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

Kameryn Denaro,[†]* Brian Sato,[‡§]* Ashley Harlow,[‖] Andrea Aebersold,[§] and Mayank Verma[§]

[†]Teaching and Learning Research Center, [‡]Department of Molecular Biology and Biochemistry, [§]Division of Teaching Excellence and Innovation, and [‖]School of Education, University of California, Irvine, CA 92697

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 how researchers determine whether groups of courses or instructors have unique classroom characteristics. One approach uses cluster analysis, highlighted by a recently developed tool, the COPUS Analyzer, 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 examine a novel 250 course data set and present evidence that a predictive cluster analysis tool may not be appropriate for analyzing COPUS data. We perform a de novo cluster analysis and compare results with the COPUS Analyzer output and identify several contrasting outcomes regarding course characterizations. Additionally, we present two ensemble clustering algorithms: 1) *k*-means and 2) partitioning around medoids. 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 (2012), the President's Council of Advisors on Science and Technology (2012), and the Association of American Universities (2019). These organizations highlight the benefits of active-learning pedagogies (Chickering and Gamson, 1987; Hake, 1998; Crouch and Mazur, 2001; Ruiz-Primo *et al.*, 2011; Prince, 2004; Knight and Wood, 2005; Maciejewski, 2015; Smith *et al.*, 2005; Ong *et al.*, 2011; Singer and Smith, 2013; Freeman *et al.*, 2014; Tomkin *et al.*, 2019) as practices that improve learning for all students, particularly those from diverse backgrounds (Handelsman *et al.*, 2004; Ong *et al.*, 2011; Eddy and Hogan, 2014; Theobald *et al.*, 2020).

Despite these findings, the implementation of evidence-based teaching practices is generally not 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 widely available, there is often a disconnect between instructor perception of implementation of active-learning pedagogies and what is actually occurring in the classroom (Ebert-May *et al.*, 2011; Derting *et al.*, 2016). 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, 2014; Wieman, 2016).

David Feldon, *Monitoring Editor*

Submitted Apr 24, 2020; Revised Oct 23, 2020; Accepted Nov 10, 2020

CBE Life Sci Educ March 1, 2021 20:ar3

DOI:10.1187/cbe.20-04-0077

*Address correspondence to: Kameryn Denaro or Brian Sato (kdenaro@uci.edu, bsato@uci.edu).

These observations give a more standardized assessment of the class compared with 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 and frameworks have been developed over the past two decades to better describe what is occurring within a higher education classroom (Sawada *et al.*, 2002; Chi and Wylie, 2014; Wieman and Gilbert, 2014; Wieman, 2015; Frey *et al.*, 2016; Reimer *et al.*, 2016; Owens *et al.*, 2017). One of the most commonly used protocols is the Classroom Observation Protocol for Undergraduate STEM (COPUS; Smith *et al.*, 2013; Lund *et al.*, 2015; Lund and Stains, 2015; Weaver *et al.*, 2015; Wieman and Gilbert, 2015; Velasco *et al.*, 2016; Akiha *et al.*, 2017; McVey *et al.*, 2017; Daher *et al.*, 2018; Jiang and Li, 2018; Liu *et al.*, 2018; Stains *et al.*, 2018; Ludwig and Prins, 2019; Tomkin *et al.*, 2019; Wolyniak and Wick, 2019; Deligkaris and Chan, 2020; Reisner *et al.*, 2020; Riddle *et al.*, 2020). COPUS consists of 25 distinct codes that classify instructor and student behaviors (see Table 1, taken from Smith *et al.*, 2013) that are recorded in 2-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 (Smith *et al.*, 2013; Weaver *et al.*, 2015; Lewin *et al.*, 2016; Akiha *et al.*, 2017; McVey *et al.*, 2017; Jiang and Li, 2018; Liu *et al.*, 2018; Solomon *et al.*, 2018; Kranzfelder *et al.*, 2019; Riddle *et al.*, 2020; Reisner *et al.*, 2020). For example, Lewin *et al.* (2016) highlighted the frequency of the Instructor Lecturing code for classes that used 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) used multiple linear regression models, Poisson regression models, and zero-inflated Poisson regression models with the individual codes serving as the outcome variables to identify differences in the use of various COPUS codes between faculty who did and did not participate in the program. A third technique used to analyze COPUS data is cluster analysis. Cluster analysis is a data-mining technique that allows researchers to cluster a set of observations into similar (homogeneous) groupings based on a set of features. This technique, which 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, has been used by the Stains group (Lund *et al.*, 2015; Stains *et al.*, 2018). Additionally, cluster analysis is used when researchers are in the exploratory phase of their analysis (Kaufman and Rousseeuw, 1990; Ng and Han, 1994) and allows for identification of groups of observations when you do not have a particular response variable of interest (Fisher, 1958; MacQueen, 1967; Hartigan and Wong, 1979; Pollard, 1981; Kaufman and Rousseeuw, 1987; Hastie *et al.*, 2001).

