SURFING THE SEMESTER: A STUDY OF THE FLOW OF ACTIVE LEARNING IMPLEMENTATION

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Rebecca Sue Davidson Reichenbach

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Biological Sciences

October 2017

Fargo, North Dakota

North Dakota State University

Graduate School

Title

SURFING THE SEMESTER: A STUDY OF THE FLOW OF ACTIVE LEARNING IMPLEMENTATION

By

Rebecca Sue Davidson Reichenbach

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Lisa Montplaisir

Chair

Dr. Julia Bowsher

Dr. James Nyachwaya

Approved:

11/13/17	Dr. Wendy Reed
Date	Department Chair

ABSTRACT

Instructional reforms have been called for on a national level. Little data exists as to how changes take place. This study explored the implementation of active learning practices by non-tenure and tenure track faculty over the course of a semester. Faculty were introduced to evidence based pedagogy through workshops and faculty learning communities. Their instructional practices within a semester were tracked through observations conducted using the Classroom Observation Protocol for Undergraduate STEM (COPUS). Interviews were conducted to gain insight into reasons for instructional trends. A general trend downward was observed through the semester but was not found to be statistically significant at different time-points. A possible Simpson's paradox was detected in the collapsed COPUS categories of Instructor Guiding (G) and Students Working (SW) that may also be interfering with broad interpretation. Recommendations are made for further data collection to increase power and to use care before interpreting collapsed COPUS categories.

ACKNOWLEDGMENTS

Thanks are extended to Dr. Jared Ladbury, Department of Psychology, NDSU; Dr. James Grier, Department of Biological Sciences, NDSU; Dr. Mark Hanson, Associate Director of the Office of Institutional Research and Analysis, NDSU; and Melody McConnell, fellow graduate student, Monplaisir Lab Group, Department of Biological Sciences, NDSU, for their assistance during the preparation of this manuscript.

This study was conducted with funding provided by NSF DUE 1525056.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

INTRODUCTION

It is generally agreed that reform needs to occur in the American education system at all levels (American Association for the Advancement of Science, 1994; American Association for the Advancement of Science, 2010; Olson & Riordan, 2012). The most efficacious of the reformed instructional methodologies fall under the umbrella of "active learning" (Freeman et al., 2014). Active learning practices are usually defined as ones that cause the learner to engage with the material more deeply instead of traditional listening and rote memorization (Budd, van der Hoeven Kraft, McConnell, & Vislova, 2013). However, gaps still exist in our knowledge of what occurs in classrooms. Amount and quality of implementation of reformed instructional practices at universities is currently unknown (Budd et al., 2013). Researchers must have metrics by which to measure progress towards reformed instruction (Olson & Riordan, 2012).

One such metric is use of observation protocols (Couch, Brown, Schelpat, Graham, & Knight, 2015). Such protocols reduce the likelihood of instructors misidentifying their level of implementation of reformed instruction (Ebert-May et al., 2011) or adjusting responses to be more socially desirable (Hamilton, et al., 2003) via self-report. Several observation protocols have been developed in an attempt to provide objective measures of reformed instruction. Examples are the Reformed Teaching Observation Protocol (Piburn & Sawada, 2000; Sawada et al., 2002), the Teaching Dimension Observation Protocol (Hora & Ferrare, 2010), and the Classroom Observation Protocol for Undergraduate STEM (Smith, Jones, Gilbert, & Wieman, 2013). All of these provide a vehicle for snapshotting what occurs in classrooms but the Classroom Observation Protocol for Undergraduate STEM (COPUS) is generally deemed the most objective as its use requires fewer judgment calls and less content specific expertise on the part of the observer (Smith, Vinson, Smith, Lewin, & Stetzer, 2014). Paired observers using

COPUS can be ready to enter classrooms after just two hours of training (Smith et al., 2013). These factors make COPUS a very attractive tool for researchers and is the reason it was chosen as a tool for the study described in this paper.

Like any tool, observation protocols need to be used correctly. No evidence-based practices or suggestions exist for their implementation beyond general training. This is similar to being taught how to use a hammer but not knowing how to tell if the nail is sufficiently sunk. In other words, the use of the tool is known but not how often we should use it when gauging level of reform. Could the picture developed by researchers be missing details of instructors moving along a slower than anticipated slope towards reform? Could the subtleties of change be masked by assumptions researchers are making with data collection? It is likely that some cycle of implementation exists and that implementation varies as the semester progresses. Without accounting for existing variation within a semester, measurements taken at differing semester time-points may over or underestimate amount of instructional change. Strong indications already exist that level of use of such instructional strategies is overestimated (Henderson & Dancy, 2010).

Research Questions

This project seeks to track the implementation of active learning practices by faculty in large enrollment STEM classrooms. For this project, large enrollment is defined as enrollment > 50 at the fourth week of the semester.

1. What are the observable characteristics of implementation of active learning practices throughout the semester?
2. How consistent is observation data collected at different points in a semester?

LITERATURE REVIEW

Large enrollment courses were chosen as the specific subjects of study in this research project because it is already known that the level of active learning is significantly less common in them (Macdonald, Manduca, Mogk, & Tewksbury, 2005; Budd et al., 2013). Class size is often perceived as a large barrier to the implementation of active learning strategies (Henderson & Dancy, 2007). Using the observation instrument as a feedback tool can help overcome this barrier by providing instructors with an objective measure of their methodological changes. The importance of objectivity in the observation should not be underestimated. It is important for observers to be aware that feedback can be taken as criticism (Donnelly, 2007). Use of objective measurement instruments may reduce this perception. Prior research also notes that faculty often modify research-based pedagogies and tools (Henderson & Dancy, 2007). Such alterations can be reflected in observation feedback and corrections implemented when needed. Encouragement and support like this should be at the center of reform efforts (Woodbury & Newsome, 2002). A greater contribution to the Discipline Based Education Research (DBER) community could be made by revealing more details about transitioning in large enrollment classes than those of smaller size as large class size in traditional lecture halls is one of the barriers to active learning implementation most cited by faculty (Henderson & Dancy, 2010).

Most universities report average class sizes based on enrollment for courses at all undergraduate levels combined, however, it is widely acknowledged that typical 100 – 200 level courses are the largest. These introductory gatekeeper courses often reach sizes of over 300 and are likely to remain large due to budgetary considerations. Attrition in STEM majors is highest during the first two years of university enrollment while students are encountering these large enrollment courses (Labov, 2004; Gasiewski, Eagan, Garcia, Hurtado, and Chang, 2012).

Performing poorly in STEM courses in comparison to non-STEM has a strong association with students leaving STEM majors (Chen, 2013). Meta-analysis of instructional methods shows that those courses taught with active learning pedagogies have lower DFW (Withdrawals and Grades of D or F) rates (Freeman et al., 2014). Emergent data from the courses taught by participants in this study show a drop in DFW occurring (M. Hanson, personal communication, July 20, 2017). To better translate the approach for wide application, a more detailed understanding of how to transition STEM courses and what that process looks like is needed.

The question remains as to how often supportive observations should be conducted to generate a profile of the transition of an instructor and perform valid comparisons across semesters. It is known that even within one week of a course, COPUS results can fluctuate broadly (Wieman & Gilbert, 2014; Lund et al., 2015) but this is not accounted for in current research practices. A scan of the literature through Google Scholar conducted in August 2017 revealed 115 extant documents that cite COPUS. Eleven used COPUS to collect data for their studies. A summary of these published research articles is illustrated in Table 1. The most common methodology among researchers is to collect observations of a course at any point in the semester. Observation time-points are chosen for convenience either to match the observer’s schedule (Wieman & Gilbert, 2014; Smith et al., 2014; Lewin, Vinson, Stetzer, & Smith, 2016) or from videos chosen and submitted by the instructors themselves from any point in the semester (Lund et al., 2015; Stains, Pilarz, & Chakraverty, 2015; Lax, Morris, & Kolber, 2016; Jones, 2017).

Table 1
Extant articles citing COPUS as a portion of their data collection.

