CBE Life Sciences Education
Vol. 14, 1–12, Summer 2015

Article

The Best of Both Worlds: Building on the COPUS and RTOP Observation Protocols to Easily and Reliably Measure Various Levels of Reformed Instructional Practice

Travis J. Lund,* Matthew Pilarz,† Jonathan B. Velasco,‡ Devasmita Chakraverty,§ Kaitlyn Rosploch,‡ Molly Undersander,‡ and Marilyne Stains‡

*Department of Natural Sciences, Oregon Institute of Technology, Klamath Falls, OR 97601; †Department of Chemistry and Biochemistry, Rowan University, Glassboro, NJ 08028; ‡Department of Chemistry, University of Nebraska–Lincoln, Lincoln, NE 68588; §Biology Education, IPN Leibniz Institute for Science and Mathematics Education, Kiel, Germany

Submitted October 12, 2014; Revised January 15, 2015; Accepted January 15, 2015
Monitoring Editor: Jennifer Momsen

Researchers, university administrators, and faculty members are increasingly interested in measuring and describing instructional practices provided in science, technology, engineering, and mathematics (STEM) courses at the college level. Specifically, there is keen interest in comparing instructional practices between courses, monitoring changes over time, and mapping observed practices to research-based teaching. While increasingly common observation protocols (Reformed Teaching Observation Protocol [RTOP] and Classroom Observation Protocol in Undergraduate STEM [COPUS]) at the postsecondary level help achieve some of these goals, they also suffer from weaknesses that limit their applicability. In this study, we leverage the strengths of these protocols to provide an easy method that enables the reliable and valid characterization of instructional practices. This method was developed empirically via a cluster analysis using observations of 269 individual class periods, corresponding to 73 different faculty members, 28 different research-intensive institutions, and various STEM disciplines. Ten clusters, called COPUS profiles, emerged from this analysis; they represent the most common types of instructional practices enacted in the classrooms observed for this study. RTOP scores were used to validate the alignment of the 10 COPUS profiles with reformed teaching. Herein, we present a detailed description of the cluster analysis method, the COPUS profiles, and the distribution of the COPUS profiles across various STEM courses at research-intensive universities.

INTRODUCTION

Instructional reforms in science, technology, engineering, and mathematics (STEM) courses at the college level have been intensifying in recent years. For example, there have been two major national initiatives since 2011: the Widening Implementation and Demonstration of Evidence-Based Reforms program from the National Science Foundation (NSF) and the STEM Education Initiative from the Association of American Universities. These reforms have been focused on broadening the adoption of evidence-based instructional practices by educating and training STEM faculty members in their implementation. Critical to the success of these initiatives is the ability to reliably measure and describe classroom instructional practices.

The American Association for the Advancement of Science (AAAS) convened a group of 60 faculty members, evaluators, researchers, and administrators involved in reform efforts at the higher education level to identify the set of tools available for measuring instructional practices (AAAS, 2012). These tools include surveys, interviews, classroom observations, and teaching portfolios. Observations have

CBE Life Sci Educ June 1, 2015 14:ar18
DOI:10.1187/cbe.14-10-0168
Address correspondence to: Marilyne Stains (mstains2@unl.edu).

become increasingly popular, since they provide the most direct and reliable measures of teaching practices (Kane *et al.*, 2002; Ebert-May *et al.*, 2011). They can provide a means to understand the adaptations faculty members make to evidence-based instructional practices, tailor faculty development, and evaluate instructional change due to reform efforts.

However, observations are significant only if they are accompanied by the use of observation protocols that provide valid and reliable data. From a practical perspective, observation protocols should provide meaningful feedback that faculty members can understand and act upon. From a research perspective, the protocols need to align with research on effective teaching and to have enough resolution to identify small but significant changes. In both contexts, observation protocols need to be easily implementable. The AAAS report identifies two types of protocols: holistic and segmented. Holistic protocols require coders to evaluate each item for the class period as a whole. The most commonly implemented holistic protocol at the college level has been the Reformed Teaching Observation Protocol (RTOP; Piburn *et al.*, 2000; Sawada *et al.*, 2002). Segmented protocols require the coders to evaluate each item over short periods of time (e.g., every 2 min). An increasingly popular segmented protocol for the college level is the Classroom Observation Protocol in Undergraduate STEM (COPUS; Smith *et al.*, 2013), which was built from the Teaching Dimensions Observation Protocol (TDOP; Hora and Ferrare, 2010, 2012).

The RTOP is an instrument that is grounded in the literature on inquiry-based teaching (Piburn *et al.*, 2000; Sawada *et al.*, 2002). Specifically, it measures the extent to which students are actively constructing knowledge. It consists of five subscales (Lesson Design and Implementation; Content: Propositional Pedagogic Knowledge; Content: Procedural Pedagogic Knowledge; Classroom Culture: Communicative Interactions; and Classroom Culture: Student/Teacher Relationships), containing five items each for a total of 25 items. Each item is evaluated through a five-point Likert scale. The scale is based on the extent to which the practice described in the item is present throughout the whole class period (1: never occurred; 5: very descriptive). The RTOP has been shown to have high interrater reliability within a single research team (Sawada *et al.*, 2002; Marshall *et al.*, 2011), although some studies have also found a high level of variation among RTOP coders within the same study (Amrein-Beardsley and Osborn Popp, 2011). The RTOP has three major shortcomings. First, the original protocol did not provide specific descriptions for each level of the Likert scale, which creates difficulties in interpreting intermediate scores and comparing RTOP scores between studies (Marshall *et al.*, 2011). For example, an RTOP score of 30 may reliably describe straight lecturing for one research team but may reliably describe lecturing with some student interaction for another research team. McConnell and colleagues recently developed a rubric to address this weakness, which also enabled them to achieve high interrater reliability (Budd *et al.*, 2013). The second shortcoming of the RTOP is the difficulty in interpreting RTOP scores. Specifically, the analysis of a class period via RTOP yields a number between 0 and 100, with the high end of the scale indicating that student-centered instructional practices were implemented for the majority of the class. Analysis of the five subscales among which the 25 items are distributed can offer meaning behind this number, but it does not provide a detailed description of the instructional practice and thus lacks resolution. Finally, the high end of the RTOP scale seems difficult to achieve in lecture-based environments. Indeed, studies implementing RTOP in STEM lecture courses in higher education have documented a limited number of lectures with RTOP scores greater than 70 (Piburn *et al.*, 2000; Ebert-May *et al.*, 2011; Budd *et al.*, 2013).

The COPUS (Smith *et al.*, 2013) addresses the second shortcoming by focusing on the behaviors of instructors and students on a small timescale. Specifically, observers identify from a list of 25 codes (12 and 13 codes for instructors and students, respectively) which behaviors took place within each 2-min time frame. Instructor behaviors include lecturing, asking questions, or writing on the board, while student behaviors include listening, working in groups, and answering questions (see Table 1 for codes). The COPUS thus provides a high resolution of the instructional practices enacted in the classroom. Moreover, it has been demonstrated to be easily implementable by various types of observers (e.g., K–12 teachers, researchers) and to provide high interrater reliability (Smith *et al.*, 2013). The analysis of a class period via COPUS yields two pie charts (one for student behaviors and one for instructor behaviors) describing the prevalence of each code. This prevalence is calculated by dividing the total number of 2-min time blocks in which a certain code was used by the total number of *codes* that were used. There

Table 1. Abbreviated definitions of COPUS codes

Student codes		Instructor codes	
AnQ-S	Student answering instructor's question	PQ	Posing nonrhetorical, nonclicker question
SQ	Student asking a question	AnQ-I	Answering student question
CG	Discuss CQ in groups	CQ	Asking a clicker question (CQ)
WG	Work on worksheet in groups	FUp	Follow-up on CQ or activity
OG	Other group activities	W-I	Instructor waiting
L	Listening to instructor	Lec	Lecturing
Ind	Individual thinking/problem solving	RtW	Real-time writing on board, etc.
Prd	Making a prediction about a demo, experiment	MG	Moving through class, guiding work
WC	Whole-class discussion	1o1	One-on-one extended discussion with student(s)
T/Q	Test or quiz	D/V	Showing/conducting a demo, experiment, etc.
SP	Student presentation	Adm	Administration
W-S	Students waiting	O-I	Other
O-S	Other		

are two shortcomings associated with this analysis. First, it is difficult to identify and compare instructional styles at that level of resolution (i.e., 25 codes). The developers of the COPUS recently addressed this limitation by combining codes into a smaller set of four categories for the instructor and the student behaviors (Smith *et al*., 2014).

