practice collecting COPUS data. Trained observers then completed 2-3 classroom observations in pairs to ensure reliability between the two-raters of at least 90% and Cohen's Kappa above 0.85 for each pair.

Instructors were notified at the beginning of each academic term that they would be observed during two lecture periods. Dates were assigned based on observer availability without any prior knowledge regarding what would occur in that lecture period. Observations were re-scheduled only if the originally selected date was an exam day. Instructor and student codes were collected for each class period and then summarized as percent of two-minute intervals during which a given code was occurring. COPUS data for the two classroom observations for a given course were averaged prior to data analysis. This study was approved by the University of California, Irvine, Institutional Review Board as exempt (IRB 2018-4211).

Data Analysis

To characterize the types of instructional practices observed in our 250 course dataset, we performed a variety of cluster analyses and compared them to the COPUS profiles resulting from the COPUS Analyzer (http://www.copusprofiles.org/). To address research question 1, we compared the COPUS profiles to a *de novo* cluster analysis using the same restrictions established by Stains et al. (2018), including using the same subset of codes (group worksheet, group other, group clicker, student question, work

1-on-1, clicker question, teacher question, and lecture) and performing k-means clustering with k = 7 using a Fisher's Exact Test. To address research question 2, we performed three separate k-means algorithms; one on the Analyzer codes (group worksheet, group other, group clicker, student question, work 1-on-1, clicker question, teacher question, and lecture), one on the collapsed codes (instructor presenting, instructor guiding, instructor administration, instructor other, student receiving, students talking to the class, students working, and student other), and one on all 25 COPUS codes. We compared the COPUS profiles to the *de novo* ensemble of the three k -means algorithms using a Fisher's Exact Test. To address research question 3, we performed three separate PAM (Partitioning Around Medoids) algorithms; one on the Analyzer codes, one on the collapsed codes, and one on all 25 COPUS codes. We compared the *de novo* ensemble of the three k-means algorithms to the *de novo* ensemble of the three PAM algorithms using a Fisher's Exact Test.

k-Means Clustering

In order to partition the data into distinct groups where the observations within the subgroups are quite similar and the observations in different clusters are quite different, we used k-means clustering. This is a simple and elegant approach for partitioning a data set into k distinct, non-overlapping clusters (James et al., 2013). k -means clustering is an unsupervised statistical learning technique that does not require the data to have a response variable (Fisher, 1958; Hartigan & Wong 1979; MacQueen, 1967). Among all classroom observations, there is heterogeneity across the observations, and we used clustering to find distinct homogeneous subgroups among

the COPUS observations. Our dataset includes n = 250 classroom observations with p equal to the number of COPUS features we are considering. For example, using the collapsed codes we have p = 8 features (instructor guiding, instructor presenting, instructor administration, instructor other, student receiving, student talking, student working, and student other).

To specify the desired number of clusters, k, the *NbClust* package in R was used (Charrad et al., 2014). This R package determines the relevant number of clusters in a data set by performing 30 different indices (see Table S3 for a complete list) while varying the cluster size and distance measures. For further discussion of the indices, see Charrad et al. (2014). After determining the relevant number of clusters, the k -means algorithm will assign each observation to exactly one of the k clusters. k-means clustering was performed using the *stats* package in R (R Core Team, 2018). k-means clustering partitions the observations into k clusters such that the total within-cluster variation, summed over all k clusters, is as small as possible. That is, k-means clustering solves the following minimization problem:

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\},$$

where $C_1,\ldots,C_K$ denote sets containing the indices of the observations in each cluster, p is the number of features, and K is the number of clusters. The algorithm for k-means clustering is as follows: (1) randomly assign a number, from 1 to k, to each of the observations. These serve as initial cluster assignments for the observations. (2) Iterate

until the cluster assignments stop changing: (2a) for each of the k clusters, compute the cluster centroid. The k th cluster centroid is the vector of the p feature means for the observations in the k th cluster. (2b) Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance). We used 20 random starts for the k-means clustering algorithm, since it is suggested to have the number of random starts to be greater than 1 (Gareth et al., 2013).

Partitioning Around Medoids Clustering

Partitioning around medoids (PAM) is a more robust method to cluster data compared to the more commonly used k-means algorithm (Kaufman & Rousseeuw 1987; Kaufman & Rousseeuw, 1990; Ng & Han, 1994). The main difference between the k-means algorithm and the PAM algorithm is that in the PAM algorithm, a data point within the cluster defines the medoid, whereas in k-means, the cluster center is the average of all the data points. The algorithm follows the work of Conrad and Bailey (2015) and uses the *cluster* (Maechler et al., 2018) and *randomForest* (Liaw & Wiener, 2002) package in R. The PAM analysis proceeds as follows: (1) unsupervised Random Forests (RF) is used to generate a proximity matrix using the COPUS variables; (2) PAM uses the dissimilarity matrix (1-proximity) to cluster the observations. RF dissimilarity measures have been successfully used in several unsupervised learning tasks (Breiman & Cutler, 2003; Hastie et al., 2001; Liu et al., 2000; Shi & Horvath, 2006). RF is a modern statistical learning method that involves a collection or ensemble of classification trees. Each tree is grown based on a different bootstrap sample of the original data. For the RF, each tree votes for a class and the final prediction for each

observation is based on the majority rule. In unsupervised RF, synthetic classes are randomly generated, and the trees are grown. Despite the synthetic classes, similar samples end up in the same leaves due to the tree's branching process. The proximity of the samples can be measured, and the proximity matrix is constructed. In the second step of the PAM analysis, the clustering is found by assigning each observation to the nearest medoid with the goal of finding k representative objects which minimize the sum of the dissimilarities of the observations to their closest representative object (Maechler et al., 2018). To determine the relevant number of clusters, the Silhouette index was used (Rousseeuw, 1987).