Article	**Observation Time-point**	**Number of Faculty Sampled**	**Number of courses / sections**	**Observations per Course Section**	**Main Subject of Research**
Achen & Lumpkin (2015)	All semester	1	1 course 2 sections	22	Instructor
Cleveland, Olimpo, & DeChenne-Peters (2017)	All semester	2	1 course 2 sections	9	Student Learning Gains
Connell, Donovan, & Chambers (2016)	All quarter	1	1 course 2 sections	5	Student Learning Gains
Jones (2017)	Not reported	Not reported	54	1 week**	Student Learning Gains
Lax et al. (2016)	Not reported*	Not Reported	1 course 2 sections	1 unit	Student Learning Gains
Lewin et al. (2016)	Feb, April, Nov	119	138	2***	Instructor
Lund et al. (2015)	All semester	73	Not reported	1 week	Instructor
Maciejewski (2016)	Not reported	6	1 course 7 sections	2	Student Learning Gains
Smith et al. (2014)	Feb, April	43	51	1 to 3	Instructor
Stains et al. (2015)	All semester	27	27	1 week	Instructor
Wieman & Gilbert (2014)	Not reported	49	49	1	Instructor

Note. *No reported time-point but confined study to one unit. ** Reports methods as "per Lund" so this is assumed to be 1 week. ***Calculated number of observations per course from other information reported in the article.

The on-going research practice of ignoring time-point in the semester encompasses several underlying assumptions that have not been tested. It fails to account for any underlying trends that may occur naturally within the course of a semester. How do holidays and mid-terms affect instructional practices? Do review days affect an instructor profile? Are all units and topics taught similarly by a given instructor? More personal factors may also be at play and cycle within a semester. A widely acknowledged trend among teachers known as the “Teaching Cycle” exists in K-12 education (Moir, 1990). The Teaching Cycle is illustrated in Figure 1. The impact of such affective cycles on classroom instruction choices is unknown. No evidence exists that suggests university faculty would not undergo some form of the Teaching Cycle as they move through a renovation of their course. It is also possible that the Gartner Hype Cycle (Linden & Fend, 2003) illustrated in Figure 2 may be influencing the adoption and fidelity of implementation for active learning techniques. While developed for business applications, it has been applied to describe adoption of new techniques in other areas (Lamb, Frazier, & Adams, 2008).

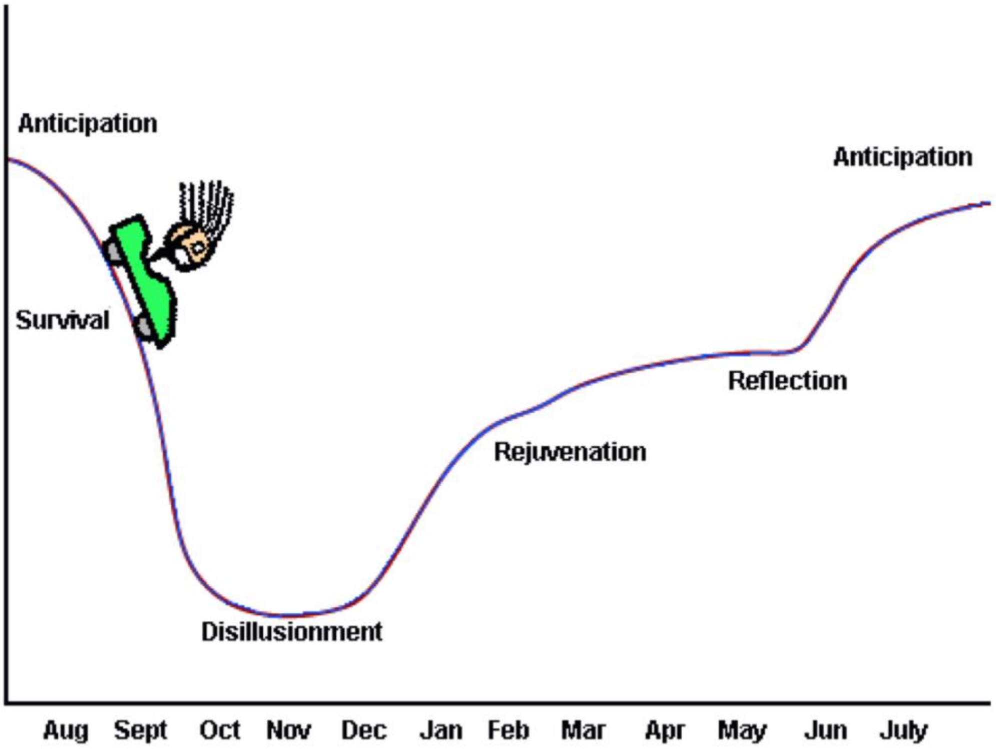


Figure 1. An illustration of the Teaching Cycle as it pertains to K-12 teachers (Ginsburg, 2011). Reprinted with permission of Ginsburg Educational Consulting and Coaching, LLC.

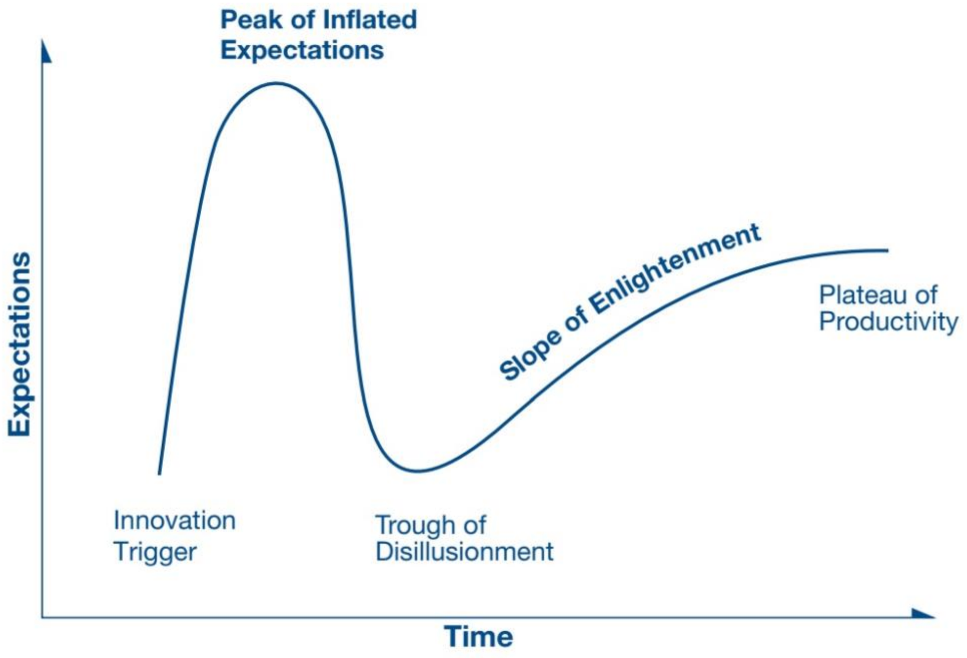


Figure 2. The Gartner Hype Cycle (Gartner Methodologies, 2017) has been applied to describe adoption of new technologies as they come on the market but might be able to describe adoption of new techniques in other areas. Reprinted with permission of Gartner Methodologies.

Enough preliminary and anecdotally reported evidence exists to indicate a need to investigate instructional cycles in undergraduate STEM courses. These cycles may inadvertently affect results reported in studies seeking to chart instructor profiles as they transition to active learning. For example, if instructors are more likely to engage in active learning based instruction early in the semester but measurements are taken late in the semester, the amount of transition could be underestimated. It is also possible that programs could miss points where instructors most need support. Without timely and appropriate support, abandonment or alteration of research based practices is more likely. This, in turn, would decrease their effectiveness and impact.

METHODS AND MATERIALS

Study Site

This project was conducted at North Dakota State University, Fargo, ND. Data were collected as part of the Gateways-ND program (NSF DUE 1525056). Gateways-ND is a two-year professional development program in which faculty apply for participation. One factor that makes this program unique is that it seeks to change classroom methodologies of current instructors instead of instituting change by hiring new instructional fellows as has been reported by some (Wieman & Gilbert, 2014). Five successive cohort groups of tenure and non-tenure track instructional faculty will be recruited during the duration of Gateways-ND. Each will sign up for two years of training specifically introducing Evidence Based Instructional Practices (EBIPs) through the project. EBIPs are generally classified in the category of active learning as previously defined. Priority in selection for the project was given to applicants with large enrollment undergraduate courses and those with a traditionally high DFW rate within the institution.

Thirty-six volunteer faculty members comprised Cohort I. Cohorts are groups of people who share a commonality. In this case, that commonality is the experience of pedagogical exposure and attempts at implementation. An illustration of the planned timeline of training and measurement is shown in Figure 3. Data were collected during the 2016 – 2017 instructional year. This coincided with the second year of their enrollment in the Gateways-ND program. Introduction to various evidence based practices occurred in January, May, and August 2016 as well as in January 2017. Supportive Faculty Learning Community (FLC) meetings were held once a month during Spring 2016, once every three weeks during Fall 2016, and once every three weeks during Spring 2017. This touchpoint combination of workshops and FLCs comprised the main EBIP exposure for participants.