Second, the pie chart analysis does not always provide independently consistent results when making comparisons between classes with certain parallel behavioral codes. For example, if one instructor lectures throughout most of a 50-min class (during 21 of the 25 2-min time blocks) and runs clicker questions during five of the 25 2-min time blocks, his or her pie chart will indicate 81% for lecture (21/26 codes) and 19% clicker questions (5/26 codes). However, if another instructor writes on the board *as* he or she is lecturing (with both lecturing and board-writing marked in 21 of the 25 2-min time blocks) and similarly runs clicker questions during five of the 25 2-min time blocks, his or her pie chart will indicate 45% lecture (21/47 codes), 45% board-writing, and only 11% clicker questions (5/47). This analysis neglects to capture the important fact that each of the instructors lectured and used clickers for the same percentage of *time periods*. If one calculates the prevalence of behavioral codes based on the percentage of time intervals they are marked in rather than on the percentage of total codes, both instructors are found to lecture during 84% of the 2-min time periods in their classes and to ask clicker questions during 20%, and, *in addition*, the second instructor is found to write on the board during 84% of his or her time periods. Thus, by reporting the prevalence of each behavioral code as the percentage of *time periods* in which it is observed, the same data are captured, but temporal information important for meaningful cross-classroom comparisons is also included.

Recognizing the strengths and weaknesses of these two protocols, we sought to identify typical sets of COPUS behaviors as a way to provide a better insight into RTOP scores and to facilitate comparisons of teaching practices between instructors and over time. Specifically, we intended to statistically characterize COPUS profiles that represent the typical instructional styles enacted in STEM courses at the college level and that identify the extent to which students are engaged in constructing their own knowledge.

METHODS

Study Context/Participants

The classroom video recordings used in this analysis were collected in the context of two related research studies. The first study is an evaluation of the Cottrell Scholars Collaborative New Faculty Workshop (CSC NFW), a national workshop designed to enhance the teaching knowledge and skills of newly hired chemistry faculty members (Council of Scientific Society Presidents and American Chemical Society, 2013; Baker *et al*., 2014). Participants in these workshops are chemistry faculty members entering their first or second years as assistant professors at research-intensive universities across the country. Classroom video recordings of workshop participants were collected during the Fall semester following faculty participation in the CSC NFW, which is offered during the summer. In addition, we collected classroom recordings from a control group of new chemistry assistant professors who had not attended the CSC NFW but were comparable in other characteristics (e.g., working at research-intensive institutions). Potential control faculty members were identified through exploration of chemistry departments' websites. Once identified, they were recruited by email.

The second study is an evaluation of a local workshop series targeting STEM faculty members at a single research-intensive university in the Midwest. This series of semester-long faculty workshops were designed to promote faculty awareness and adoption of a variety of evidence-based instructional practices. In most cases, classroom recordings were collected from workshop participants both before and after their participation in the teaching workshops. We also collected classroom video recordings from control faculty members on campus who had not attended the workshops. In total, we were able to collect classroom video recordings from a significant percentage of the faculty members in the departments of chemistry (41%), biology (37%), physics (17%), and mathematics (11%), and of several additional faculty members in various bioscience fields (e.g., biochemistry, plant pathology) and the school of engineering.

Data Collection

In both studies, video data were collected by recording whole class periods for 1 week (two to three sequential class periods, depending on the class schedule) in a course that each study participant was teaching. In many cases, we revisited faculty members in different semesters or different courses to collect additional sets of video recordings of two to three sequential class periods. In total, we visited the classes of 73 separate faculty members via 102 weeklong classroom visits, collecting video recordings of 269 individual class periods. These 73 faculty members represent 28 different research-intensive institutions across the United States. Because separate class recordings were collected from many faculty members before and after workshop participation and/or in very different courses, we did not cluster our recordings by faculty member. In addition, we noted that some instructors used widely varying techniques and strategies for managing a class period, even across the course of a single week of classes. That is, on Monday, Instructor A might teach exactly like Instructor B, who always lectures; yet on Friday, Instructor A might implement group work, using a strategy very similar to how Instructor C usually teaches. Because we were interested in describing this very range of instructional strategies, we did not cluster each weeklong set of two to three classroom recordings together but instead treated *each* of the class periods as an individual sample of how an instructor might structure a class period.

The 269 class period observations from these two studies encompass a wide variety of instructors, disciplines, course levels, class sizes, and instructional methods, as described in Table 2. All observations are of STEM faculty members at institutions in the United States categorized by the Carnegie classification system (Carnegie Foundation for the Advancement of Teaching, 2014) as having high or very high research activity.

Video Coding

To capture the variety of practices implemented in the class periods we visited, our research group first used the RTOP

Table 2. Number of observed class periods with indicated characteristics

Characteristics	Number of class periods	Number of faculty members[a]
Department		
Chemistry	134	39
Biology	80	17
Physics	16	5
Mathematics	11	4
Other	28	8
Course level		
Freshman undergraduate	88	20
Sophomore undergraduate	75	20
Upper-division undergraduate	50	19
Graduate	56	19
Class size		
1–25 students	69	25
26–50 students	48	15
51–100 students	53	14
101–150 students	34	11
>150 students	65	16
Classroom type		
Fixed seating	177	50
Nonfixed desks	45	16
Tables	47	17
Years of faculty experience		
0–1 prior years as faculty member	78	28
2–5 prior years as faculty member	40	11
6+ prior years as faculty member	151	36
Observation type		
Nonworkshop faculty	68	21
Workshop faculty		
Preworkshop	81	31
Postworkshop	120	45

[a]Some faculty members were recorded more than once ($N = 24$) and in different courses ($N = 4$) across a 2-yr period; therefore, the total number of faculty members per characteristic may be greater than the total number of individual faculty members involved in the study ($N = 73$).