As a product of their cluster analysis, Stains *et al.* (2018) generated the COPUS Analyzer tool based on an original data set of 2008 individual class periods collected from more than 500 STEM instructors across 25 institutions in the United States. They note that the COPUS Analyzer (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. Because 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 (Fisher, 1958; Hartigan, 1975; Hartigan and Wong, 1979; Wong, 1979; Hastie *et al.*, 2001; Ben-David *et al.*, 2006; Gareth *et al.*, 2013). Due to this nature of cluster analysis, using 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 an analysis.

Study Aims

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

1. Do clustering results for our data set vary when using 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 before clustering?
3. How do de novo clustering results differ when using 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 2018–2019 academic year at a research-intensive university in the western United States. 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

TABLE 1. COPUS code description

	Observation	Description	All codes	Analyzer codes	Collapsed codes
Student codes	Listening	Listening to instructor/taking notes, etc.	Student.L	—	S.Receiving
	Answer question	Student answering a question posed by the instructor with rest of class listening	Student.AnQ	—	S.Talking
	Asking	Student asking question	Student.SQ	Student.SQ	S.Talking
	Whole class	Engaged in whole-class discussion by offering explanations, opinion, judgment, etc., to whole class, often facilitated by instructor	Student.WC	—	S.Talking
	Presentation	Presentation by student(s)	Student.SP	—	S.Talking
	Thinking	Individual thinking/problem solving: only marked when an instructor explicitly asks students to think about a clicker question or another question/problem on their own	Student.Ind	—	S.Working
	Clicker	Discuss clicker question in groups of two or more students	Student.CG	Student.CG	S.Working
	Worksheet	Working in groups on worksheet activity	Student.WG	Student.WG	S.Working
	Other group	Other assigned group activity, such as responding to instructor question	Student.OG	Student.OG	S.Working
	Prediction	Making a prediction about the outcome of demo or experiment	Student.Prd	—	S.Working
	Test/quiz	Test or quiz	Student.TQ	—	S.Working
	Waiting	Waiting (instructor late, working on fixing AV problems, instructor otherwise occupied, etc.)	Student.W	—	S.Other
	Other	Other: explained in comments	Student.Other	—	S.Other
Instructor codes	Lecturing	Lecturing (presenting content, deriving mathematical results, presenting a problem solution, etc.)	Instructor.Lec	Instructor.Lec	I.Presenting
	Writing	Real-time writing on board, doc. projector, etc. (often checked off along with Lec)	Instructor.RtW	—	I.Presenting
	Demo/video	Showing or conducting a demo, experiment, simulation, video, or animation	Instructor.DV	—	I.Presenting
	Follow-up	Follow-up/feedback on clicker question or activity to entire class	Instructor.FUp	—	I.Guiding
	Pose question	Posing non-clicker question to students (nonrhetorical)	Instructor.PQ	Instructor.PQ	I.Guiding
	Clicker question	Asking a clicker question (mark the entire time the instructor is using a clicker question, not just when first asked)	Instructor.CQ	Instructor.CQ	I.Guiding
	Answer question	Listening to and answering student questions with entire class listening	Instructor.AnQ	—	I.Guiding
	Moving/guiding	Moving through class guiding ongoing student work during active-learning task	Instructor.MG	—	I.Guiding
	One on one	One-on-one extended discussion with one or a few individuals, not paying attention to the rest of the class (can be along with MG or AnQ)	Instructor.1o1	Instructor.1o1	I.Guiding
	Administration	Administration (assign homework, return tests, etc.)	Instructor.Adm	—	I.Administration
	Waiting	Waiting when there is an opportunity for an instructor to be interacting with or observing/listening to student or group activities and the instructor is not doing so	Instructor.W	—	I.Other
	Other	Other: explained in comments	Instructor.Other	—	I.Other
		Total number of codes:	25	8	8