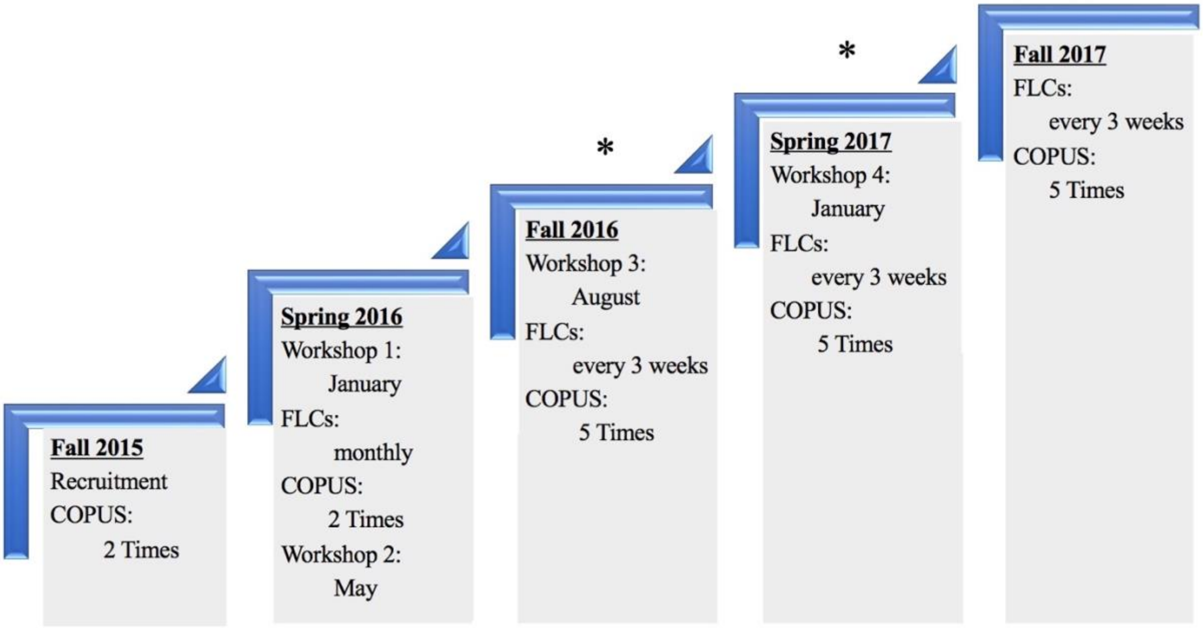


Figure 3. Cohort 1 Timeline and Training Plan. * Indicates semesters of data collection reported in this paper.

Participants in the research sample instructed the same course in paired semesters, either Fall 2015 and Fall 2016 or Spring 2016 and Spring 2017. Other criteria for inclusion in the sample were course enrollment greater than 50, attendance at a minimum of two of the four offered workshop training sessions, and attendance at half or more of the FLC meetings each semester. Ten met criteria for Fall 2016 and six for Spring 2017, four of which were also in the Fall 2016 sample. The instructors represent a diverse cross-section of the university (Table 2).

Two informational streams were gathered: classroom observations and interviews. Similar streams have been suggested in the literature to capture classroom change (Hamilton, et al., 2003). The data streams also align with the taxonomy of observable scientific teaching practices developed by Couch et al. (2015) to aid evaluation of course transformations.

Table 2
Demographic information of instructors (n = 12).

University College	Agriculture	3
	Engineering	4
	Science & Mathematics	5
Years Instructing in a University Faculty Position	0 to 5	3
	6 to 10	5
	11+	4
Rank	Non-tenure track	5
	Pre-tenure	2
	Tenured	5
Percentage of Appointment Designated as Research	0 to 10	5
	20 to 40	6
	50+	1

Data Stream I: Classroom Observations

A record of employed classroom practices was aggregated for each instructor utilizing the Classroom Observation Protocol for Undergraduate STEM (COPUS). The COPUS instrument was developed for use in situations similar to our investigation: characterizing the general state of STEM instruction, providing feedback to instructors who desired information about how they and their students spend time in class, identifying faculty professional development needs, and checking the accuracy of the faculty reporting of practices (Smith et al., 2013). The COPUS instrument provides a more objective record of what occurs in the classroom compared to other available instruments such as the Reformed Observation Teaching Protocol (Smith et al., 2014). The objectivity of the COPUS was vital to this project due to the wide range of STEM subjects encompassed by Gateways–ND. It allowed those not expert in all topics in an observed lesson to codify classroom activities.

Observers were trained by attending two, one-hour workshops introducing the instrument where initial agreement on code interpretations was reached. Observations were then conducted in the classrooms of volunteer instructors known to employ active learning practices but who were not members of the Gateways project. Interrater agreement was reached via discussion and reliability was checked through calculation of Cohen's kappa for observing pairs (training Cohen's kappa > 0.80). Observers also reported difficult to interpret codes and situations back to a general group where further discussion took place until agreement on such interpretation was reached. Observations began once sufficient interrater agreement was reached for all possible combinations of observers. Implementation observations were conducted by individual observers or pairs of observers. Each possible pairing of observers occurred at least once during research semesters as an interrater agreement check to ensure code interpretation remained consistent (Cohen's kappa ratings > 0.80).

The timing of observations varied based on observer availability but began after the first ten days of the semester and continued throughout the semester concluding at least one week before final exams. They were unannounced and spaced a minimum of three class sessions (approximately one week) but no more than three weeks apart to get a broader picture of the implementation landscape. Observations did not occur on days when more than half of a class session was given to exams, guest speakers, or student presentations. Five of each instructor were conducted during Fall 2016 and six during Spring 2017. One observation for each instructor was dropped from the Spring 2017 set to maintain consistency between the semesters. Dropped observations were chosen for elimination first to exclude those that occurred too late in the semester, next to maintain a spread as close to two weeks as possible, and finally to maximize data points within individual weeks. Dates of observations were mapped onto equivalent semester week producing 13 blocks. The blocks were divided into time-points of Early, Middle, and Late (Table 3).

Table 3
Mapping of observation dates into categorical time blocks.

Date of Observation		Semester Week	Time Period
Fall 2016	**Spring 2017**		
8/22/16 – 8/26/16	1/9/17 – 1/13/17	1	Excluded
8/29/16 – 9/2/16	1/16/17 – 1/20/17	2	
9/5/16 – 9/9/16	1/23/17 – 1/27/17	3	Early
9/12/16 – 9/16/16	1/30/17 – 2/3/17	4	Early
9/19/16 – 9/23/16	2/6/17 – 2/10/17	5	Early
9/26/16 – 9/30/16	2/13/17 – 2/17/17	6	Early
10/3/16 – 10/7/16	2/20/17 – 2/24/17	7	Middle*
10/10/16 – 10/14/16	2/27/17 - 3/3/17	8	Middle
10/17/16 – 10/20/16	3/6/17 – 3/10/17	9	Middle
10/24/16 – 10/28/16	3/20/17 – 3/24/17**	10	Middle
10/31/16 – 11/4/16	3/27/17 – 3/31	11	Middle
11/7/16 – 11/11/16	4/3/17 – 4/7/17	12	Late
11/14/16 – 11/18/16	4/10/17 – 4/14/17	13	Late
11/21/25/16 – 11/25/16	4/17/17 – 4/21/17	14	Late
11/28/16 – 12/2/16	4/24/17 – 4/28/17	15	Late
12/5/16 – 12/9/16	5/1/17 – 5/5/17	16	Excluded

Note. *The "Middle" category has an additional week as Week 8 corresponded to midterms and had the fewest observations conducted. **Missing dates correspond to Spring Break.

COPUS Calculations

A simple mathematical average of individual codes for each observation was calculated in instances where observations were conducted by pairs or triplets of observers. COPUS codes were then grouped based on protocol established by Smith et al., 2014. This method was introduced because it is difficult to determine trends when using the 25 individual codes generated by COPUS.

There are four collapsed COPUS categories for the Instructor and four for the Student. The Instructor codes are Presenting (P), Guiding (G), Administration (Adm), and Other (O). The student codes are Receiving (R), Talking to class (STC), Working (SW), and Other (OS). The details of each category are presented in Table 4.

Table 4

Description of the collapsed COPUS codes (Smith et al., 2014).