instrument described earlier to code all of our classroom video recordings. After completing the initial training provided by the RTOP developers (Piburn *et al*., 2000), all of the video coders (three postdoctoral research assistants, one graduate student, and one assistant professor) coded a total of 20 videos using the descriptive rubric created by McConnell and colleagues (Budd *et al*., 2013), periodically discussing scoring discrepancies and making minor clarifications and modifications to the rubric where necessary for our context. After this training process, a preconsensus (i.e., independent) intraclass correlation coefficient of 0.849 ± 0.095 was achieved. We thereafter coded the data using the following system to ensure precise item-level data and good continued interrater reliability: Each video was always coded independently by two coders. If, on any video, the two coders' scores were more than 10 points apart, *or* if they disagreed by two or more points on more than two RTOP items, a third coder was assigned to also code the video. All three coders then discussed the video and reached a final consensus score for all 25 RTOP items. For the videos not requiring a consensus, the final scores were determined by averaging the independent coders' scores for each of the 25 items. Videos coded using this system achieved a very high level of preconsensus (i.e., independent coding) interrater reliability: the average of intraclass correlation coefficients for exact agreement among 10 different pairs of coders was 0.875 ± 0.085. This level of interrater reliability is above the one achieved in other studies using intraclass correlation coefficients (Ebert-May *et al*., 2011). We chose the intraclass correlation coefficient over the correlation coefficient, which has been typically used in prior RTOP studies (Sawada *et al*., 2002; Park *et al*., 2010; Budd *et al*., 2013), for two reasons (see Jones *et al*., 1983). First, the correlation coefficient describes the extent to which two raters rank observations in similar order but does not provide information about the level of agreement between the two raters. Therefore, a high correlation coefficient is not necessarily indicative of high interrater reliability. Second, we are interested in the reliability of the mean of the RTOP scores achieved by all raters rather than the reliability of each individual rater. To achieve these two goals and based on our rating system (the same subset of raters rated each video), we calculated the average measure intraclass correlation coefficient in SPSS by choosing a two-way random analysis of variance model with absolute agreement.

In a second, subsequent round of data analysis, our research group, which included two undergraduate students, one graduate student, three postdoctoral research assistants, and one assistant professor, used the COPUS instrument described earlier to code all of the classroom videos. We first completed a brief (2 h) training period as described by Smith *et al*. (2013). We then independently coded six of our videos and established an average Cohen's kappa score of 0.868 ± 0.084 for the set of student codes and an average Cohen's kappa score of 0.827 ± 0.072 for the set of instructor codes using all pairs of observers. This small set of videos helped us further refine our understanding of the COPUS codes. Finally, we coded 11 videos independently, establishing an average Cohen's kappa of 0.908 ± 0.045 for the set of student codes and an average Cohen's kappa of 0.852 ± 0.069 for the set of instructor codes. Each class recording was then coded by a single coder. These levels of interrater reliability are on par with those reported by the team who designed the COPUS (Smith *et al*., 2013). Cohen's kappa is preferred over percent agreement when coding categorical data, because it takes into account the possibility that two raters agree by chance (Jones *et al*., 1983).

All the videos were anonymized so that the coders did not know whether the videotaped instructor belonged to the pre- or postworkshop and treatment or control categories.

Data Analysis

To identify common COPUS profiles, we performed a cluster analysis of the COPUS codes across our 269 classroom recordings. COPUS code analysis was based on the percentage of each class's 2-min time segments in which they appeared, rather than the percentage of codes, for reasons provided in the *Introduction*.

The goal of a cluster analysis is to sort cases (here, individual class periods) into an arbitrary number of relatively homogenous clusters based on a given set of variables (here,

Table 3. Average percentage of 2-min intervals per class period containing each of the COPUS codes

Students			Instructor		
Codes	Average	SD	Codes	Average	SD
L	95%	10%	Lec	81%	20%
Ind	3%	8%	RtW	40%	37%
CG	7%	13%	FUp	14%	18%
WG	1%	7%	PQ	23%	18%
OG	4%	11%	CQ	9%	15%
GW	13%	18%	AnQ-I	13%	13%
AnQ-S	21%	17%	MG	3%	8%
SQ	11%	12%	1o1	3%	8%
WC	1%	5%	D/V	3%	6%
Prd	0%	2%	Adm	6%	6%
SP	0%	2%	W-I	12%	16%
T/Q	1%	5%	O-I	3%	7%
W-S	2%	4%			
O-S	2%	4%			

COPUS codes). The first step in a cluster analysis is the selection of the clustering variables. Although no firm consensus exists regarding minimum sample sizes for cluster analysis, some suggest (Mooi and Sarstedt, 2011) a minimum of 2^n cases, where n is the number of clustering variables. Given our sample size of 269 observations, this suggested that an approximate maximum of eight of our 25 COPUS codes should be used to perform the cluster analysis. Furthermore, the variables should be as informative and nonredundant as possible.

To identify the most redundant codes, we first measured the correlation of the 25 variables with one another using Pearson's r. Unsurprisingly, the instructor codes for posing and answering questions (PQ and AnQ-I) were very highly correlated ($r > 0.95$, $p < 0.01$), with the student codes for answering and posing questions (AnQ-S and SQ), respectively. In addition, the instructor code for waiting (W-I) was somewhat highly correlated ($r > 0.80$, $p < 0.01$) with the instructor codes for clicker question use and follow-up (CQ and FUp). Although follow-up was also highly correlated ($r = 0.806$, $p < 0.01$) with clicker question use, the use of student responses or group activity results to inform the subsequent classroom activities is a critical indicator of an active, student-centered instructor and was thus retained.

In selecting the student codes for cluster analysis, we noted that three student codes (clicker groups, CG; worksheet groups, WG; and other groups, OG) measure fine distinctions in what can clearly be considered group work. We thus labeled the time intervals that contained any of these codes with a new student code, group work (GW). Listening (L) was a student code that was coded in an average of 95 ±10% of all time segments and was thus deemed to be a relatively uninformative variable for the purpose of our cluster analysis (see Table 3). Similarly, all other student codes (Ind, WC, Prd, SP, T/Q, W, and O) were specific activities that occurred in an average of ≤3% of the 2-min time segments and were therefore less useful in the context of our cluster analysis. We were thus left with our three most useful student codes: student questions (SQ), students answering instructor questions (AnQ-S), and student group work (GW).

After removing the instructor codes for posing questions (PQ), answering student questions (AnQ-I), and waiting (W-I), due to their high correlation with other codes (see paragraph describing the identification of redundant codes), we noted that lecturing (L), writing on the board (RtW), follow-up (FUp), and the use of clicker questions (CQ) were the most prominent instructor behaviors (see Table 3). Of the remaining instructor behaviors, three are descriptive of an instructor's behavior during student group work: moving through class (MG), one-on-one discussions (1o1), and waiting (W-I). As noted earlier, W-I is highly redundant with both CQ and FUp. In addition, 1o1 is moderately well correlated ($r > 0.73$, $p < 0.01$) with both GW and MG; however, MG is not well correlated with any other codes, aside from 1o1. We thus elected to include MG as a unique measure of instructor behavior during student group work.

After selection of our eight most descriptive, nonredundant COPUS codes (GW, SQ, AnQ-S, Lec, RtW, FUp, CQ, and MG; see Table 4), the next step in our cluster analysis was the selection of a clustering method. Although good arguments could be made for a variety of clustering methods (Mooi and Sarstedt, 2011), we opted to use the *k*-means procedure, a nonhierarchical partitioning method, due to its high tolerance of outliers and irrelevant clustering variables, its emphasis on the minimization of variability within each cluster, and its usefulness as an exploratory tool due to the ability to fine-tune its solution to the number of clusters desired. In addition, our standardized (0–100%), continuous variables were already a good fit for the requirements of the *k*-means procedure.