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).

with capacity for 60 students or more. Courses were spread across STEM and non-STEM disciplines (in this work, the traditional definition of STEM excluding social sciences is used) and were taught by faculty holding various positions (tenured and non-tenured, including research track and teaching track) who were or were not active-learning certified ("active-learning certified" 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. Course and instructor characteristics of COPUS data set[a]

Course/instructor characteristics	Percent of sample
Large enrollment (>100)	50
STEM course	58
Instructor gender (female)	46
Research tenure-track faculty	53
Teaching tenure-track faculty	18
Teaching non–tenure track faculty	28
Active-learning certified faculty	53

[a]COPUS data were collected from 250 courses. Large enrollment was defined as a course with more 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 2. Summary statistics for the individual COPUS codes are in Supplemental Table S1.

We documented classroom behaviors in 2-minute intervals throughout the duration of the class sessions using the 25 COPUS codes. For each class session, we created three different data sets as previously described: 1) we used the subset of codes as described in Stains *et al.* (2018), 2) we collapsed the 25 codes into eight codes as described in Smith *et al.* (2014), and 3) we used 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 (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 in which more than 80% of the class period included the Instructor Lecturing code. The interactive instructional style was characterized by course periods in which instructors supplemented lecturing with other group activities or clicker questions with group work. The student-centered instructional style encompasses classes in which even larger portions of the course period were dedicated to group activities relative to the interactive instructional style.

Even though 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. First, because our data set was restricted to large-enrollment lecture courses, this eliminated the presence of course types (e.g., 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 two-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 settings $\alpha^* = \frac{0.05}{25} = 0.002$. We found that, of the 25 codes, COPUS code usage for STEM and non-STEM courses differed for only two codes (Student Individual Thinking/Problem Solving and Instructor Real-Time Writing on the Board). These data are presented in Supplemental 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 use 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, University of California, Davis, 2019), which was used to collect COPUS data, and presentation of lecture videos that observers used to practice collecting COPUS data. Trained observers then completed two to three 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 rescheduled 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 2-minute intervals during which a given code was occurring. COPUS data for the two classroom observations for a given course were averaged before 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 data set, we performed a variety of cluster analyses and compared them with the COPUS profiles resulting from the COPUS Analyzer (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 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

To partition the data into distinct groups wherein 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 and 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 data set 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, we used the NbClust package in R (Charrad *et al.*, 2014). This R package determines the relevant number of clusters in a data set by performing 30 different indices (see Supplemental 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, performed using the stats package in R (R Core Team, 2018), 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:

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

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 kth cluster centroid is the vector of the p feature means for the observations in the kth 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, because it has been suggested that the number of random starts should be greater than 1 (Gareth *et al.*, 2013).

PAM Clustering

PAM is a more robust method to cluster data compared with the more commonly used k-means algorithm (Kaufman and Rousseeuw 1987, 1990; Ng and Han, 1994). The main difference between the k-means algorithm and the PAM algorithm is that a data point within the cluster defines the medoid in the PAM algorithm, whereas the cluster center is the average of all the data points in k-means. The algorithm follows the work of Conrad and Bailey (2015) and uses the cluster (Maechler *et al.*, 2018) and randomForest (Liaw and Wiener, 2002) packages in R. The PAM analysis proceeds as follows: 1) unsupervised Random Forests (RF) is used to generate a proximity matrix using the COPUS variables; and 2) PAM uses the dissimilarity matrix (1-proximity) to cluster the observations. RF dissimilarity measures have been successfully used in several unsupervised learning tasks (Liu *et al.*, 2000; Hastie *et al.*, 2001; Breiman, 2001; Breiman and Cutler, 2003; Shi and 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 that 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, we used the Silhouette index (Rousseeuw, 1987).