Who Acts	Collapsed Codes	Individual Codes
Instructor is:	Presenting (P)	Lec: Lecturing or presenting information
		RtW: Real-time writing
		D/V: Showing or conducting a demo, experiment, or simulation
	Guiding (G)	FUp: Follow-up/feedback on clicker question or activity
		PQ: Posing nonclicker question to students (nonrhetorical, **entire time, not just when first-asked)
		CQ: Asking clicker question (entire time, not just when first-asked)
		AnQ: Listening to and answering student questions to entire class.
		MG: Moving through the class guiding ongoing student work ***or actively monitoring student progress
		1o1: One-on-one extended discussion with individual students
	Administration (A)	Adm: Administration (assign homework, return tests, etc.)
	Other (OI)	O: Other
Students are:	Receiving (R)	L: Listening to instructor
	Talking to Class (STC)	AnQ: Student answering question posed by instructor
		SQ: Student asks question
		WC: Students engaged in whole-class discussion
		SP: Students presenting to entire class
	Working (SW)	Ind: Individual thinking/problem solving
		CG: Discussing clicker question in groups of students
		WG: Working in groups on worksheet activity
		OG: Other assigned group activity
		Prd: Making a prediction about a demo or experiment
		TQ: Test or quiz
	Other (OS)	W: Waiting (instructor late, working on fixing technical problem)
		O: Other

Note. *PQ interpretation altered to balance inherent bias of instrument towards CQ codes. **MG interpretation altered to include instructors who monitor versus those who engage in other tasks during student activities.

Statistical tests have been restricted to changes in G on the instructor side and SW on the student side as best representing learner centered practices. The G category encompasses those instructor activities most likely to occur in an active learning environment. SW codes were chosen for analysis as best indicators of the maximum number of students actively engaged with the classroom content. These measures were normalized to interval by dividing the raw COPUS score for each observation by number of possible intervals during one class session. This allowed direct comparison between courses with differing class lengths of 50 and 75 minutes.

Final averages for each instructor per each time period were used to generate the "Average" collapsed code data for repeated measures ANOVA. Statistical tests were not conducted on individual codes to avoid generating Type I errors. Because the process of creating an average also reduces variation, a random sampling of one observation per instructor per time period was pulled for a separate "Random" repeated measures ANOVA. The purpose of testing the Random set was twofold. First, to determine how much the greater variation would impact the p-value of the data set. Second, to help gauge if one observation per time period would be sufficient to capture variation between time-periods. All calculations were performed using SPSS Version 24.

Data Stream II: Semi-structured Interviews

Semi-structured interviews were conducted with all participating instructors during Spring 2017. A downward trend in active learning implementation had already been noted in the COPUS data. One question was specifically designed with designed to reveal views on instructional trends within the semester: "What do you think could cause the general trend downward in active learning implementation that we are observing in some of the courses?" Additional thoughts were re-solicited with the question, "Do you have any other insights to offer?" Transcripts were coded in an iterative process. Relevant themes were identified and samples are reported here.

RESULTS

COPUS Data

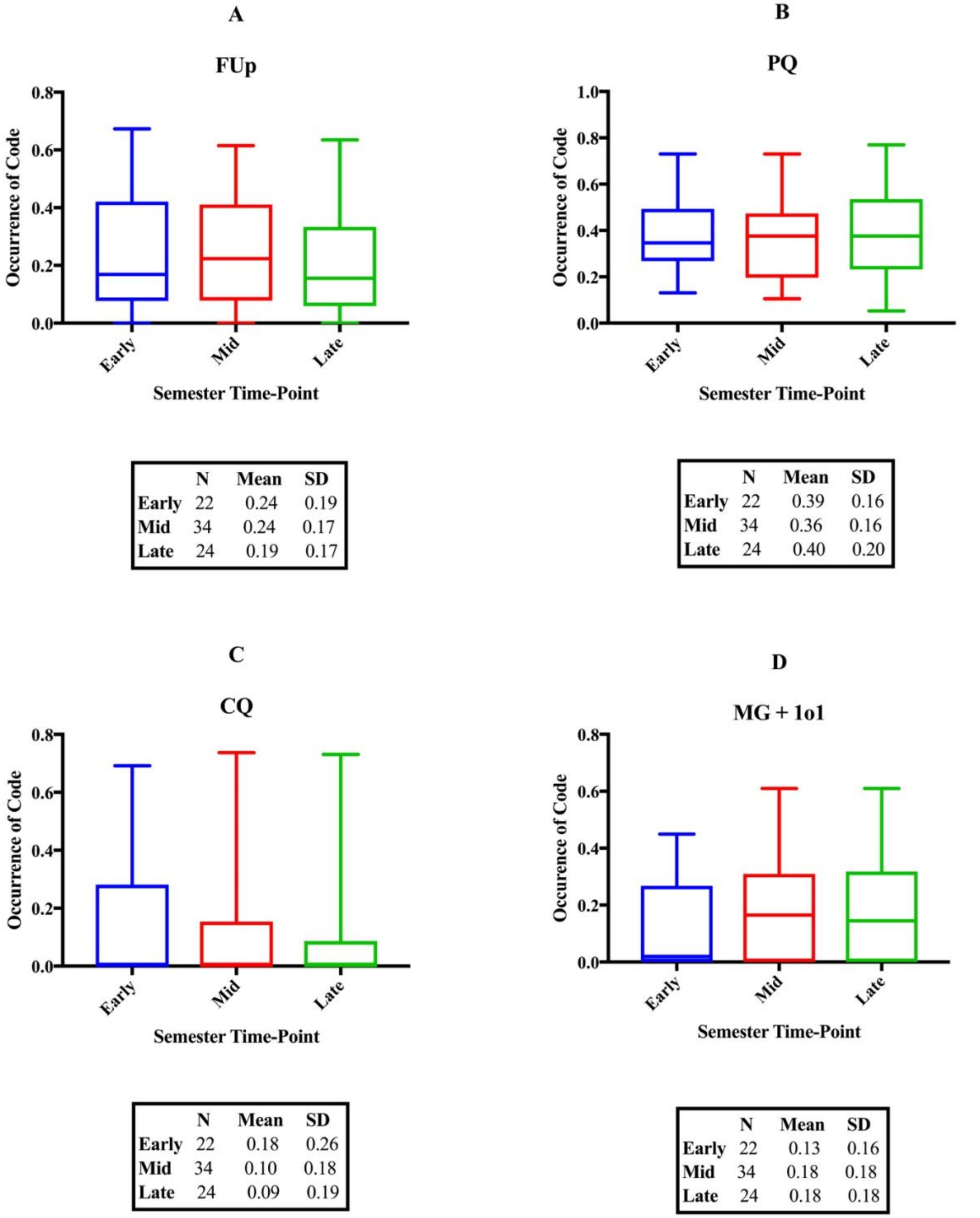
Individual COPUS codes were studied strictly through descriptive statistics to reduce the probability of generating a Type I error. Different patterns of implementation are visible for each member of the G code set (Figure 4; Table 5). The FUp (Figure 4, Panel A) and PQ (Figure 4, Panel B) codes seem to have the most stable implementation with little change in their medians, means, or SDs. The interquartile ranges do shift around the medians with PQ showing less shifting than FUp (Table 5). While the upper range stayed steady for both, PQ's minimum progressed further downward through the semester. In contrast, CQs (Figure 4, Panel C) were only in use among one quartile of participants and its mean reveals a steady decrease. The codes MG and 1o1 (Figure 4, Panel D; Table 5) were combined because 1o1 is a specific activity most often done while an instructor performs MG. Therefore, it represents a subset of MG. Original observation sheets were examined and code counts were corrected to ensure the combination was not tallied twice in instances where MG and 1o1 occurred during the same time interval. The mean of this combined category holds fairly steady but the median jumps upward abruptly during the Mid portion of the semester and remains at that height. Results from the final code comprising the G code set, AnQ, has been excluded from this study as it was found to not represent an activity that the instructor plans for independently. It is simply a response to a Student Question (SQ), which also does not lie among the codes of interest for the purposes of this research.

Patterns in the majority of the SW code set are shown in Figure 5 and Table 5. Ind (Figure 5, Panel A; Table 5) shows the most dramatic drop with the vast majority of its occurrence during the Early time-point. CG (Figure 5, Panel B) follows a downward stepped pattern without a dramatic drop. The WG and OG codes were combined for comparison purposes (Figure 5, Panel C; Table 5) because many of the instructors project their activities instead of printing out physical worksheets. This forces the WG code to be a subset of the OG code similar to the relationship between MG and 1o1. No occurrences of these codes sharing a single time interval existed so no double tallies needed correction. The medians and means of this category show a rise throughout the semester (Figure 5; Table 5). This is opposite other observed patterns. No representation for the Prd category is shown as no occurrence of Prd was observed in any course. T/Q is also unrepresented as efforts were made to avoid observing on exam days.

Table 5

Median and Interquartile Range (IQR) for Individual COPUS Codes at each semester time-point.