It should be noted that the process of cluster analysis does not produce a single, objectively correct solution. That is, different solutions may be equally accurate or desirable, depending on the purposes of the clustering and particularly on the real-world relevance of the resulting clusters, i.e., "clustering is in the eye of the beholder" (Estivill-Castro, 2002). Moreover, our cluster analysis is necessarily limited to the observations present in our own data set. However,

Table 4. The eight COPUS codes used for the cluster analysis that lead to the 10 COPUS profiles

Student codes		Instructor codes	
AnQ-S	Student answering instructor's question	CQ	Asking a clicker question
SQ	Student asking a question	FUp	Follow-up on CQ or activity
GW[a]	Students working in group though various means (worksheet, clicker, others)	Lec	Lecturing
		RtW	Real-time writing on board, etc.
		MG	Moving through class, guiding work

[a]GW is not a code in the original set of 25 COPUS codes; it is a new code that groups the original COPUS codes WG, CG, and OG.

as noted in Table 2, our 269 observations include a variety of STEM disciplines, course levels, class sizes, and classroom settings. In addition, our pool of instructors ranges from new hires to experienced faculty members, with varying levels of contact with innovative teaching methods, including two separate faculty development workshops. This variability in our data increases the chances of capturing many of the most common broad instructional styles in STEM education at research-intensive institutions and improves the likelihood that a given class period in a STEM field would be categorizable in one of the COPUS clusters that our analysis identifies.

The *k*-means partitioning method requires the input of the desired number of clusters at the start of the clustering process. To gain a sense of the number of clusters that might be meaningful, we first performed several exploratory hierarchical clustering analyses on our data, which suggested that clustering our data into fewer than four groups or more than 20 would not lead to meaningful clusters. We thus explored the *k*-means output for clustering our 269 observations into four to 20 clusters. We examined each of the clustering solutions for homogeneity within clusters, diversity between clusters, and the real-world interpretability of each of the clusters. Although as few as four clusters and as many as 20 could indeed be interpreted to have some practical meaning in terms of teaching styles, we generally found that the clustering solutions with very few clusters included excessive variability within the clusters and that those with very large numbers of clusters yielded groups with unnecessarily nuanced distinctions between the teaching styles depicted in the clusters.

We also used RTOP scores as an independent measure of cluster homogeneity, since RTOP scores had not been used as a clustering variable. In addition, we explored the various clustering solutions produced when including a ninth variable (D/V), as few as six variables (excluding MG and/or RtW), and even including all 25 COPUS codes as variables. Once several of the most promising cluster solutions were identified, they were tested for robustness by repeatedly randomizing the list of observations and rerunning the *k*-means analysis, a process that accounts for the sensitivity of the clustering to the initial order of the cases (Mooi and Sarstedt, 2011). In the end, we selected a clustering solution that was statistically rigorous; that yielded homogenous, diverse, and meaningful clusters; and that was representative of the general patterns we observed repeatedly across the majority of the clustering solutions. The results of this cluster analysis will be described in the following sections.

RESULTS

Description of COPUS Profiles

Table 5 presents the output of the cluster analysis, including our label for each cluster, the number (*N*) of class periods that are found in each cluster, and the average percentage of 2-min time intervals in which each of the eight classroom behaviors is present in each cluster. We provide here an interpretation of these clustered behaviors.

First, several of our clusters can generally be described as a lecturing instructional style since the lecture code was selected on average for more than 80% of the 2-min intervals per class period. Lecture (with slides) is the simplest cluster, consisting primarily of the instructor lecturing, with occasional student questions or student answers to instructor questions. Similarly, lecture (at board) consists of heavy use of whiteboards, chalkboards, or document cameras to capture real-time writing while lecturing and is associated with a slightly higher percentage on average of student questions and answers. Finally, transitional lecture is still primarily characterized by instructor lecturing, but clicker questions and group work are beginning to be used by instructors at a noticeable rate.

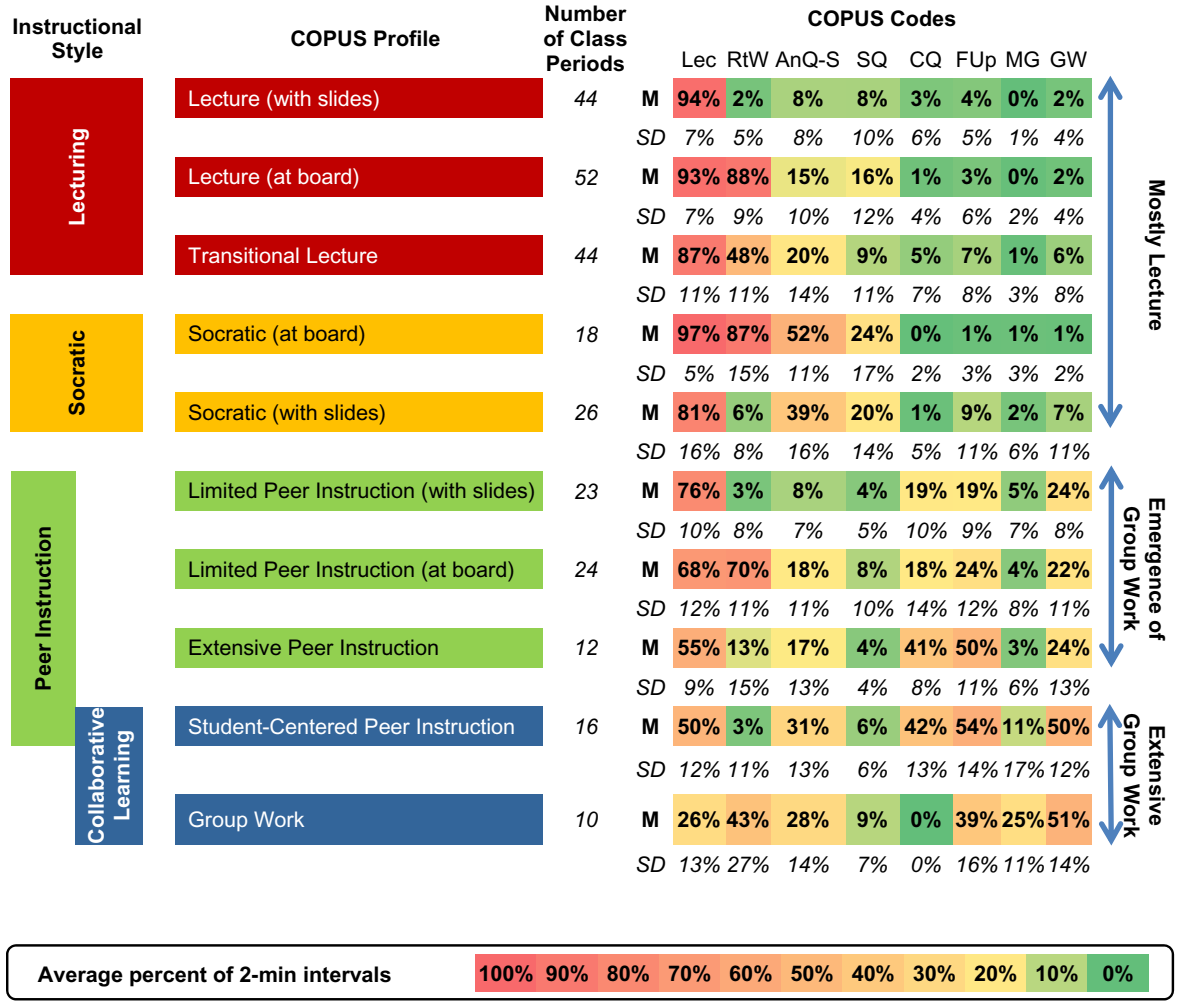
A second set of class periods can generally be described as a Socratic instructional style. These class periods are still characterized by a very high percentage of 2-min time intervals containing instructor lecturing (greater than 80%), yet a relatively high percentage also contain student questions and student answers to instructor questions, indicating regular, short instructor–student interactions. Obviously, the Socratic instructional style can be implemented at the board or with slides; use of slides to focus Socratic questioning is also associated with a slight increase in the use of student small-group breakout discussions and subsequent follow-up when compared with the Socratic at the board method, although these behaviors are still only present on average for 10% or fewer of the 2-min time intervals.