Ensemble of Algorithms

Instead of relying on a single "best" clustering, we used an ensemble of algorithms applied to our data set, including both k-means clustering ensemble of algorithms and a PAM clustering ensemble of algorithms. We applied 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). A handful of applications of ensemble algorithms can be found in the educational literature (Kotsiantis *et al.*, 2010; Pardos *et al.*, 2011; Beemer *et al.*, 2018).

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 eight codes as described in Smith *et al.* (2014), and 3) we used 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 Using 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 data set, we performed a de novo cluster analysis. To start, we used 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 in the work by Stains *et al.* (2018), including the same subset of codes and performing k-means clustering with

TABLE 3. COPUS Analyzer versus de novo clustering of study data[a]

		De novo *k*-means clustering							Total	Percent
	COPUS Analyzer	A	B	C	D	E	F	G		
Didactic	1	51	21	5	0	0	0	0	77	31
	2	12	5	1	5	4	0	0	27	11
Interactive	3	6	13	20	0	8	0	20	67	27
	4	10	6	3	6	6	0	0	31	12
Student centered	5	1	4	0	0	0	1	0	6	2
	6	1	4	2	0	0	5	0	12	5
	7	1	6	9	2	6	3	3	30	12
Total		82	59	40	13	24	9	23	250	
Percent		33%	24%	16%	5%	10%	4%	9%		

[a]*k*-means algorithm with $k = 7$ was applied to our COPUS data and compared with the outcome of analyzing the same data using the COPUS Analyzer tool. The rows indicate the number of courses that clustered into the seven categories of instruction as defined by the COPUS Analyzer. The columns represent the clustering of our data into seven 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.

$k = 7$. 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) as sorted by the COPUS analyzer, those 67 observations were split into five 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) with 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 ($p = 0.004$).

TABLE 4. *k*-Means ensemble of algorithms applied to our data set[a]

k-Means clustering ensemble					
Analyzer codes	Collapsed codes	All codes	Cluster vote	*n*	Percent
1	**1**	**1**	**1**	**138**	
1	2	1	1	20	65
2	1	1	1	3	
1	1	2	1	1	
2	**2**	**2**	**2**	**78**	
2	2	1	2	3	35
2	1	2	2	2	
1	2	2	2	5	

[a]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 two (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, two using the collapsed codes, and one 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.

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 before 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 two (according to majority rule; Table 4), as opposed to the seven identified from the Stains *et al.* (2018) work (Figure 1). Eighty-six percent 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, Students Working in Groups, and Student Asking a Question codes. Table 5 presents 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 and the majority of cluster 4 from the Analyzer being designated as traditional lecture.

RQ3. How Do de Novo Clustering Results Differ when Using *k*-Means Algorithms versus PAM Algorithms?

Another means to identify the most appropriate number of clusters for our data set is the robust clustering algorithm PAM. PAM also identified two as the optimal number of clusters (using both the Analyzer codes and all 25 codes, with similar