COPUS Code		Semester Time-Point								
		Early			Mid			Late		
		Median	IQR		Median	IQR		Median	IQR	
			25th %	75th %		25th %	75th %		25th %	75th %
G	FUp	0.17	0.08	0.42	0.22	0.08	0.41	0.16	0.06	0.33
	PQ	0.35	0.27	0.49	0.38	0.20	0.47	0.38	0.23	0.54
	CQ	0	0	0.28	0	0	0.15	0	0	0.09
	MG +1o1	0.02	0	0.27	0.17	0	0.31	0.14	0	0.32
SW	Ind	0.03	0	0.27	0	0	0.03	0	0	0.04
	CG	0	0	0.18	0	0	0.15	0	0	0.09
	WG + OG	0	0	0.19	0.07	0	0.34	0.17	0	0.31



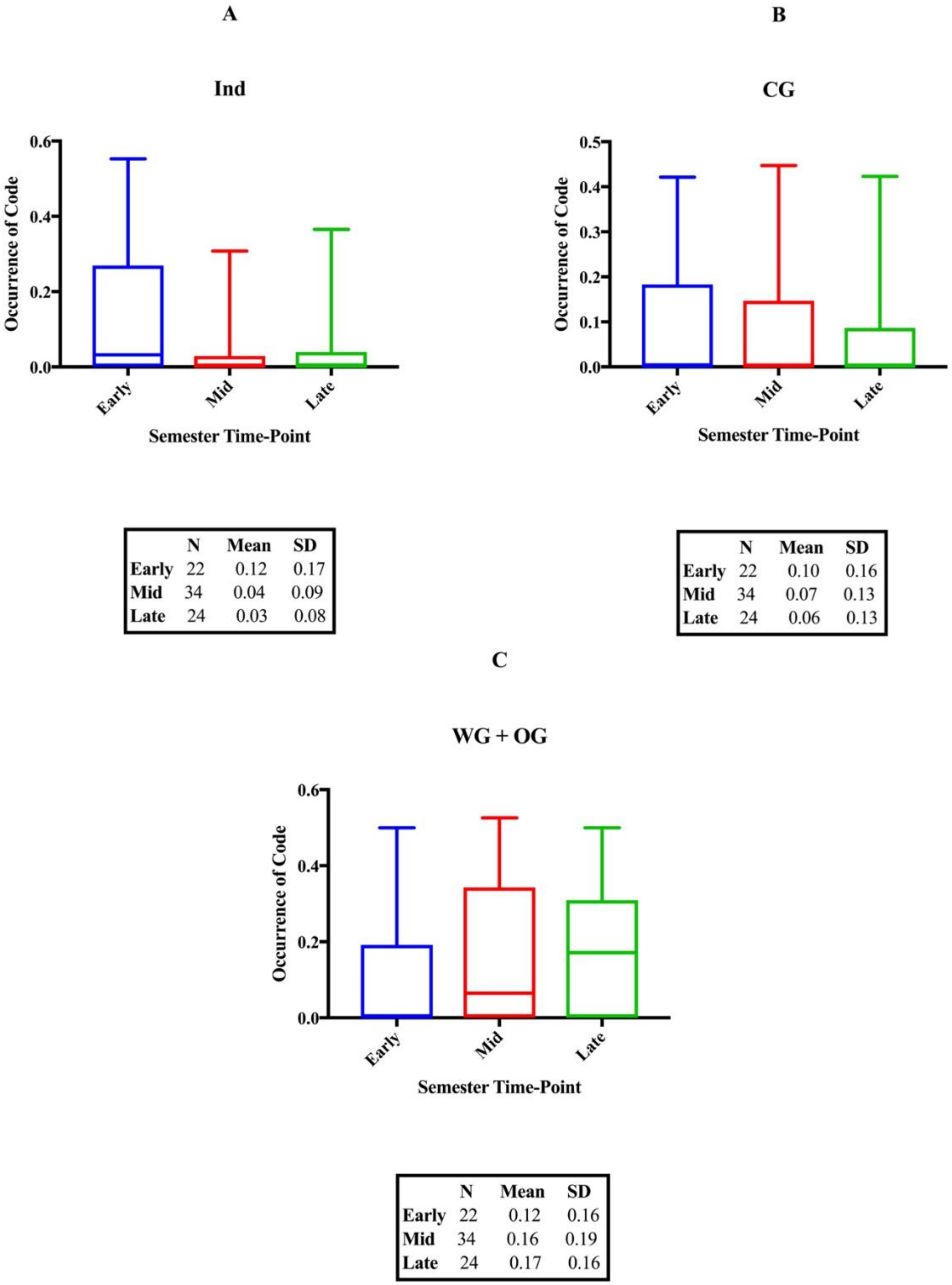
	N	Mean	SD
Early	22	0.24	0.19
Mid	34	0.24	0.17
Late	24	0.19	0.17

	N	Mean	SD
Early	22	0.39	0.16
Mid	34	0.36	0.16
Late	24	0.40	0.20

	N	Mean	SD
Early	22	0.18	0.26
Mid	34	0.10	0.18
Late	24	0.09	0.19

	N	Mean	SD
Early	22	0.13	0.16
Mid	34	0.18	0.18
Late	24	0.18	0.18

Figure 4. Box-and-whisker plots display individual descriptive statistics for four of the codes that make up the G code set. Values on Y axes are left to vary based on the natural spread of the data.



	N	Mean	SD
Early	22	0.12	0.17
Mid	34	0.04	0.09
Late	24	0.03	0.08

	N	Mean	SD
Early	22	0.10	0.16
Mid	34	0.07	0.13
Late	24	0.06	0.13

	N	Mean	SD
Early	22	0.12	0.16
Mid	34	0.16	0.19
Late	24	0.17	0.16

Figure 5. Box-and-whisker plots display individual descriptive statistics for the SW code set. Values on Y axes are left to vary based on the natural spread of the data.

Descriptive statistics for the averaged collapsed code sets revealed a downward trend in means with increasing variance for both G and SW as the semester progressed (Table 6). An examination of the data using range, median, and quartiles also revealed steady trends downward for most of these measures as the semester progressed (Figure 6). Greater variance was observed in the Random sample in all cases except the Mid measurement of the SW category (Table 6 and Figure 6) but similar trends were present in all sections except SW at the Mid time-point.

Table 6
Descriptive statistics across collapsed code categories and samples.

Sample	Category	Time-point	Mean	SD	Variance
Averaged	G	Early	0.978	0.372	0.138
		Mid	0.942	0.402	0.162
		Late	0.913	0.409	0.168
	SW	Early	0.329	0.246	0.061
		Mid	0.274	0.220	0.048
		Late	0.277	0.221	0.049
Random	G	Early	1.006	0.405	0.152
		Mid	0.894	0.490	0.240
		Late	0.896	0.417	0.174
	SW	Early	0.346	0.229	0.052
		Mid	0.274	0.268	0.072
		Late	0.282	0.239	0.057

One-way repeated measures ANOVAs were conducted to compare the occurrence of G and SW codes at different semester time points. All data met standard assumptions of normality including skewness, kurtosis, Shapiro-Wilk, and Q-Q plot. Data for the G category did not meet Mauchly’s Test of Sphericity so a Greenhouse-Geisser correction was used. No statistically significant differences were found for either code category (G code set F(1.5) = 0.36, p = 0.63; SW code set F(1.7) = 1.6, p = 0.23) . Results for all ANOVA tests performed are summarized in Table 7.