Ensemble of Algorithms

Instead of relying on a single "best" clustering, we used an ensemble of algorithms applied to our dataset, including both k-means clustering ensemble of algorithms and a PAM clustering ensemble of algorithms. We utilized the ensemble method (Strehl and Ghosh, 2002), using the *NbClust* package in R to cluster our data using different subsets of the COPUS codes to run multiple clusterings and then combine the information of the individual algorithms. Use of the ensemble of algorithms gives us a robust cluster assignment, as our cluster assignment does not rely on a single choice of variables to input into the cluster, and the number of clusters does not rely on a single choice for determining the best number of clusters. For classification, an ensemble average will perform better than a single classifier (Moon et al., 2007). In the educational literature, there have been a handful of applications of ensemble algorithms (Beemer et al., 2018; Kotsiantis et al., 2010; Pardos et al., 2011).

The k-means ensemble and PAM ensemble are based on individual algorithms that relied on different transformations of the COPUS codes; (1) we used the subset of codes described in Stains et al. (2018), (2) we collapsed the 25 codes into 8 codes as described in Smith et al. (2014), and (3) we utilized all COPUS codes (Table 1). The final k-means clustering ensemble gives each of the three individual k-means algorithms a vote for the final cluster. The final PAM clustering ensemble gives each of the three individual PAM algorithms a vote for the final cluster.

Results

RQ1. Do clustering results for our data set vary when utilizing the COPUS Analyzer versus *de novo* cluster analysis guided by the parameters established by the Analyzer?

To characterize the types of instructional practices observed in our 250 course dataset, we performed a variety of cluster analyses. To start, we utilized the existing COPUS Analyzer created by Stains et al. (2018). We first ran our COPUS data through the COPUS Analyzer and compared these results to those obtained with a *de novo* cluster analysis using the same restrictions set out by the work by Stains et al. (2018), including the same subset of codes and performing k-means clustering with k = 7. These two means of clustering the COPUS data resulted in differing cluster patterns (Table 3) with only 36% agreement between the two sets of clusters. Sending our data through the COPUS Analyzer resulted in 42% of our classroom observations being labeled didactic, 39% interactive, and 19% student-centered. The *de novo* cluster analysis using our classroom observations gives a different breakdown of didactic

(57%), interactive lecture (21%), and student-centered lecture (23%). The similarities in the COPUS profiles and the *de novo* clustering varied by cluster. For example, 67% of the cluster 1 (didactic instructional style) observations were clustered together in the *de novo* clustering. On the other hand, for the 27% of our classroom observations that fell into cluster 3 (interactive instructional style) by the COPUS analyzer, those 67 observations were split into 5 different clusters and had at most 30% of the observations clustered together in the *de novo* clustering. And the observations falling under cluster 7 (student-centered instructional style) from the COPUS Analyzer were almost evenly split in the *de novo* clustering. The instability of the clustering algorithm can be seen from the very different results obtained when comparing the COPUS Analyzer and *de novo* clustering using the same clustering technique (k-means clustering), the same number of clusters (k = 7), and the same data (n = 250 classroom observations). Using a Fisher's Exact test for count data, we found that there was a significant difference in the clustering results from the analyzer and our *de novo* cluster analysis (p = 0.004).

RQ2. How do *de novo* clustering results differ when the COPUS data are transformed (i.e. combining the codes into a condensed set or using a subset of the COPUS codes) in the various ways presented in the literature prior to clustering?

We performed k-means clustering with the data transformed into the Analyzer codes (Stains et al., 2018), collapsed according to Smith et al. (2014), or left as the original 25 COPUS codes. In each case, the optimal number of clusters for our data was 2 (according to majority rule) (Table 4), as opposed to the seven identified from the

Stains work (2018) (Figure 1). 86% of our classroom observations had perfect agreement across the individual algorithms.

Cluster 1 can be characterized as a traditional lecture cluster; primarily driven by the “Instructor Presenting” and “Student Receiving” codes. Cluster 2 can be characterized as an active learning cluster; with greater usage of the “Student Other Group Work”, “Student Working in Groups”, and “Student Asking a Question” codes. Table 5 shows the comparison of the k-means ensemble and the COPUS Analyzer which shows a significant difference in the results of the two ensembles (Fisher’s Exact Test, $p < 0.001$). One interesting outcome is that the k-means ensemble is split on the COPUS Analyzer classification of “interactive” lectures (clusters 3 and 4) with the majority of cluster 3 from the Analyzer being designated as active learning classes while the majority of cluster 4 from the Analyzer being designated as traditional lecture.

RQ3. How do *de novo* clustering results differ when utilizing k-means algorithms versus partitioning around medoids (PAM) algorithms?