A third set of class periods clustered into what we labeled a peer instruction (PI) style of teaching. PI is a particular instructional strategy that prompts students to think and answer conceptual questions individually; this is followed by a discussion of their answers with their peers and an opportunity to provide their final answers back to the instructor (Mazur, 1997; Crouch and Mazur, 2001; Vickrey *et al.*, 2015). PI is commonly used with classroom response systems or voting cards. Although many of the class periods in these clusters display these PI strategies, it should be noted that many display PI-*like* patterns, without necessarily adhering strictly to PI best practices. A Kruskal-Wallis *H* test showed that there was a significantly higher proportion of group work and a significantly lower proportion of lecture in the PI set of clusters when compared with the Socratic set of clusters, $\chi^2(3269) = 173.118$, $p < 0.001$ and $\chi^2(3269) = 140.274$, $p < 0.001$, respectively.

We labeled the first two PI clusters as limited PI (with slides or at board). These two clusters are comparable with the first two lecturing clusters, except that group work occurs for about a quarter of the 2-min intervals compared with 2% in the two lecturing clusters; group work is often facilitated with the use of clickers, as the increase in the percentage of the CQ code indicates. We labeled the next cluster as extensive PI, since the percentages of time intervals in this cluster containing clicker questions and instructor follow-up (to clicker questions or group work) are twice those in the previous two PI clusters. Interestingly, the percentage of intervals containing group work is not higher than the previous PI clusters, suggesting a similar level of student–student engagement.

The final PI-based instructional style can be characterized as student-centered PI. Although the average percentage of time intervals containing lecturing, clicker questions, and instructor follow-up is not significantly different between this cluster and the extensive PI cluster, the frequency of group work and *students* answering questions doubled (50 and 31% of 2-min intervals, respectively); there is also the first prominent appearance of instructors moving among student groups. That is, more time is provided for students

Table 5. COPUS profile characteristics[a]



[a]The number of class periods contained in each COPUS profile is presented, followed by the average (M) of 2-min intervals per class period containing each of the eight COPUS codes used for the cluster analysis, with the SDs.

to explore the material and articulate their reasoning to one another and to the class. Although this cluster exhibits exemplary PI strategies, the high percentage of time dedicated to student–student interactions also places it firmly in our final general instructional style, collaborative learning. A Kruskal-Wallis H test showed that this instructional style has a significantly higher percentage of 2-min intervals for group work and a significantly lower percentage of 2-min time intervals devoted to lecture when compared with all three previous sets of clusters ($\chi^2(3269) = 173.118$, $p < 0.001$ and $\chi^2(3269) = 140.274$, $p < 0.001$, respectively). In addition to the student-centered PI cluster, the collaborative learning instructional style includes a group work cluster that involves various methods (such as handouts or questions posed via PowerPoint) to prompt group interactions. This final cluster is characterized by the lowest average percentage of lecturing (26% of the 2-min intervals) and the highest average percentage of group work and moving among students (51 and 25% of 2-min intervals, respectively); these instructors clearly planned for a student-centered, active-learning classroom. It should be noted that the cluster analysis actually outputs four distinct clusters with high percentages of group work, which we have combined into two categories for the sake of brevity and practical relevance. The student-centered PI cluster combines the class periods from a small cluster, including a high percentage of MG ($N = 4$) with those from a cluster without prominent MG behavior ($N = 12$). Percentages of all other behavioral codes are similar. Similarly, the group work cluster combines the class periods from a cluster that does not include writing on the board ($N = 2$) with those from a cluster that does ($N = 8$). Again, the percentages of all other behavioral codes are similar. Thus, for the purposes of our analysis, we considered these clusters to be similar enough in their instructional strategy to combine them.

Thus, our cluster analysis identified 10 specific instructional strategies (i.e., COPUS profiles) that represent four general instructional styles (lecturing, Socratic, peer instruction, and collaborative learning). In turn, these instructional styles

Table 6. Comparisons of categorization of RTOP scores in prior studies with the average RTOP scores of each COPUS profiles

Ebert-May *et al.* (2011)		COPUS profiles		Budd *et al.* (2013)	
Category	RTOP range	Profile: RTOP average (SD)		Category	RTOP range
Straight lecture	0–30	Lecture (at the board): 28 (5) Lecture (with slides): 29 (7) Transitional lecture: 33 (7)		Teacher centered	0–30
Lecture with some demonstration and minor student participation	31–45	Limited PI (with slides): 37 (8) Limited PI (at board): 42 (6) Extensive PI: 46 (5)	Socratic (at board): 34 (8) Socratic (with slides): 44 (10)	Transitional	31–49
Significant student engagement with some minds-on as well as hands-on involvement	46–60	Student-centered PI: 52 (8)	Group work: 50 (5)		
Active student participation in the critique as well as the carrying out of experiments	61–75			Student centered	50–100
Active student involvement in open-ended inquiry, resulting in alternative hypotheses, several explanations, and critical reflection	76–100				

represent statistically significant increments of student-centered instructional behaviors (e.g., GW; Table 5). The lecturing and Socratic styles are primarily characterized by a very high percentage of lecture. In the peer instruction instructional style, we see the emergence of group work (and related behaviors), accompanied by a respective decrease in lecture. Finally, in collaborative learning, we observe extensive group work and related behaviors (Table 5).

We can use the 10 COPUS profiles to characterize classroom practices in conjunction with the RTOP score. Table 6 presents a comparison of our COPUS profiles with the RTOP categories used by other authors. In particular, we feel that these COPUS profiles provide helpful insight into what a *transitional* RTOP classroom can look like, beginning with the transitional lecture cluster of lecture-focused classrooms that have begun to include some student-centered instructional strategies. It is important to note that our cluster analysis identifies two basic "tracks" for moving a lecture-based classroom toward greater student-centeredness: 1) through the use of PI to prompt student–student interactions or 2) through the use of heavy Socratic questioning and/or occasional "turn-to-your-neighbor" strategies.

In addition, it is interesting to note that the average RTOP score for each cluster is associated with a small but significant variation in RTOP scores. That is, within each of the 10 COPUS profiles, there can be variation in student-centeredness that the RTOP identifies via some of its items. Although many RTOP items measure student–student (RTOP items: 2, 16, 18, 19, 20, 23) and student–instructor (RTOP items: 5, 21, 25) interactions, other RTOP items (1 3, 4, 6–15, 17, 22, 24) measure aspects of the classroom not captured on the COPUS. Thus, just as the COPUS profiles provide insight into