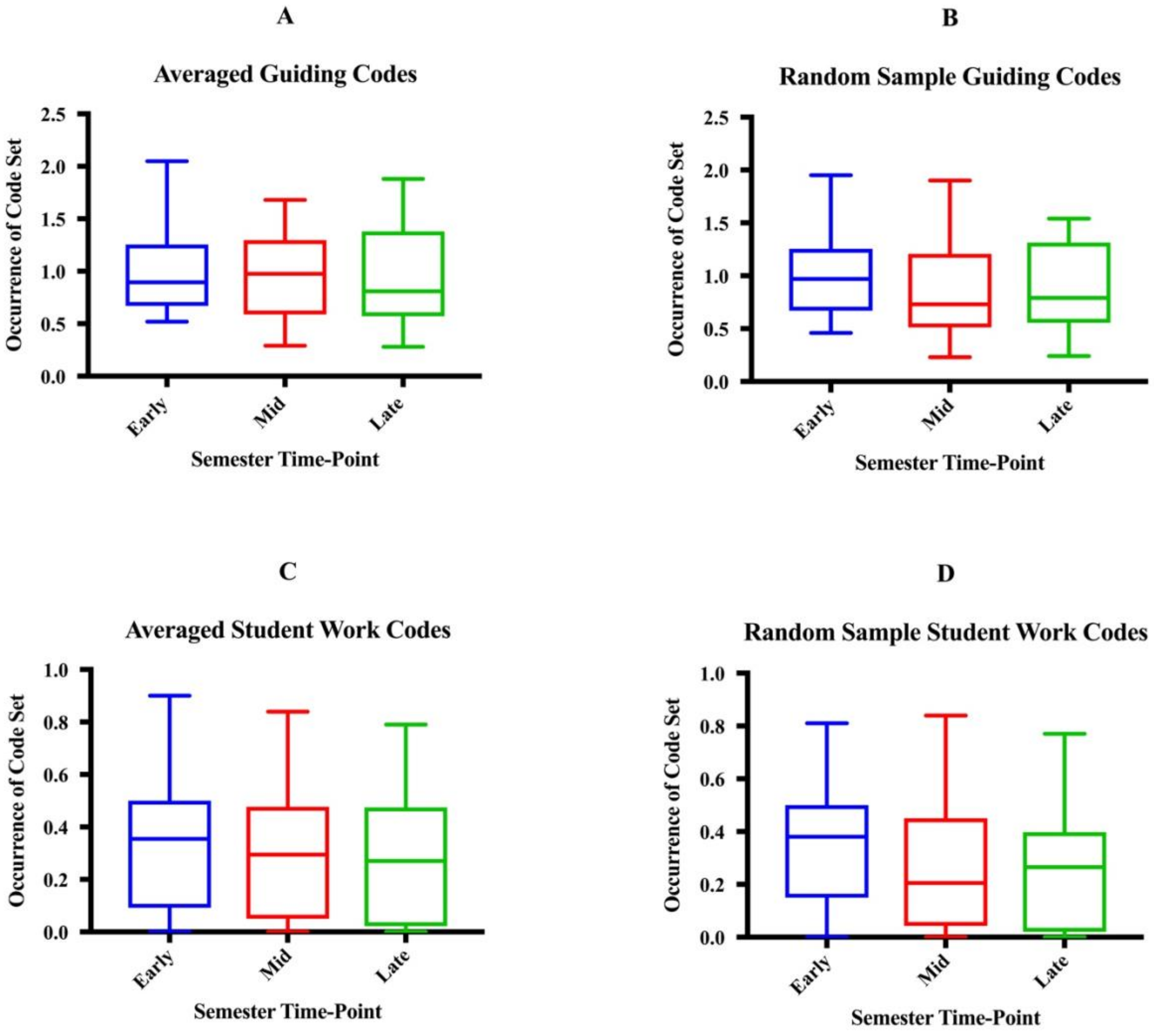


Figure 6. Box-and-whisker plots display descriptive statistic semester trends in codes for averaged data and randomly sampled data.

Table 7
Repeated Measures ANOVA results.

Sample	Code Group	df	F	p value	Observed Power
Averaged	G	1.5	0.36	0.63	0.10
	SW	1.7	1.6	0.23	0.28
Randomly Drawn	G	1.8	1.3	0.28	0.25
	SW	2.0	2.0	0.16	0.37

Interview Data

No interviewees expressed surprise at the idea of a general trend downward in active learning as the semester progressed. Many voluntarily expressed surprise at the thought that it wouldn't trend downward. Emergent reasons for why this would occur are pressure to cover material increases towards the end of the semester, being tired, falling back to the familiar, and having tried active learning in the early portion of the semester but choosing to abandon it (Figure 7). One comment regarding abandonment of CQs due to coverage pressure was offered. No other insights regarding specific code types or other trends were offered.

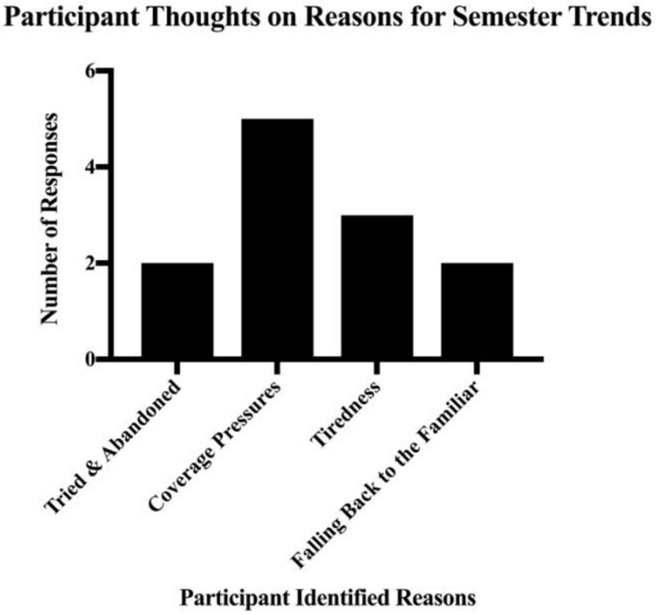


Figure 7. Emergent themes identified from participants asked for their thoughts on implementation trends.

DISCUSSION

Research Question 1

Previous studies have discussed the difficulty in examining individual COPUS codes as their interrelationships are quite complex (Smith et al., 2014; Lund et al., 2015; Lewin et al., 2016). Code categories were devised to partially overcome this challenge (Smith et al., 2014). However, differential implementation of the codes within a category could cause interference. If one or more of the codes in a category set trend downward while others trend upward, they could cancel each other out within the averaged data producing an artificially flattened line or possibly a reversal. Trends in individual COPUS codes from the G and SW categories were examined as a portion of this study to check for these possibilities.

The greatest consistency in implementation among the six codes that comprise the G code sets was found in FUp and PQ (Figure 4, Panels A and B). All instructors displayed some level of PQ at each time-point in the semester with a fairly steady mean, median, and interquartile range. All other codes revealed that each time-point held at least one course where that practice did not appear. This is seen as the min range line is at zero for all other codes. In addition, CG, which did not have a broad adoption among instructors, does not illustrate a lowest quartile or a median. For these reasons, analysis is best served by examination of the interquartile range (IQR), median where possible, mean, and variance.

The clearest trends in median and mean are seen in the SW codes Ind, CG, and WG+OG (Figure 5). Ind and CG dropped as the semester progressed while WG+OG increased. A similarly opposite movement is seen in the corresponding G codes of CQ and MG+1o1 (Figure 4, Panels C and D) where CQ descends and MG+1o1 rises. The problem of reversal in averaged data sets similar to this one is known as Simpson’s paradox. This particular statistical anomaly is defined as a situation where a trend appears in data sets but disappears or reverses when the groups are combined (Pearl, 2014). It may indicate the presence of a separate confounding or lurking variable (Pearl, 2014). Caution in interpretation is recommended as the paradox can lead to incorrect conclusions (Pearl, 2014; Pollet, Stulp, Henzi, & Barrett, 2015). The presence of Simpson's paradox in both code sets verifies the need to examine trends in these codes individually as well as in the collapsed form.

One recommendation for dealing with the presence of Simpson's paradox is to pay close attention to variance (Pollet et al., 2015). The largest shifts in variance occurred in the codes Ind, CG, and CQ. These are the codes that trended down in implementation, indicating abandonment by instructors toward the end of the semester. Movement downward is predicted by both the Teaching Cycle (Moir, 1990) and the Gartner Hype Cycle (Gartner, 2017). Ind is likely influenced by CQ as it is often linked pedagogically with Mazur's Peer Instruction Model and clickers (Crouch & Mazur, 2001). The trend here could indicate instructors abandoning techniques that require more planning before implementation. Interviewee M highlighted time as an important factor, "Especially toward the end of the semester, sometimes time management is the issue." The PQ IQR rose during the Late time-point and its variance increased while CQs reached their lowest level. Why more PQs and fewer CQs? Participants identified pressure to cover material or catch up as the most common reason to abandon active learning practices later in the semester. "What drove me to use the flashcards less at the end was that I felt time pressured so much. Pressures in what I wanted to get through," said Interviewee G. The late rise in PQ is another detail likely lost to Simpson's paradox. CQ implementation is the only COPUS code that has been the independent subject of research scrutiny (Lewin et al., 2016). The findings in this report indicate the possibility that time-point of observation may account for some of the reported wide variance in CQ.

The reasons behind the revealed patterns have yet to be explored. A lurking variable influencing implementation may be at work as well as more complex interactions. CQs may have a larger than anticipated influence on the other codes. Being technologically based in clicker systems, it is also the code most likely to be matched with the Gartner Hype Cycle of adoption. If so, support could be better targeted in future pedagogical training exercises to help bypass the disillusionment phase.