Another means to identify the most appropriate number of clusters for our dataset is the robust clustering algorithm partitioning around medoids (PAM). PAM also identified two as the optimal number of clusters (using both the Analyzer codes and all 25 codes, with similar traditional lecture and active learning profiles as previously identified from the k-means clustering). The cluster assignment for our data that arose from the three different individual algorithms (Analyzer codes, collapsed codes, and all codes) and the vote of the ensemble are presented in Table 6. 57% of our classroom observations had perfect agreement between the three individual algorithms.

The comparison of the PAM ensemble clustering and the k-means ensemble clustering is presented in Table 7. The vast majority of the classes which clustered as active learning from the k-means ensemble were also categorized as active learning under the PAM ensemble, whereas 53 of the traditional lecture classes from the k -means ensemble were also categorized as active learning under the PAM ensemble (20% of the total classroom observations). There is a difference in the two ensembles (Fisher's Exact Test, $p < 0.001$). Through the more robust PAM clustering, we were able to identify more classes that clustered in the active learning instruction profile.

Discussion

The increased push to improve undergraduate STEM education has led to greater interest in collecting independent (not from the student or instructor perspective) classroom data to describe what is occurring in the classroom, as evidenced by a number of recent COPUS-utilizing publications (Liu et al., 2018; Ludwig & Prins, 2019; Reisner et al., 2020; Stains et al., 2018). There are a variety of arenas in which COPUS data can be valuable, for supporting faculty merit and promotion cases (as suggested by Smith et al., 2013), illustrating the effectiveness of professional development activities, or to connect these data to other student or instructor outcomes for research purposes. Thus, it becomes increasingly important that we analyze such data in a rigorous manner following best practices established by other fields. Typical ways that COPUS data are presented in published literature include in a descriptive fashion to highlight the average presence of various codes among different instructor populations (Lewin et al., 2016; McVey et al., 2017; Akiha et al., 2017; Jiang & Li, 2018; Kranzfelder

et al., 2019; Liu et al., 2018; Reisner et al., 2020; Riddle, 2019; Smith et al., 2013; Solomon et al., 2018; Weaver et al., 2015), identification of particular course or instructor characteristics that may correlate with specific COPUS codes using regression analyses (Tomkin et al., 2019), and clustering of COPUS course profiles (Stains et al., 2018). The benefit of cluster analysis is that it allows researchers to take a deeper and more holistic look at the COPUS data rather than rely on drawing conclusions from select COPUS codes. Furthermore, cluster analysis can also be combined with the regression analyses used in works like Tomkin et al. (2019) to identify particular course or instructor characteristics that correlate with a course being found in a particular cluster. This would allow one to identify variables that correlate with a course being characterized as falling in an active learning cluster, for example. In future work, we would like to identify course level data (e.g. enrollment size, taught in an active learning versus traditional classroom space) and instructor level data (e.g. research versus teaching track, gender, active learning certification status) that is associated with distinct forms of classroom instruction.

Before discussing our findings, we recognize this work contains certain limitations. First, while our dataset consists of COPUS observations from 250 courses, these were collected at a single institution, which may represent course experiences that are unique to this setting. Second, as COPUS data collection is labor intensive, we are making general conclusions regarding a course based on data from only a fraction of the meeting periods, a limitation less prevalent for other classroom observation protocols (Owens et al., 2017). And third, our dataset includes observations from both

STEM and non-STEM courses, albeit all which were large lecture enrollments. While COPUS is intended for STEM courses, the fact that frequency of COPUS codes varied minimally between STEM and non-STEM courses (Table S2) leads us to believe the usage of this protocol in these settings is appropriate.

In this work, we used cluster analysis as a statistical learning algorithm to describe how our data is related across the COPUS codes. As clustering algorithms are not meant to be predictive, we suggest that researchers perform a *de novo* cluster analysis with each new data set collected, and when doing so, use an ensemble of clusters as the ensemble improves the accuracy over a single classifier (Moon et al., 2007). Clusters can change with new data, are affected if there are outliers in the data, and are dependent on the choice of variables included in the analysis. The information from different clusterings does not need to be thrown out; the cluster assignments from previous and current clusterings can be combined by methods presented in Strehl and Ghosh (2002) or using an ensemble combining the information from the different clustering as in this paper. We prefer using the PAM algorithm since COPUS data often have outliers. For our particular dataset, all COPUS codes had outliers with the exception of “Instructor Lecturing”.

Another approach we believe may be beneficial is latent class analysis (LCA) clustering techniques and mixture distribution models (Hagenaars & McCutcheon, 2009; Lubke & Luningham, 2017), which is a theory-driven approach as opposed to the distance-based approaches of this paper (PAM and k-means). It has been noted that LCA may be more appropriate to use versus PAM in cases where one’s data set has a

large number of variables, fewer clusters, larger sample sizes, and non-uniform cluster sizes (Anderlucci & Hennig, 2014). Numerous education research studies (Maull et al. 2010; Talavera & Gaudioso, 2004; Vermunt & Magidson, 2002; Xu, 2011) have compared LCA to k-means, concluding that the main advantages of LCA over k-means for traditional clustering are that LCA uses probability based modeling, utilizes the BIC statistic to calculate the best number of clusters, and does not require the user to standardize variables before the clustering process. Brusco et al. (2016) performed a simulation study of k-means, PAM, and LCA and found that both PAM and LCA outperform k-means. Pelaez and colleagues (2019) used LCA and a random forest ensemble to identify at-risk students in introductory Psychology courses; they found that they were able to discriminate between the most at-risk and least at-risk students by identifying characteristics that had a large difference between the clusters that could be related to the students' risk level. Because we may expect to see non-uniform cluster sizes and small numbers of clusters in our COPUS dataset, we would like to compare the PAM ensemble to LCA clustering in future work (Anderlucci & Hennig, 2014; Conrad & Bailey, 2015; Magidson & Vermunt, 2002).