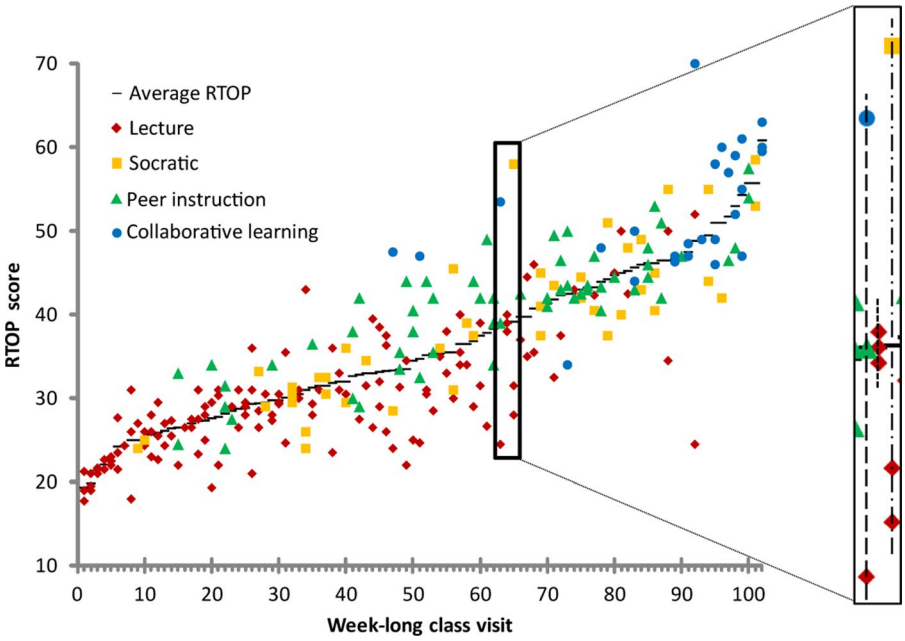


Figure 1. Distribution of RTOP scores and COPUS instructional styles across single weeks of instruction. In this figure, each data point represents one of the 269 class periods observed in this study. Each vertical stack of data points contains the two to three class periods (M/W/F or T/Th) observed during each of our 102 weeklong classroom visits. Scanning horizontally along different RTOP scores illustrates that the same RTOP score can often encompass different COPUS instructional styles. The inset, which is an enlargement of a small portion of the figure, presents the class period data from three of the weeklong classroom visits; the vertical variation in class period characteristics illustrates that the same instructor may (or may not) teach using very different instructional styles within the same week.

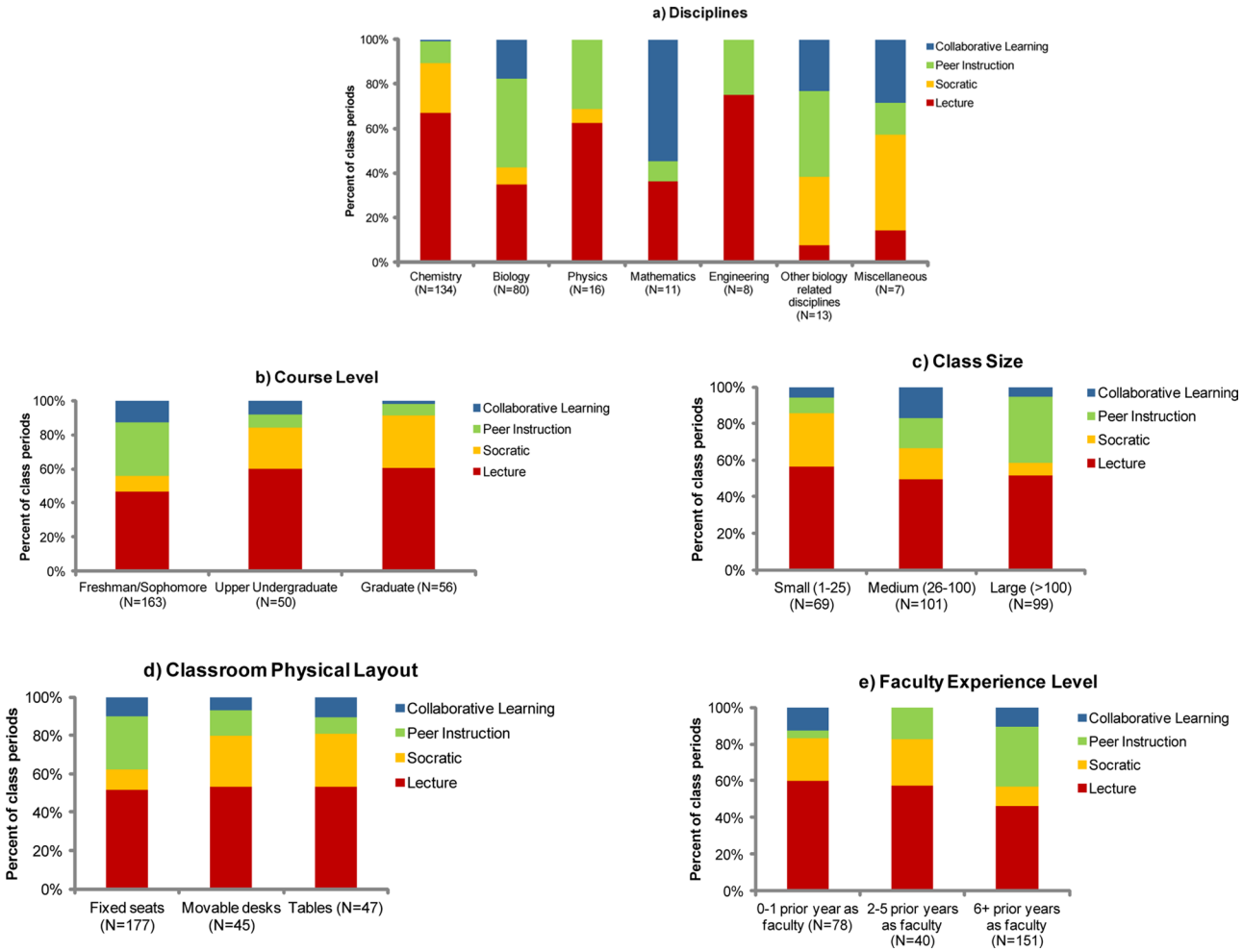


Figure 2. Distribution of the four instructional styles by (a) disciplines, (b) course level, (c) class size, (d) classroom physical layout, and (e) faculty teaching experience. *N* refers to the total number of class periods that fell into the specific category.

transitional RTOP scores, RTOP scores can measure variability in student-centeredness within the COPUS profiles. Figure 1 presents this orthogonality in the RTOP and COPUS data. Looking horizontally along RTOP scores, it is notable that a variety of instructional styles can achieve a similar RTOP score. These data also highlight the variations in instructional practices that faculty members employ. Looking vertically across a week's worth of recordings (M/W/F or T/Th), it is notable that some instructors teach using the same instructional style, while others may use two or even three classroom instructional strategies across the course of just a single week (see insert within Figure 1). This provides good evidence that at least two or three successive classroom visits are necessary to adequately characterize an instructor's classroom practices; it may be that additional visit could demonstrate additional instructional variability in some of our faculty members.

Representation of COPUS Profiles in STEM Courses at Research-Intensive Institutions

Approximately half of the 269 observed STEM class periods clustered into the lecturing instructional style, while around a quarter clustered as some form of peer instruction, around fifteen percent into Socratic, and roughly a tenth into collaborative learning. Figure 2 presents the characteristics of the class periods present in each of these four instructional styles.

Figure 2a highlights major differences in the distribution of instructional styles by STEM disciplines. Chemistry, physics, and engineering courses are most often taught through lecturing (notably, the engineering data only represent eight class periods). PI can be found in ~40% of the biology and biology-related class periods. Interestingly, half of the mathematics class periods were taught with collaborative learning; however, since the total number of periods in that discipline is small, these results should be taken with caution. Further studies using this methodology are required to understand differences in instructional practices between STEM disciplines.

Figure 2b presents the distribution of the four instructional styles across various course levels. A chi-square analysis indicates a statistically significant higher proportion of PI in freshman/sophomore courses compared with upper-undergraduate and graduate courses, which are dominated by lecturing and Socratic instructional styles; $\chi^2(6269) = 37.94$, $p < 0.001$, Cramer's $V = 0.0266$.