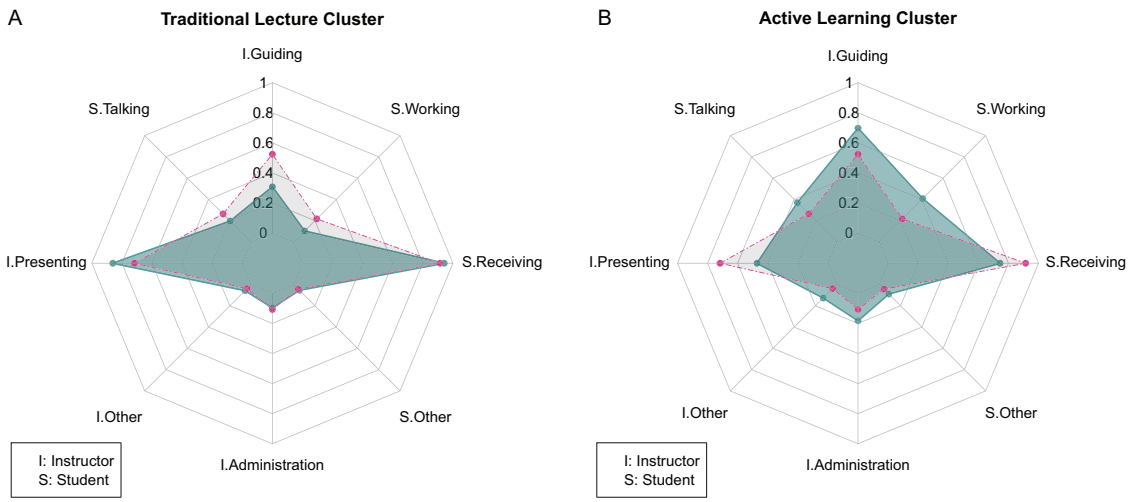

FIGURE 1. Radar plots highlighting the resulting clusters (A, cluster 1: traditional lecture; and B, cluster 2: active learning) from the 250 course COPUS data set. Red lines indicate the average fraction of 2-minute intervals a given code was selected across the entire data set. Green lines indicate 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, instructor behaviors; S, student behaviors.

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. Fifty-seven percent of our classroom observations had perfect agreement among 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 that 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.

TABLE 5. Comparison of COPUS Analyzer results versus *k*-means ensemble (*k* = 2)[a]

		k-Means cluster vote		
	COPUS Analyzer	Traditional 1	Active 2	Total
Didactic	1	74 (96%)	3 (4%)	77 (31%)
	2	23 (85%)	4 (15%)	27 (11%)
Interactive	3	22 (33%)	45 (67%)	67 (27%)
	4	21 (68%)	10 (32%)	31 (12%)
Student centered	5	5 (83%)	1 (17%)	6 (2%)
	6	6 (50%)	6 (50%)	12 (5%)
	7	11 (37%)	19 (63%)	30 (12%)
	Total	162 (65%)	88 (35%)	250

[a]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.

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-using publications (Liu *et al.*, 2018; Stains *et al.*, 2018; Ludwig and Prins, 2019; Reisner *et al.*, 2020). There are several arenas in which COPUS data can be valuable: for supporting faculty merit and promotion cases (as suggested by Smith *et al.*, 2013), for illustrating the effectiveness of professional development activities, or for connecting 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: descriptively, to highlight the average presence of various codes among different instructor populations (Smith *et al.*, 2013; Weaver *et al.*, 2015; Lewin *et al.*, 2016; Akiha *et al.*, 2017; McVey *et al.*, 2017; Jiang and Li, 2018;

TABLE 6. PAM ensemble of algorithms applied to our data set[a]

PAM clustering ensemble					
Analyzer codes	Collapsed codes	All codes	Cluster vote	n	Percent
1	**1**	**1**	**1**	**91**	
1	2	1	1	23	46
2	1	1	1	1	
2	**2**	**2**	**2**	**51**	
1	2	2	2	79	54
2	2	1	2	4	
2	1	2	2	1	

[a]The optimal number of clusters was also two 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[a]

			PAM cluster vote		
			Traditional	Active	
			1	2	Total
***k*-Means**	Traditional	1	**112 (45%)**	50 (20%)	162 (65%)
cluster vote	Active	2	3 (1%)	**85 (34%)**	88 (35%)
			115 (46%)	135 (54%)	250

[a]How each particular course clustered using either *k*-means or PAM ensembles is indicated. The *k*-means ensemble and PAM ensemble had perfect agreement in cluster assignment for 79% of the classroom observations (shown in bold).