In addition to its methodological implications, we feel this work also highlights the value of cross-disciplinary research. With the push to decrease silos often seen in discipline-based education research fields (Henderson et al., 2017; Reinholz and Andrews, 2019), and the rise of data science across many disciplines, STEM education researchers have an opportunity to leverage collaborations with statisticians and computer scientists to better understand educational data and identify new ways to

improve teaching and learning. Collaborations can be formed for specific research projects, but can also be expanded to create research teams aimed at viewing existing problems in the field through new lenses and to train the next generation of researchers to have expertise spanning multiple fields. In this particular instance, by broadening one's research team, it may be possible to answer novel questions using existing COPUS data or expand one's research design when embarking on a study that relies on classroom observation data.

Acknowledgements

The authors would like to thank the team of faculty (Shannon Alfaro and Paul Spencer) and students (Albert Bursalyan, Andrew Defante, Amy Do, Heather Echeverria, Samantha Gille, Emily May, Dominic Pyo, and Emily Xu) who collected the COPUS data as well as the vast array of faculty who allowed us into their classrooms to collect these data. This work was supported by the National Science Foundation (NSF DUE 1821724).

References

Achen, R.M. & Lumpkin, A. (2015). Evaluating Classroom Time through Systematic Analysis and Student Feedback. *Journal for the Scholarship of Teaching and Learning* 9 (2) ar4. https://doi.org/10.20429/ijsotl.2015.090204 PMID:25976654.

Akiha, K., Brigham, E., Couch, B.A., Lewin, J., Stains, M.,... & Smith, M.K. (2017), What Types of Instructional Shifts do Students Experience? Investigating Active Learning in STEM Classes across Key Transition Points from Middle School to the University Level. *Electronic Theses and Dissertations.* https://digitalcommons.library.umaine.edu/etd/2795

Anderlucci, L. & Hennig, C. (2014). The Clustering of Categorical Data: A Comparison of a Model-based and a Distance-based Approach, Communications in Statistics - Theory and Methods, 43:4, 704-721, DOI: 10.1080/03610926.2013.806665

Association of American Universities (AAU), (2019). Undergraduate STEM Education Initiative. https://www.aau.edu/education-community-impact/undergraduate-education/undergraduate-stem-education-initiative-3 accessed on July 18, 2019.

Beemer, J., Spoon, K., He, L., Fan, J. & Levine, R. A. (2018). Ensemble Learning for Estimating Individualized Treatment Effects in Student Success Studies. *International Journal of Artificial Intelligence in Education 28*, 315-335.

Ben-David, S., von Luxburg, U. & P´al, D. (2006). A sober look at clustering stability. In *Proceedings of the Conference on Computational Learning Theory*, pages 5–19.

Breiman, L. (2001), “Random Forests,” Machine Learning, 45, 5–32.

Breiman, L., & Cutler, A. (2003), “Random Forests Manual v4.0”, Technical report, UC Berkeley, available online at ftp://ftp.stat.berkeley.edu/pub/users/breiman/Using random forests v4.0.pdf.

Brusco, M.J., Shireman, E., & Steinley, D. (2016). A comparison of latent class, k-means, and k-median methods for clustering dichotomous data. *Psychological methods* 22 (3) 563.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). “NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set.” *Journal of Statistical Software* 61: 1–36. http://www.jstatsoft.org/v61/i06/paper.

Chi, M.T.H. & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist*, 49 (4), 219-243. DOI: 10.1080/00461520.2014.965823

Chickering, A.W. & Gamson, Z.F. (1987). Seven principles for good practice in undergraduate education. *American Association for Higher Education Bulletin*, 3, p.7.

Conrad D.J. & Bailey B.A. (2015). Multidimensional Clinical Phenotyping of an Adult Cystic Fibrosis Patient Population. *PLoS ONE* 10(3): e0122705. https://doi.org/10.1371/journal.pone.0122705

Crouch, C. H., & Mazur, E. (2001). Peer Instruction: Ten years of experience and results. *American Journal of Physics*, 69(9), 970-977.

Daher, T., Pérez, L. C., Babchuk, W. A., & Arthurs, L. A. (2018), Exploring Engineering Faculty Experiences with COPUS: Strategies for Improving Student Learning Paper presented at 2018 *ASEE Annual Conference & Exposition*, Salt Lake City, Utah. https://peer.asee.org/30486

Deligkaris, C. & Chan Hilton, A.B. (2020). COPUS: A non-evaluative classroom observation instrument for assessment of instructional practices. url = http://hdl.handle.net/20.500.12419/136

Derting, T.L., Ebert-May, D., Henkel, T.P., Maher, J.M., Arnold, B., Passmore, H.A. (2016). Assessing faculty professional development in STEM higher education: Sustainability of Outcomes. *Science Advances*, 2(3). https://doi.org/10.1126/sciadv.1501422.