The number of students enrolled in a course (i.e., class size) and the physical layout of a classroom are often cited as barriers to the implementation of student-centered instructional practices (Gess-Newsome *et al.*, 2003; Henderson and Dancy, 2007; Hora, 2012). However, Figure 2c (class size) and 2d (physical layout) demonstrate that instructional styles that include various levels of student–student interactions can be implemented in large classes with amphitheater-style layouts. For example, 38% of the class periods in fixed-seat classrooms and 41% of the class periods with more than 100 students were taught through PI or collaborative learning. PI is significantly overrepresented in the large classes and underrepresented in the small classes; $\chi^2(6269) = 37.08$, $p < 0.001$, Cramer's $V = 0.263$. Similarly, this instructional style is significantly underrepresented in classroom with tables; $\chi^2(4269) = 18.61$, $p < 0.005$, Cramer's $V = 0.186$.

Finally, we explored differences in instructional styles by years of faculty teaching experience (Figure 2e). A chi-square analysis indicates a significantly higher proportion of PI among the most experienced faculty members; $\chi^2(6269) = 33.63$, $p < 0.001$, Cramer's $V = 0.250$. The proportion of lecturing decreases from 60% for faculty members with 0–1 yr of experience to 46% for faculty members with 6+ yr of experience. Faculty members thus seem to integrate instructional strategies involving student–student interactions as they gain teaching experience. This is consistent with results from prior studies published in physics and geosciences, which found that more experienced faculty members are in general interested in implementing evidence-based instructional strategies and achieve higher RTOP scores (Dancy and Henderson, 2010; Budd *et al.*, 2013).

DISCUSSION

The Identification of COPUS Profiles: An Efficient Method to Provide a Reliable and Valid Description of the Level of Reformed Teaching Enacted in STEM Courses

Within the current climate of instructional reform in STEM courses at the undergraduate level, there is a critical need to develop tools that easily but reliably measure instructional practices. Moreover, these tools need to reflect our current theoretical perspective on effective teaching. Several observation protocols have been developed that address these various criteria separately. In this study, we set out to leverage the two most promising observation protocols, the RTOP and the COPUS, in order to identify typical teaching practices in STEM courses that also reflect various levels of reformed teaching. We conducted a cluster analysis on eight COPUS codes over 269 individual class periods collected from a variety of STEM courses. This analysis led to the identification of 10 clusters, which we refer to as COPUS profiles. These profiles were then validated by comparing the RTOP scores of the class periods falling within each cluster. The resulting 10 COPUS profiles provide a fine-grained description of teaching styles ranging from more teacher centered to more student centered. The number and variety of STEM courses observed and analyzed to define these profiles is unprecedented. The COPUS profiles thus represent the best characterization of the typical instructional practices enacted in STEM courses at research-intensive institutions to date. The COPUS profiles will be further tested and refined by applying them to a new set of video recordings conducted in STEM courses at the college level. We will focus on selecting class periods with a high level of student engagement in order to better resolve the collaborative learning clusters.

Our study demonstrates that, in order to establish the instructional style in use in a given classroom, it may be sufficient to use only eight key codes of the original 25 provided in the COPUS protocol. Moreover, the COPUS profiles we have identified reflect incremental levels of student-centered instructional practices and thus provide a better resolution of reformed teaching than is currently available with the RTOP. Therefore, the strategy described in this paper drastically facilitates the analysis of classroom observations, while providing reliable and valid results.

To facilitate the process of categorizing classroom observations into the 10 COPUS profiles, we have constructed a rubric that summarizes several defining code cutoffs. This simple rubric successfully categorizes 87% of our own classrooms into the cluster they were placed in by the original *k*-means clustering process. Although outliers and unique class periods may not adhere to this rubric, and personal judgment should be used where necessary, we suspect that the majority of STEM class periods can be categorized into our clusters using these rules. We have integrated this rubric to enable instructors and researchers to easily convert their COPUS codes into one of the 10 COPUS profiles (www.copusprofiles.org).

From a professional development perspective, we envision that these profiles will provide meaningful feedback and guidance to faculty members who are interested in understanding and changing their own instructional practices. An observation protocol based on the eight COPUS codes can also be easily implemented as part of a peer-observation program. The profiles can also provide a tool for professional development staff to identify the specific needs of their particular population of faculty members. From a research perspective, these profiles can be used to characterize the state of instructional practices in various STEM disciplines and measure the extent of changes in instructional practices as a result of instructional reforms.

First Comprehensive Look at STEM Instructional Practices in Research-Intensive Institutions

Our analysis of 269 class periods collected from 73 faculty members representing 28 different research-intensive institutions provided new insights into the instructional practices of STEM faculty members at this type of institution.

First, we found that many faculty members employ different types of instructional practices within the course of a single week, demonstrating the need to observe faculty members for at least a week in order to adequately characterize their teaching styles.

Second, we found an increase in the level of student-centeredness of COPUS profiles as the teaching experience of the faculty members increases and as the level of the course decreases (from graduate to lower-level undergraduate). Interestingly, faculty members with six or more years of experience were much more likely to teach lower-level undergraduate courses than first- and second-year faculty members (71.5%

of the class periods observed for these experienced faculty members were at the lower level vs. 29.5% of the class periods observed for the first- and second-year faculty members). On the other hand, 43% of the class periods taught at the lower level by first- and second-year faculty members and 51% of the class periods taught at the lower level by the experienced faculty members belong to PI or collaborative learning styles; only 3 and 2% of the upper-undergraduate and graduate-level class periods taught, respectively, by first- and second-year faculty members and the experienced faculty members belong to these more student-centered instructional styles. These findings indicate that the classroom environment or curricula associated with the lower-level undergraduate courses may be perceived by faculty members to be better suited for the inclusion of student–student interactions than the upper-level undergraduate courses and graduate courses, regardless of the faculty member's level of teaching experience.

Finally, we found that fixed-seat classrooms and large-enrollment courses do not necessarily constitute barriers to the implementation of more student-centered instructional practices, despite the fact that faculty members often cite these contextual variables as constraints to engagement of students in peer discussion and group activities (Gess-Newsome *et al*., 2003; Henderson and Dancy, 2007; Hora, 2012). Interestingly, we found that only a small portion of the class periods we observed in the more ideal environments (movable desks, small class size) clustered into the more student-centered COPUS profiles. These data highlight that, while attention should be paid to physical (classroom) infrastructure, upgrading it will not automatically lead to uptake of student-centered instructional practices. An expensive new sports car is not useful if the driver does not know how to use a gearshift. Similarly, faculty members need proper training in student-centered instructional practices; otherwise, the expenditure on infrastructure will have minimal impact.

Taken together, these findings highlight the need for further research on the decision-making processes of STEM faculty members in all instructional contexts.

Limitations

The characterization of faculty instructional practices described in this study was solely based on observations made during regular lectures. However, faculty members can engage students in a meaningful manner outside lecture with activities such as workshops (Gafney and Varma-Nelson, 2008), homework (Novak, 1999; Simkins and Maler, 2009), or laboratory sessions. Future research endeavors should triangulate the data collected through the method developed in this study with other course-related data. This would provide a more accurate description of the manner by which and extent to which faculty members are helping their students construct an understanding of the subject matter.

The sample of classrooms used for this study provided a limited number of student-centered environments. Indeed, only 26 class periods fell into the collaborative learning instructional style. The average of the RTOP scores of the class periods falling under this category indicates a moderate level of student-centeredness. We thus had limited ability to discriminate student-centered instructional styles. To address this limitation, we plan on increasing the sample size of student-centered classrooms, which should provide a more complete resolution of the teacher-centered/student-centered continuum.