Liu *et al.*, 2018; Solomon *et al.*, 2018; Kranzfelder *et al.*, 2019; Riddle *et al.*, 2020; Reisner *et al.*, 2020), to identify particular course or instruct or characteristics that may correlate with specific COPUS codes using regression analyses (Tomkin *et al.*, 2019), and to cluster 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 within 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 vs. traditional classroom space) and instructor-level data (e.g., research vs. teaching track, gender, active-learning certification status) that are associated with distinct forms of classroom instruction.

Before discussing our findings, we acknowledge that this work contains certain limitations. First, while our data set 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 data set includes observations from both STEM and non-STEM courses, albeit all of which were large-enrollment lectures. While COPUS is intended for STEM courses, the fact that frequency of COPUS codes varied minimally between STEM and non-STEM courses (Supplemental 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 are 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 by using an ensemble combining the information from the different clustering, as in this paper. We prefer using the PAM algorithm, as COPUS data often have outliers. For our particular data set, 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 and McCutcheon, 2002; Lubke and 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 nonuniform cluster sizes (Anderlucci and Hennig, 2014). Many education research studies (Vermunt and Magidson, 2002; Talavera and Gaudioso, 2004; Maull *et al.* 2010; Xu, 2011) have compared LCA with *k*-means, concluding that the main advantages of LCA over *k*-means for traditional clustering are that LCA uses probability-based modeling and 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 nonuniform cluster sizes and small numbers of clusters in our COPUS data set, we would like to compare the PAM ensemble to LCA clustering in future work (Vermunt and Magidson, 2002; Anderlucci and Hennig, 2014; Conrad and Bailey, 2015).

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 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.

ACKNOWLEDGMENTS

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.

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*, Retrieved September 1, 2018, from 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. https://doi.org/10.1080/03610926.2013.806665

Association of American Universities. (2019). *Undergraduate STEM Education Initiative*. Retrieved July 18, 2019, from www.aau.edu/education-community-impact/undergraduate-education/undergraduate-stem-education-initiative-3.

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 (pp. 5–19).

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

Breiman, L., & Cutler, A. (2003). *Random Forests Manual* v4.0 (Technical report). Berkeley: University of California, Berkeley. Retrieved October 1, 2019, from 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. Retrieved October 1, 2019, from 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. https://doi.org/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*, 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, UT). Retrieved October 1, 2019, from https://peer.asee.org/30486

Deligkaris, C., & Chan Hilton, A. B. (2020). *COPUS: A non-evaluative classroom observation instrument for assessment of instructional practices*. Retrieved October 1, 2019, from 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), e150142. 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), 550–558. 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), 453–468.

Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association*, *53*(284), 789–798. https://doi.org/10.1080/01621459.1958.10501479

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 USA*, *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*. (pp. 20–26) New York: Springer.

Hagenaars, J., & McCutcheon, A. (Eds.) (2002). Applied latent class analysis. Cambridge, UK: Cambridge University Press. https://doi:10.1017/CBO9780511499531.

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

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

Hartigan, J. A. (1975). Clustering algorithms. Hoboken, NJ: John Wiley & Sons, Inc.

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.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

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. https://doi.org/10.1002/jee.20168

James, G., Witten, D., Hastie, T. & Robert Tibshira, R. (2013). *An introduction to statistical learning: With applications in R*. New York: Springer.

Jiang, Y., & Li, A. J. (2018). Observation and analysis on Chinese and American college classroom. In *International conference on education reform, management and applied social science*. https://doi.org/10.12783/dtssehs/ermas2018/26988

Kaufman, L., & Rousseeuw, P. (1987). Clustering by means of medoids. *Data analysis based on the L1-norm and related methods* (pp. 405–416).

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction toclusteranalysis*. NewYork: Wiley. https://doi.org/10.1002/9780470316801

Knight, J. K., & Wood, W. B. (2005). Teaching more by lecturing less. *Cell Biology Education*, *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: https://doi.org/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. Retrieved October 1, 2019, from www.jstor.org/stable/43632000

Laugger, S., Stewart, J., Tilghman, S. M., & Wood, W. B. (2004). Scientific teaching. *Science*, *304*(5670), 521–522. Retrieved October 1, 2019, from www.jstor.org/stable/3836701

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 Sciences Education*, *15*(1), ar6. doi: https://doi.org/10.1187/cbe.15-10-0224

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

Liu, B., Xia, Y., & 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). In *2018 IEEE Frontiers in Education conference (FIE) San Jose, CA* (pp. 1–4).