Ebert-May, D., Derting, T.L., Hodder, J., Momsen, J.L., Long, T.M., & Jardeleza, S.E. (2011). What We Say is Not What We Do: Effective Evaluation of Faculty Professional Development Programs. *BioScience*, 61(7). https://doi.org/10.1525/bio.2011.61.7.9

Eddy, S.L. & Hogan, K.A. (2014). Getting under the hood: how and for whom does increasing course structure work?. *CBE—Life Sciences Education*, *13*(3), pp.453-468.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M.K., Okoroafor, N., Jordt, H., & Wenderoth, M.P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410-8415.

Frey, R.F., Fisher, B.A., Solomon, E.D., Leonard, D.A., Mutambuki, J.M.,...& Pondugula, S. (2016). A Visual Approach to Helping Instructors Integrate, Document, and Refine Active Learning. *Journal of College Science Teaching,* 45 (5).

Gareth, J., Witten, D., Hastie, T. & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R. New York: Springer.

Hake, R.R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *Am. Jour. Phys.* 66 (1): 64-74.

Handelsman, J., Ebert-May, D., Beichner, R., Bruns, P., Chang, A., … & Wood, W.B. (2004). Scientific Teaching. *Science,* 521-522..

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, New York: Springer.

Hartigan, J., & Wong, M. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* 28(1), 100-108. DOI:10.2307/2346830

Hastie, T., Tibshirani, R.,, Friedman, J. (2001). The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc..

Henderson, C., Connolly, M., Dolan, E., Finkelstein, N., Franklin, S., & John, K.S. (2017). Towards the STEM DBER Alliance: Why We Need a Discipline-Based STEM Education Research Community. Journal of Engineering Education, 106: 349-355. doi:10.1002/jee.20168

Jiang, Y. & Li, A.J. (2018). Observation and Analysis on Chinese and American College Classroom. 2018 *International Conference on Education Reform, Management and Applied Social Science.* ISBN: 978-1-60595-012-9. DOI 10.12783/dtssehs/ermas2018/26988

Kaufmann, L. & Rousseeuw, P. (1987). Clustering by Means of Medoids. Data Analysis based on the L1-Norm and Related Methods, 405-416.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley. ISBN: 978-0-47031680-1. DOI: 10.1002/9780470316801

Knight, J.K., & Wood, W.B. (2005). Teaching more by lecturing less. *CBE Life Sci Educ.* 4 (4):298-310.

Kotsiantis, S., Patriarcheas, K., & Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting student's performance in distance education. *Knowledge-Based Systems,* 23, 529-535, DOI: 10.1016/j.knosys.2010.03.010

Kranzfelder P., Bankers-Fulbright J.L., Garcı´a-Ojeda M.E., Melloy M., Mohammed S., & Warfa, A.-R.M. (2019) The Classroom Discourse Observation Protocol (CDOP): A quantitative method for characterizing teacher discourse moves in undergraduate STEM learning environments. *PLoS ONE* 14(7): e0219019. https://doi.org/ 10.1371/journal.pone.0219019

Lane, E.S. & Harris, S.E. (2015). A New Tool for Measuring Student Behavioral Engagement in Large University Classes. *Journal of College Science Teaching* 44 (6), 83–91. http://www.jstor.org/stable/43632000.

Laugger, S., Stewart, J., Tilghman, S.M., & Wood, W.B. (2004). Scientific Teaching. *Science* 304 (5670), 521-522. http://www.jstor.org/stable/3836701.

Liaw, A. & Wiener, M. (2002). Classification and Regression by randomForest. *R News* 2(3), 18-22.

Lewin J.D., Vinson E.L., Stetzer M.R., & Smith M.K. (2016). A Campus-Wide Investigation of Clicker Implementation: The Status of Peer Discussion in STEM Classes. *CBE Life Sci Educ.*, 15 (1): ar6. DOI: 10.1187/cbe.15-10-0224

Liu, B., Xia, Y., and Yu, P. (2000), “CLTree-Clustering Through Decision Tree Construction,” Technical report, IBM Research.

Liu, S.C., Lang, C.K., Merrill, B.A., Leos, A., Harlan, K. & Froyd, J. (2018). "Developing Emergent Codes for the Classroom Observation Protocol for Undergraduate STEM (COPUS)," *2018 IEEE Frontiers in Education Conference (FIE)*, San Jose, CA, USA, 2018, pp. 1-4.

Ludwig, P. M., & Prins, S. (2019). A Validated Novel Tool for Capturing Faculty-Student Joint Behaviors with the COPUS Instrument. *Journal of microbiology & biology education*, *20*(3), 20.3.55. https://doi.org/10.1128/jmbe.v20i3.1535

Lund, T.J., Pilarz, M., Velasco, J.B., Chakraverty, D., Rosploch, K., Undersander, M., & Stains, M. (2015). The Best of Both Worlds: Building on the COPUS and RTOP Observation Protocols to Easily and Reliably Measure Various Levels of Reformed Instructional Practice. *CBE-Life Sciences Education*, 14(2), ar18. https://doi.org/10.1187/cbe.14-10-0168arXiv:https://doi.org/10.1187/cbe.14-10-0168 PMID: 25976654.