Research Question 2

Averaged Sample vs. Random Sample

As expected, a greater variance is displayed in the Random Sample than in the Averaged Sample (Figure 6). The Random Sample was generated to test the idea that such a sample, with its increased variance, might reveal a statistically significant difference between time-points. The p values did trend closer to the line of statistical significance but did not cross it (Table 7). The power of both data sets was very low but similar data trends are evident in both sample sets. The similarity in trends may indicate that a single observation conducted during each time point might be sufficient to capture time-point trends. As most studies use Averaged data, that set is the subject of further discussion.

Averaged Code Sets G and SW

Averaged COPUS code sets are the starting point for analysis of change within the larger Gateways-ND project. The ultimate goal of the program is to positively affect student experiences and achievement in STEM courses at the university level. Statistically significant change in student achievement as measured by DFW rate (Figure 8) has been linked to Large (X^2

(1, N = 1913) = 23.9, p < 0.0001) and Moderate (X^2 (1, N = 941) = 11.0, p <0.001) changes in the G code set and Large (X^2 (1, N = 1413) = 23.6, p < 0.0001) changes in the SW code set among courses with enrollment > 75 in the Gateways-ND project (M. Hanson, personal communication, July 20, 2017). These comparisons are based on baseline data collected during the Late time-point and implementation data collected during the Early and Mid time-points. Increasing student achievement is the ultimate goal of reform so identifying trends within the semester that could cause conclusions about impacts of the program to be over- or underestimated need to be identified.

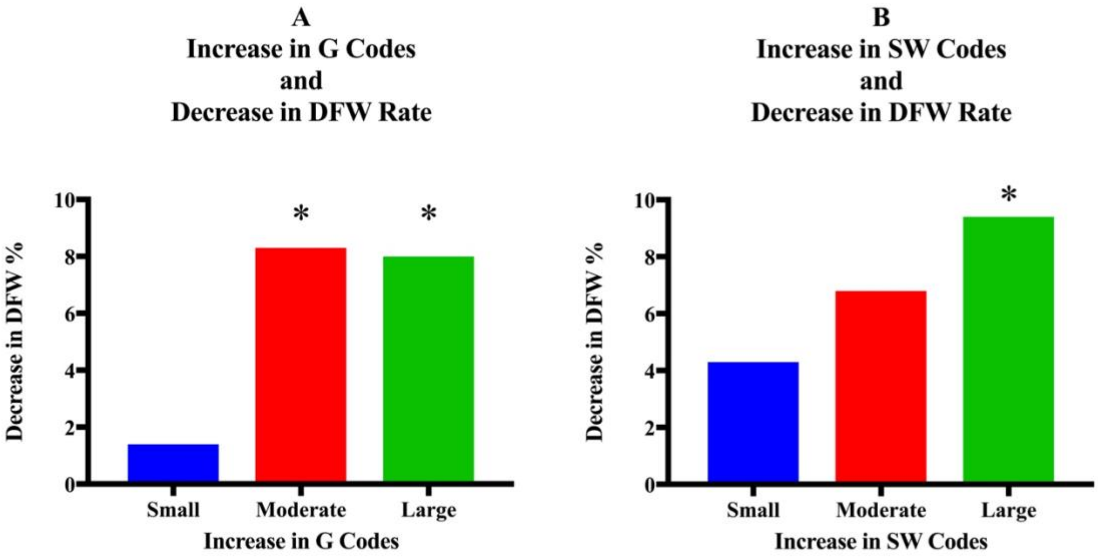


Figure 8. Statistically significant decrease in DFW Rates linked to degree of increase in G and SW code sets (M. Hanson, personal communication, July 20, 2017).

While the RM-ANOVA did not reveal any statistically significant difference in time-point (Table 7), the low power of the test and the presence of Simpson's paradox warranted a closer look at the visible trends within the data. Thought needs to be given to lurking variables that could interact with the results. The upper quartile and median both held relatively steady in the Averaged data. This signals some instructors may be holding steady or, at least, that changes are in an equilibrium situation where increases in implementation equal decreases. Downward movement is in the min-max points and the lowest quartile of implementation. One explanation for the downward trend is instructors transitioning to learner centered practices more slowly than expected. This concurs with findings of prior research that described pedagogies and instructional methodologies not implemented at a predicted level (Budd et al., 2013). Instructors found in the lowest quartile may need more support to prevent abandonment of active-learning practices. General fatigue toward the end of the semester may have caused instructors to fall into old, traditional habits. Interviewee I offered the following insight: "… they start doing it right out of the gate and then as the semester becomes overwhelming they start to fall back to lecturing at the end." This was independently supported by Interviewee F: "When you're rushed, it's easy to fall back on the old ways." Again, this would indicate a need for additional support.

It is also possible that workshops produced a recency effect. Interviewees identified workshops as the most influential piece of training within the program. "The workshops helped me pull out the most pedagogy," said Interviewee C. Advice given during the workshops could have compounded the recency effect. During their first year of training, participants were cautioned against making too many changes in one semester. It was suggested that they redesign their course in pieces. Interviewee K described classroom implementation as, "It's really kind of little, little baby steps." It is likely that the participants chose to alter lessons falling closest to a workshop. "In September, you're so early on in the course you are trying things out," offered Interviewee C. This would cause an elevation at the Early time-point. Later lessons may have been left to languish in traditional lecture format causing a decrease. These actions would equate to the predicted beginning height on both the Teaching Cycle (Moir, 1990) and the Gartner Hype

Cycle (Linden & Fend, 2003). The Gartner Hype Cycle specifically predicts the height as a reaction to the introduction of innovation. The second step on both cycles is a downward slope that is consistent with the slight general downward trend. Interviewee I, who did try to redesign an entire course in one semester related the experience as, "I just kind of did it anyways. I learned from my mistake." This opinion was likely shared within the FLC. Abandonment of active learning methodologies at different rates based on individual preference would have caused the increased variance. However, this does not account for the rise seen in the subsets of codes most closely related to group work not involving clickers (MG+1o1 and WG+OG).

It is probable that a lurking variable is the cause in the rise of the group work related codes. This could be a factor based on department, incoming experience, or even the room layout. Room layout is particularly attractive as the lurker in this case. Instructors scheduled into the most traditional rooms, large lecture auditoriums, may have required longer processing and planning time before implementing active learning strategies. In other words, the codes displayed a learning curve for implementation in those rooms. Effects of department, experience, and room layout are unknown as those lay outside the scope of this study.

In addition to the slight downturn in implementation, variance is observed to increase on all graphs as the semester progresses. This is indicative of decreased fidelity in implementation of active learning practices. While variance in COPUS data has been given some slight acknowledgement (Wieman & Gilbert, 2014) it has been generally dismissed or overlooked by researchers. Our data give indication that some care may be needed when comparing COPUS data collected at different semester time-points because the variance increases. The underlying presence of the Simpson's paradox in the code sets makes an examination of the changes in variance a valuable step when interpreting the data (Pollett et al., 2015). This could be significant to researchers as at least some studies (Smith et al., 2014; Lewin et al., 2016) report pooled COPUS data collected at what would here be classified as Early and Late semester. If only collapsed code sets are examined, the increasing variance could interfere with findings and finer details in the data's story may be missed.

Other details in trends are also missed because G codes do not map exclusively to the SW codes. Some portions of G map to "Students Talking to Class" (STC) not SW (Table 3). For example, if a student asks a question (SQ) it falls within STC codes but the instructor response of Answering a Question (AnQ) would fall within G. That subtlety of interaction is lost in the SW grouping, which groups codes reflective of the activity of large segments of the class. Expansion to include interactions of the STC codes is likely to provide a richer picture even though these are usually rated as less active (Lund et al., 2015).

While imperfect, the COPUS remains a standard for observations even as new protocols such as the Practical Observation Rubric to Assess Active Learning (Eddy, Converse, & Wenderoth, 2015) come online. Perhaps the need is not for additional observation protocols but more care in the interpretation.

Limitations

The low power of this study is the major limitation (Table 5). It is possible and, in fact, highly probable, that it has returned a false negative. This could be ameliorated by continuing data collection with Cohort 2 in its second year (Fall 2017 – Spring 2018) to increase n. An increase in n would also help verify the possible Simpson's paradox revealed in the data. A test probe of Ind, the code with the largest shift involved in the Simpson's paradox, via RM-ANOVA revealed a significant change through the semester (F (1.3, 20) = 3.9; p = 0.05). Further statistical tests were not completed with this data to avoid a Type I error.