Lubke, G. H., & Luningham, J. (2017). Fitting latent variable mixture models. *Behaviour Research and Therapy*, *98*, 91–102. https://doi.org /10.101 6/j.brat.2017.04.003

Ludwig, P. M., & Prins, S. (2019). A validated novel tool for capturing faculty-student joint behaviors with the COPUS instrument. *Journal of Microbiology and Biology Education*, *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.

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: https://oup.prod.sis.lan/teamat/article-pdf/35/4/187/8387911/hrv019.pdf.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, *Statistics* (pp. 281–297). Berkeley: University of California Press, Berkeley. Retrieved October 1, 2019, from https://projecteuclid.org/euclid.bsmsp/1200512992

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, Pittsburgh, PA*.

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, OH). Retrieved October 1, 2019, from https://peer.asee.org/28471

Meilă, M. (2003). Comparing Clusterings by the Variation of Information. In Schölkopf, B., & Warmuth, M. K. (Eds.), *Learning Theory and Kernel Machines. Lecture Notes in Computer Science* (vol. 2777). Heidelberg, Berlin: Springer

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.

National Research Council. (2012). *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering*, Washington, DC: National Academies Press.

Ng, R. T., & Han, J. (1994). Efficient and effective clustering methods for spatial data mining. VLDB.

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 USA*, *114*(12), 3085–3090. doi: https://doi.org/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., & 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. *Annals of Statistics*, *9*(1), 135–140.

President's Council of Advisors on Science and Technology. (2012). *Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics*. Retrieved October 1, 2019, from https://files.eric.ed.gov/fulltext/ED541511.pdf

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

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved October 1, 2019, from 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. (Eds.), *Transforming institutions: Undergraduate STEM education for the 21st century* (pp. 383–398). West Lafayette, IN: 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), mr3. doi: https://doi.org/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. https://doi.org/10.1021/acs.jchemed.9b01066

Riddle, E., Gier, E., & Williams, K. (2020). Utility of the flipped classroom when teaching clinical nutrition material. *Journal of the Academy of Nutrition and Dietetics*, *120*(3), 351–358. https://doi.org/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*, 53–65. doi: https://doi.org/10.1016/0377-0427(87)90125-7

Ruiz-Primo, M. A., Briggs, D., Iverson, H., Talbot, R., & 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., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The Reformed Teaching Observation Protocol. *School Science and Mathematics*, *102*, 245–253. https://doi.org/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: https://doi.org/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., Luo, J., & Frey, R. F. (2018). A mixed-methods investigation of clicker implementation styles in STEM. *CBE—Life Sciences Education*, *17*(2), ar30. https://doi.org/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* (pp. 17–23).

Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, N., Behling, S., & 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 USA*, *117*(12), 6476–6483. doi: https://doi.org/10.1073/pnas.1916903117

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

University of California, Davis (2019, October 1). *Generalized Observation and Reflection Platform*. Retrieved October 1, 2019, from https://ceeucdavis.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. https://doi.org/10.1021/acs.jchemed.6b00062.

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In Hagenaars, J., & McCutcheon, A. (Eds.), *Applied latent class analysis* (pp. 89–106). Cambridge University Press.

Weaver, G. C., Burgess, W. D., Childress, A. L., & Slakey, L. (Eds.). (2015). *Transforming institutions: Undergraduate STEM education for the 21st century*. (Knowledge Unlatched open access ed.). West Lafayette, IN: Purdue University Press.

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). Foreword. In Weaver, G.C., Burgess, W.D., Childress, A.L., & Slakey, L. (Eds.), *Transforming institutions: Undergraduate STEM education for the 21st century* (pp. ix–xiv). West Lafayette, IN: Purdue University Press.

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.

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. *FASEB Journal*, *33*(1), 1.

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