Lund, T.J., & Stains, M. (2015). The importance of context: an exploration of factors influencing the adoption of student-centered teaching among chemistry, biology, and physics faculty. *International Journal of STEM Education*, 2, ar13. https://doi.org/10.1186/s40594-015-0026-8.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* Volume 1: Statistics, 281--297, University of California Press, Berkeley, Calif., 1967. https://projecteuclid.org/euclid.bsmsp/1200512992

Maciejewski, W. (2015). Flipping the calculus classroom: an evaluative study. Teaching Mathematics and its Applications: *An International Journal of the IMA* 35 (4), 187–201. https://doi.org/10.1093/teamat/hrv019 arXiv:http://oup.prod.sis.lan/teamat/article-pdf/35/4/187/8387911/hrv019.pdf.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2018). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.7-1.

Maull, K.E., Saldivar, M.G., & Sumner, T. (2010). Online curriculum planning behavior of teachers. *In Proceedings of the Third International Conference on Educational Data Mining.*

McVey, M. A., Bennett, C. R., Kim, J. H., & Self, A. (2017), Impact of Undergraduate Teaching Fellows Embedded in Key Undergraduate Engineering Courses Paper presented at 2017 *ASEE Annual Conference & Exposition*, Columbus, Ohio. https://peer.asee.org/28471

Meila, M. (2003). Comparing clusterings. *In Proceedings of the Conference on Computational Learning Theory,* pages 173–187.

Moon, H., Ahn, H., Kodell, R., Baek, S., Lin, C., & Chen, J. (2007). Ensemble methods for classification of patients for personalized medicine with high-dimensional data. *Artificial Intelligence in Medicine*, 197–207.

Ng, R.T., & Han, J. (1994). Efficient and Effective Clustering Methods for Spatial Data Mining. *VLDB*.

NRC (2012). Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering, Washington, DC: National Academies Press.

Ong, M., Wright, C., Espinosa, L., & Orfield, G. (2011). Inside the double bind: A synthesis of empirical research on undergraduate and graduate women of color in

science, technology, engineering, and mathematics. *Harvard Educational Review*, 81(2), 172-209.

Owens, M.T., Seidel, S.B., Wong, M., Bejines, T.E., Lietz, S., Perez, J.R., & Tanner, K.D. (2017). Classroom sound classifies teaching practices. *Proceedings of the National Academy of Sciences* Mar 2017, 114 (12) 3085-3090; DOI: 10.1073/pnas.1618693114

Pardos, Z.A., Gowda, S.M., Baker, R.S.J.D., & Heffernan, N.T. (2011). The sum is greater than the parts: ensembling models of student knowledge in educational software. *ACM SIGKDD Explorations*, 13(2).

Pelaez, K., Levine, R. A., Guarcello, M. A., and Fan, J. (2019). Latent Class Analysis and Random Forest Ensemble to Identify At-Risk Students in Higher Education. *Journal of Educational Data Mining 11*, 18-46.

Pollard, D. (1981). Strong consistency of k-means clustering. *The Annals of Statistics*, 9(1):135–140.

President's Council of Advisors on Science and Technology (PCAST), (2012). Report to the President: Engage to Excel: Producing One Million Additional College Graduates

with Degrees in Science, Technology, Engineering, and Mathematics. https://files.eric.ed.gov/fulltext/ED541511.pdf

Prince, M. (2004). Does active learning work? A review of the research, 93(3), 223-231. *Journal of Engineering Education.*

R Core Team (2018). R: A language and environment for statistical computing. *R Foundation for Statistical Computing,* Vienna, Austria. URL: https://www.R-project.org/.

Reimer, L.C., Nili, A., Nguyen, T., Warschauer, M., & Domina, T. (2016). Clickers in the Wild: A Campus-Wide Study of Student Response Systems in Weaver, G.C., Burgess, W.D., Childress, A.L., & Slakey, L., Transforming Institutions Undergraduate STEM Education for the 21st Century (383-398). West Lafayette, Indiana: Purdue University Press.

Reinholz, D.L. & Andrews, T.C. (2019). Breaking Down Silos Working Meeting: An Approach to Fostering Cross-Disciplinary STEM—DBER Collaborations through Working Meetings. CBE—Life Sciences Education 18:3. DOI: 10.1187/cbe.19-03-0064

Reisner, B.A., Pate, C.L., Kinkaid, M.M., Paunovic, D.M., Pratt, J.M., & Smith, S.R. (2020). I've Been Given COPUS (Classroom Observation Protocol for Undergraduate

STEM) Data on My Chemistry Class…Now What? *Journal of Chemical Education*, 97 (4), 1181–1189. DOI: 10.1021/acs.jchemed.9b01066

Riddle, E., Gier, E., & Williams, K. (2019). Utility of the Flipped Classroom When Teaching Clinical Nutrition Material. *J Acad Nutr Diet*. 2020;120(3):351-358. DOI: 10.1016/j.jand.2019.09.015

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics,* 20, DOI: 10.1016/0377-0427(87)90125-7

Ruiz-Primo, M. A., Briggs, D., Iverson, H., Talbot, R., and Shepard, L. A. (2011). Impact of undergraduate science course innovations on learning. *Science*, 331(6022), 1269–1270.