This study did not look at the confounding category of “Students Talking to Class” (STC) from the COPUS. Failure to include STC codes means that some instructor G codes do not have a relevant mirror on the student side of the data. The variance this may have introduced is unknown. It is also unknown if the findings here would hold true in lower enrollment courses.

Future Directions

More data is being collected from 10 members of Cohort II which is in-progress Year 2 at this time. This training point is equivalent to when the data reported here was collected from Cohort I. The increase in sample size will increase the statistical power of this study to yield more definitive results. On-going data collection is also occurring with 9 of the members of Cohort I reported here. This will yield a comparator group of Final Year implementation to explore the question: How much does implementation variation continue?

Inclusion of STC codes in future analyses would provide a richer view of trends among instructors. Anecdotally, the College of Engineering in particular is moving in the direction of Socratic Questioning which is an instructional methodology that would not be captured by G codes but would by an examination of STC codes.

Conclusion

This study explored the implementation of learner centered instructional practices over the course of a semester. This could be a critical question for the evaluation of research projects seeking to track changes in instructional practices. While a general trend downward was observed in both G and SW code sets in these courses, no statistically significant difference was found in level of implementation at different time-points. Failure to account for the downward trend could end in over-exaggerating instructional changes among instructors when reporting statistically significant decreases in DFW rates as emerging from Gateways-ND. Additionally, a potential Simpson’s paradox was detected in the code groupings indicating a need to examine codes individually before grouping as the paradox could produce artificially flattened lines of change. The low power of this study justifies further examination of the descriptive trends.

REFERENCES

Achen, R., & Lumpkin, A. (2015). Evaluating Classroom Time through Systematic Analysis and Student Feedback. *International Journal for the Scholarship of Teaching and Learning, 9*(2), 4. https://doi.org/https://doi.org/10.20429/ijsotl.2015.090204

American Association for the Advancement of Science (1994). *Benchmarks for Science Literacy*. Oxford University Press.

American Association for the Advancement of Science (2010). Vision and Change: A Call to Action, Washington, DC.

Budd, D. A., van der Hoeven Kraft, K. J., McConnell, D. A., & Vislova, T. (2013). Characterizing Teaching in Introductory Geology Courses: Measuring Classroom Practices. *Journal of Geoscience Education, 61*(4), 461–475. https://doi.org/10.5408/12-381.1

Chen, X. (2013). STEM Attrition: College Students' Paths into and out of STEM Fields. Statistical Analysis Report. NCES 2014 – 001. *National Center for Education Statistics.*

Cleveland, L. M., Olimpo, J. T., & DeChenne-Peters, S. E. (2017). Investigating the Relationship between Instructors' Use of Active-Learning Strategies and Students' Conceptual Understanding and Affective Changes in Introductory Biology: A Comparison of Two Active-Learning Environments. *CBE-Life Sciences Education, 16*(2), ar19. https://doi.org/10.1187/cbe.16-06-0181

Connell, G. L., Donovan, D. A., & Chambers, T. G. (2016). Increasing the Use of Student-Centered Pedagogies from Moderate to High Improves Student Learning and Attitudes about Biology. *CBE-Life Sciences Education*, *15*(1), ar3. https://doi.org/10.1187/cbe.15-03-0062

Couch, B. A., Brown, T. L., Schelpat, T. J., Graham, M. J., & Knight, J. K. (2015). Scientific Teaching: Defining a Taxonomy of Observable Practices. *CBE-Life Sciences Education*, *14*(1), ar9. https://doi.org/10.1187/cbe.14-01-0002

Crouch, C. H., & Mazur, E. (2001). Peer Instruction: Ten years of experience and results. *American Journal of Physics*, *69*(9), 970–977. https://doi.org/10.1119/1.1374249

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., Converse, M., & Wenderoth, M. P. (2015). PORTAAL: A Classroom Observation Tool Assessing Evidence-Based Teaching Practices for Active Learning in Large Science, Technology, Engineering, and Mathematics Classes. *CBE-Life Sciences Education*, *14*(2), ar23. https://doi.org/10.1187/cbe.14-06-0095

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Gartner Methodologies (2017). Gartner Hype Cycle. Retrieved September 21, 2017, from http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp

Gasiewski, J. A., Eagan, M. K., Garcia, G. A., Hurtado, S., & Chang, M. J. (2012). From Gatekeeping to Engagement: A Multicontextual, Mixed Method Study of Student Academic Engagement in Introductory STEM Courses. *Research in Higher Education*, *53*(2), 229–261. https://doi.org/10.1007/s11162-011-9247-y

Ginsburg, D. (2011). The First-Year Teaching Roller Coaster. Retrieved September 21, 2017, from http://blogs.edweek.org/teachers/coach_gs_teaching_tips/2011/08/the_first-year_teaching_roller_coaster.html?cmp=SOC-SHR-FB

Hamilton, L. S., McCaffrey, D. F., Stecher, B. M., Klein, S. P., Robyn, A., & Bugliari, D. (2003). Studying Large-Scale Reforms of Instructional Practice: An Example from Mathematics and Science. *Educational Evaluation and Policy Analysis*, *25*(1), 1–29. https://doi.org/10.3102/01623737025001001

Henderson, C., & Dancy, M. H. (2007). Barriers to the use of research-based instructional strategies: The influence of both individual and situational characteristics. *Physical Review Special Topics - Physics Education Research*, *3*(2), 20102. https://doi.org/10.1103/PhysRevSTPER.3.020102

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

Jones, F. (2017). Comparing student, instructor, classroom and institutional data to evaluate a seven-year department-wide science education initiative. *Assessment & Evaluation in Higher Education*, *0*(0), 1–16. https://doi.org/10.1080/02602938.2017.1343799

Labov, J. B. (2004). From the National Academies: The Challenges and Opportunities for Improving Undergraduate Science Education through Introductory Courses. *Cell Biology Education*, *3*(4), 212–214. https://doi.org/10.1187/cbe.04-07-0049

Lamb, D. W., Frazier, P., & Adams, P. (2008). Improving pathways to adoption: Putting the right P's in precision agriculture. *Computers and Electronics in Agriculture*, *61*(1), 4–9. https://doi.org/10.1016/j.compag.2007.04.009

Lax, N., Morris, J., & Kolber, B. J. (2016). A partial flip classroom exercise in a large introductory general biology course increases performance at multiple levels. *Journal of Biological Education*, *0*(0), 1–15. https://doi.org/10.1080/00219266.2016.1257503

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. https://doi.org/10.1187/cbe.15-10-0224

Linden, A. & Fend, J. (2003) Understanding Gartner's Hype Cycles. Strategic Analysis Report No R-20-1971. Gartner, Inc.

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.14-10-0168

Macdonald, R. H., Manduca, C. A., Mogk, D. W., & Tewksbury, B. J. (2005). Teaching Methods in Undergraduate Geoscience Courses: Results of the 2004 On the Cutting Edge Survey of U.S. Faculty. *Journal of Geoscience Education*, *53*(3), 237–252. https://doi.org/10.5408/1089-9995-53.3.237

Maciejewski, W. (2016). 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

Moir, E. (1990). Phases of first-year teaching. *Originally published in California New Teacher Project Newsletter*.

Olson, S., & Riordan, D. G. (2012). *Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics. Report to* *the President*. Executive Office of the President. Retrieved from https://eric.ed.gov/?id=ED541511

Pearl, J. (2014). Comment: Understanding Simpson's Paradox. *The American Statistician*, *68*(1), 8–13. https://doi.org/10.1080/00031305.2014.876829

Piburn, M., & Sawada, D. (2000). *Reformed Teaching Observation Protocol (RTOP) Reference Manual. Technical Report*. Retrieved from https://eric.ed.gov/?id=ED447205

Pollet, T. V., Stulp, G., Henzi, S. P., & Barrett, L. (2015). Taking the aggravation out of data aggregation: A conceptual guide to dealing with statistical issues related to the pooling of individual-level observational data. *American Journal of Primatology*, *77*(7), 727–740. https://doi.org/10.1002/ajp.22405

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*(6), 245–253. https://doi.org/10.1111/j.1949-8594.2002.tb17883.x

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

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

Stains, M., Pilarz, M., & Chakraverty, D. (2015). Short and Long-Term Impacts of the Cottrell Scholars Collaborative New Faculty Workshop. *Journal of Chemical Education*, *92*(9), 1466–1476. https://doi.org/10.1021/acs.jchemed.5b00324

Wieman, C., & Gilbert, S. (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

Woodbury, S., & Gess-Newsome, J. (2002). Overcoming the Paradox of Change without Difference: A Model of Change in the Arena of Fundamental School Reform. *Educational Policy*, *16*(5), 763–782. https://doi.org/10.1177/089590402237312