Finally, this study observed classroom practices at 28 different research-intensive universities across the country. Further research is needed to determine whether the trends observed in this study are similar at smaller or teaching-focused institutions. Similarly, our population consisted primarily of chemistry and biology faculty members; observations of a greater variety of STEM faculty members are needed.

CONCLUSION

This research study leveraged an unprecedented number of classroom observations to empirically identify an efficient method to measure and describe instructional practices in college STEM classrooms. Specifically, we demonstrated that, with only eight COPUS codes and without the RTOP, we can describe 10 different types of instructional practices (i.e., COPUS profiles) and map these practices on a scale from teacher to student centered. This method thus not only provides a detailed description of how STEM faculty members teach, it also aligns this description with our current theoretical understanding of effective instructional practices.

This method was used to characterize the instructional practices enacted by 73 STEM faculty members at 28 different research-intensive universities, corresponding to a variety of disciplines, courses, class size, and levels of faculty teaching experience. It was found that faculty members, regardless of teaching experience, are more likely to implement some student-centered instructional strategies in the freshman and sophomore undergraduate-level courses than in more advanced courses. Moreover, we found that providing the adequate classroom environment, in term of layout and class size, does not necessarily imply that faculty members will implement student-centered teaching.

Finally, we have constructed a simple rubric that can be used to categorize any class period into one of the COPUS profiles described here. This allows the results of our cluster analysis to be extended to new classroom observations without the need to rerun the cluster analysis to determine cluster membership of new data. We anticipate that this rubric could prove valuable both for professional development endeavors and as a research tool.

ACKNOWLEDGMENTS

We thank all 73 faculty members who willingly provided us access to their classrooms. This work was partly supported by NSF grant 1256003 and start-up funding from the Department of Chemistry at the University of Nebraska–Lincoln.

REFERENCES

American Association for the Advancement of Science (2012). Describing and Measuring Undergraduate STEM Teaching Practices. http://ccliconference.org/measuring-teaching-practices (accessed 31 March 2015).

Amrein-Beardsley A, Osborn Popp SE (2011). Peer observations among faculty in a college of education: investigating the summative and formative uses of the Reformed Teaching Observation Protocol (RTOP). Educ Assessment Eval Account *24*, 15–24.

Baker LA, Chakraverty D, Columbus L, Feig AL, Jenks WJ, Pilarz M, Stains M, Waterman R, Wesemann JL (2014). Cottrell Scholars Collaborative New Faculty Workshop: professional development for new chemistry faculty. J Chem Educ *91*, 1874–1881.

Budd DA, Kraft KJVDH, McConnell DA, Vislova T (2013). Characterizing teaching in introductory geology courses: measuring classroom practices. J Geosci Educ *61*, 461–475.

Carnegie Foundation for the Advancement of Teaching (2014). The Carnegie Classification of Institutions of Higher Education. http://carnegieclassifications.iu.edu/descriptions/basic.php (accessed 31 March 2015).

Council of Scientific Society Presidents and American Chemical Society (2013). The Role of Scientific Societies in STEM Faculty Workshops. http://www.aapt.org/Conferences/newfaculty/upload/STEM_REPORT-2.pdf (accessed 31 March 2015).

Crouch CH, Mazur E (2001). Peer Instruction: ten years of experience and results. Am J Phys *69*, 970.

Dancy M, Henderson C (2010). Pedagogical practices and instructional change of physics faculty. Am J Phys *78*, 1056.

Ebert-May D, Derting TL, Hodder J, Momsen JL, Long TM, Jardeleza SE (2011). What we say is not what we do: effective evaluation of faculty professional development programs. Bioscience *61*, 550–558.

Estivill-Castro V (2002). Why so many clustering algorithms—a position paper. SIGKDD Explorations *4*, 65–75.

Gafney L, Varma-Nelson P (2008). Peer-led team learning: evaluation, dissemination, and institutionalization of a college level initiative. In: Innovations in Science Education and Technology, ed. CK Cohen, New York: Springer, 156.

Gess-Newsome J, Southerland SA, Johnston A, Woodbury S (2003). Educational reform, personal practical theories, and dissatisfaction: the anatomy of change in college science teaching. Am Educ Res J *40*, 731–767.

Henderson C, Dancy M (2007). Barriers to the use of research-based instructional strategies: the influence of both individual and situational characteristics. Phys Rev Spec Top Phys Educ Res *3*, 20102.

Hora M (2012). Organizational factors and instructional decision-making: a cognitive perspective. Rev High Educ *35*, 207–235.

Hora M, Ferrare J (2010). The Teaching Dimensions Observation Protocol (TDOP), Madison: Wisconsin Center for Education Research, University of Wisconsin–Madison.

Hora M, Ferrare JJ (2012). Instructional systems of practice: a multidimensional analysis of math and science undergraduate course planning and classroom teaching. J Learn Sci *22*, 212–257.

Jones AP, Johnson LA, Butler MC, Main DS (1983). Apples and oranges: an empirical comparison of commonly used indices of interrater agreement. Acad Manag J *26*, 507–519.

Kane R, Sandretto S, Heath C (2002). Telling half the story: a critical review of research on the teaching beliefs and practices of university academics. Rev Educ Res *72*, 177–228.

Marshall JC, Smart J, Lotter C, Sirbu C (2011). Comparative analysis of two inquiry observational protocols: striving to better understand the quality of teacher-facilitated inquiry-based instruction. Sch Sci Math *111*, 306–315.

Mazur E (1997). Peer Instruction: A User's Manual, Prentice Hall Series in Educational Innovation, Upper Saddle River, NJ: Prentice Hall, xv, 253.

Mooi E, Sarstedt M (2011). A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics. Media, Heidelberg, Germany: Springer, 329.

Novak GM (1999). Just-in-Time Teaching: Blending Active Learning with Web Technology, Prentice Hall Series in Educational Innovation, Upper Saddle River, NJ: Prentice Hall, 188.

Park S, Jang JY, Chen YC, Jung J (2010). Is pedagogical content knowledge (PCK) necessary for reformed science teaching?: Evidence from an empirical study. Res Sci Educ *41*, 245–260.

Piburn M, Sawada D, Turley J, Falconer K, Benford R, Bloom I, Judson E (2000). Reformed Teaching Observation Protocol (RTOP) Reference Manual, Tempe: Arizona Collaborative for Excellence in the Preparation of Teachers.

Sawada D, Piburn M, Judson E, Turley J, Falconer K, Benford R, Bloom I (2002). Measuring reform practices in science and mathematics classrooms: the Reformed Teaching Observation Protocol. Sch Sci Math *102*, 245–253.

Simkins S, Maler M (2009). Just-in-Time Teaching: Across the Disciplines, and Across the Academy, Sterling, VA: Stylus, 224.

Smith MK, Jones FHM, Gilbert SL, Wieman C (2013). The Classroom Observation Protocol for Undergraduate STEM (COPUS): a new instrument to characterize university STEM classroom practices. CBE Life Sci Educ *12*, 618–627.

Smith MK, Vinson EL, Smith JA, Lewin JD, Stetzerτ MR (2014). A campus-wide study of STEM courses: new perspectives on teaching practices and perceptions. CBE Life Sci Educ *13*, 624–635.

Vickrey T, Rosploch K, Rahmanian R, Pilarz M, Stains M (2015). Research-based implementation of peer instruction: a literature review. CBE Life Sci Educ *14*, es3.