Sawada, D., Piburn, M.D., Judson, E., Turley, J., Falconer, K., Benford, R. and Bloom, I. (2002), Measuring Reform Practices in Science and Mathematics Classrooms: The Reformed Teaching Observation Protocol. School Science and Mathematics, 102: 245-253. doi:10.1111/j.1949-8594.2002.tb17883.x

Shi, T. & Horvath, S. (2006). Unsupervised Learning With Random Forest Predictors, *Journal of Computational and Graphical Statistics*, 15:1, 118-138, DOI: 10.1198/106186006X94072

Singer, S. & Smith, K.A. (2013). Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering. *Journal of Engineering Education*, 102:4, 468-471. https://doi.org/10.1002/jee.20030

Smith, K. A., Sheppard, S. D., Johnson, D. W., & Johnson, R. T. (2005). Pedagogies of engagement: Classroom-based practices. *Journal of Engineering Education*, 94(1), 87–100.

Smith, M.K., Jones, F.H.M., Gilbert, S.L., & Wieman, C.E. (2013). The Classroom Observation Protocol for Undergraduate STEM (COPUS): A New Instrument to Characterize University STEM Classroom Practices. *CBE-Life Sciences Education* 12 (4), 618–627. https://doi.org/10.1187/cbe.13-08-0154 arXiv:https://doi.org/10.1187/cbe.13-08-0154 PMID: 24297289.

Smith, M.K., Vinson, E.L., Smith, J.A., Lewin, J.D., & Stetzer, M.R. (2014). A Campus-Wide Study of STEM Courses: New Perspectives on Teaching Practices and Perceptions. *CBE-Life Sciences Education* 13 (4) 624–635.

https://doi.org/10.1187/cbe.14-06-0108 arXiv:https://doi.org/10.1187/cbe.14-06-0108 PMID: 25452485.

Solomon, E.D., Repice, M.D., Mutambuki, J.M., Leonard, D.A., Cohen, C.A.,... & Frey, R.F. (2018). A Mixed-Methods Investigation of Clicker Implementation Styles in STEM. *CBE—Life Sciences Education*, 17 (2). https://www.lifescied.org/doi/full/10.1187/cbe.17-08-0180.

Stains, M., Harshman, J., Barker, M. K., Chasteen, S.V., Cole, R., & Young, A.M. (2018). Anatomy of STEM teaching in North American universities. *Science* 359 (6383), 1468–1470. https://doi.org/10.1126/science.aap8892 arXiv:https://science.sciencemag.org/content/359/6383/1468.full.pdf.

Strehl, A. & Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583 – 617.

Talavera, L. & Gaudioso, E. (2004). Mining student data to characterize similar behavior groups in unstructured collaboration spaces. *In Workshop on Artificial Intelligence in CSCL*. 16th European Conference on Artificial Intelligence. 17–23.

Theobald, E.J., Hill, M.J., Tran, E., Agrawal, S., Arroyo, N., … & Freeman, S. (2020). Active learning narrows achievement gaps for underrepresented students in

undergraduate science, technology, engineering, and math. *Proceedings of the National Academy of Sciences*, 117 (12) 6476-6483; DOI: 10.1073/pnas.1916903117

Tomkin, J., Beilstein, S., Morphew, J. *et al.* (2019). Evidence that communities of practice are associated with active learning in large STEM lectures. *IJ STEM Ed* 6, 1 (2019). https://doi.org/10.1186/s40594-018-0154-z

UC Davis (2019, October 1). Generalized Observation and Reflection Platform. Retrieved from https://cee.ucdavis.edu/GORP.

Velasco, J.B., Knedeisen, A., Xue, D., Vickrey, T.L., Abebe, M., & Stains, M. (2016). Characterizing Instructional Practices in the Laboratory: The Laboratory Observation Protocol for Undergraduate STEM. *Journal of Chemical Education 93* (7), 1191-1203 DOI: 10.1021/acs.jchemed.6b00062.

Vermunt, J. & Magidson, J (2002). Latent Class Cluster Analyses. *Applied Latent Class Analysis*.

Walter D. Fisher (1958) On Grouping for Maximum Homogeneity, *Journal of the American Statistical Association,* 53:284, 789-798, DOI: 10.1080/01621459.1958.10501479

Weaver, G.C., Burgess, W.D., Childress, A.L. & Slakey, L. (2015). Transforming Institutions: Undergraduate STEM Education for the 21st Century. *Purdue University Press.* (Knowledge Unlatched Open Access Edition.)

Wieman, C.E. & Gilbert, S.L. (2014). The Teaching Practices Inventory: A New Tool for Characterizing College and University Teaching in Mathematics and Science. *CBE-Life Sciences Education* 13 (3), 552–569. https://doi.org/10.1187/cbe.14-02-0023 arXiv:https://doi.org/10.1187/cbe.14-02-0023 PMID: 25185237.

Wieman, C.E., & Gilbert, S.L. (2015). Taking a Scientific Approach to Science Education, Part II—Changing Teaching: Challenges remain before universities more widely adopt research-based approaches, despite their many benefits over lecture-based teaching. *Microbe Magazine, 10*, 203-207.

Wieman, C.E. (2015). A Better Way to Evaluate Undergraduate Teaching. *Change: The Magazine of Higher Learning* 47 (1), 6–15. https://doi.org/10.1080/00091383.2015.996077 arXiv:https://doi.org/10.1080/00091383.2015.996077.

Wieman, C.E. (2016). Forward in Weaver, G.C., Burgess, W.D., Childress, A.L., & Slakey, L. Transforming Institutions Undergraduate STEM Education for the 21st Century (pp. ix - xiv). West Lafayette, Indiana: Purdue University Press.

Wolyniak, M.J., Wick, S. (2019). Sustained mentorship promotes the development of active learning strategies in undergraduate biology classrooms: Evidence gained from the Promoting Active Learning and Mentoring (PALM) Network. *The FASEB Journal*, 33: (1).

Xu, B. (2011). Clustering educational digital library usage data: Comparisons of latent class analysis and k-means algorithms. Ph.D. Thesis, Utah State University.

Figure Legends

Figure 1

Radar plots highlighting the resulting clusters (cluster 1 – Fig 1A, traditional lecture and cluster 2--Fig 1B, active learning) from the 250-course COPUS dataset. Red lines indicated the average fraction of 2-minute intervals a given code was selected across the entire dataset. Green lines indicate the average fraction of 2-minute intervals a given code was selected only for the courses that fall within that cluster. For example in cluster 1, the “students receiving” code was selected for nearly 100% of the 2-minute intervals of the courses on average. The collapsed codes (Smith et al., 2014) were used to create these clusters. I. represents instructor behaviors while S. represents student behaviors.

Table 1. COPUS Code Description.

Descriptions of the individual codes in Smith et al. (2013), collapsed codes in Smith et al. (2014), and the Analyzer codes in Stains et al. (2018).

Table 2. Course and Instructor Characteristics of COPUS Dataset.

COPUS data were collected from 250 courses. Large enrollment size was defined as a course with greater than 100 students. STEM included science, engineering, math, and informatics/computer science courses. There were three classes of instructor based on their job titles. Active learning certification status was bestowed on faculty who completed an active learning instruction professional development series.

Table 3. COPUS Analyzer versus *De Novo* Clustering of Study Data.

A k-means algorithm with k = 7 was applied to our COPUS data and compared to the outcome of analyzing the same data using the COPUS Analyzer tool. The rows indicate the number of courses that clustered into the 7 categories of instruction as defined by the COPUS Analyzer. The columns represent the clustering of our data into 7 undefined categories from our k-means analysis. Similarities and differences in the clustering are indicated. For example, of the 77 courses that the COPUS Analyzer sorted into cluster 1, 51 also clustered together with the *de novo* clustering.

Table 4. k-means ensemble of algorithms applied to our dataset.

Using the COPUS codes selected by the COPUS Analyzer (Stains et al., 2018), the collapsed COPUS codes (Smith et al., 2014) or all 25 COPUS codes, the optimal number of clusters of our data was 2 (traditional and active). Each row illustrates the number of courses that were clustered into either cluster 1 or 2 based on the different code parameters. For example, 20 courses were sorted into cluster 1 using the Analyzer codes, 2 using the collapsed codes, and 1 using all codes. Perfect agreement of the algorithms is shown in bold. The percent indicates the percent of our sample that was found in each cluster.

Table 5. Comparison of COPUS Analyzer results versus k-means ensemble (k = 2). Courses are listed based on how they sorted using both the COPUS Analyzer and the *de novo* k-means ensemble. For example, 97 courses in our traditional lecture cluster were also found in the didactic cluster, but an additional 43 were found in the Analyzer's interactive cluster.

Table 6. PAM ensemble of algorithms applied to our dataset.

The optimal number of clusters was also 2 using this ensemble. Similar to Table 4, we indicate the number of courses that were clustered in a particular pattern using the Analyzer codes, collapsed codes, or all COPUS codes. Perfect agreement of the algorithms is shown in bold.

Table 7. Comparison of k-means versus PAM ensemble results.

The manner in which each particular course was clustered using either k-means or PAM ensembles are indicated. The k-means ensemble and PAM ensemble had perfect agreement in cluster assignment for 79% of the classroom observations (shown in bold).

Supplemental Material

Table S1. Summary Statistics of COPUS codes.

The mean, standard deviation, median, and interquartile range of the 25 COPUS codes in the 250 course dataset.

Table S2. t-test for Two Independent Samples (STEM and non-STEM) for each COPUS code.

A two sample t-test for each COPUS code by STEM categorization is given along with the mean and standard error for each of the groups. To account for multiple testing (25 COPUS codes), significance was based on a Bonferroni correction of 0.05/25 = 0.002.

Table S3. Clustering Indices.

A list of the 30 indices used to choose the best number of clusters in the NbClust package.

Table S4. Correlation Matrix of collapsed COPUS codes.

The correlation between the collapsed codes is given. Collapsed code descriptions are given in Smith et al. (2014) and in Table 1. I. represents instructor behaviors while S. represents student behaviors.

Table S5. Correlation Matrix of COPUS Analyzer codes.

The correlation between the COPUS Analyzer codes is given. Individual code abbreviations and descriptions can be found in Stains et al. 2018 and in Table 1.

Table S6. Correlation Matrix of all COPUS codes.

The correlation between all of the COPUS codes is given. Individual code abbreviations and descriptions can be found in Stains et al. 2018 and in Table 1.

Table S7. Best Number of Clusters when Clustering on All Codes, Analyzer Codes, and Collapsed Codes.

Using the COPUS codes selected by the COPUS Analyzer (Stains et al., 2018), the collapsed COPUS codes (Smith et al., 2014) or all 25 COPUS codes, the values of the 30 different indices for the optimal number of clusters of the k-means algorithm are given.

Table S8. Summary of the Optimal Number of Clusters.

The number of indices choosing the best number of clusters. Out of the 30 indices, 12 of those selected 2 as the optimal number of clusters.