Effectiveness of a Low-Cost, Graduate Student–Led Intervention on Study Habits and Performance in Introductory Biology

Tyler D. Hoskins,[†]* J. D. Gantz,[†] Blake R. Chaffee,[‡] Kel Arlinghaus,[†] James Wiebler,[†] Michael Hughes,[§] and Joyce J. Fernandes[†]

[†]Department of Biology and [§]Department of Statistics, Statistical Consulting Center, Miami University, Oxford, OH 45056; [‡]Cleveland Clinic, Genomic Medicine Institute, Cleveland, OH 44106

ABSTRACT

Institutions have developed diverse approaches that vary in effectiveness and cost to improve student performance in introductory science, technology, engineering, and mathematics courses. We developed a low-cost, graduate student–led, metacognition-based study skills course taught in conjunction with the introductory biology series at Miami University. Our approach aimed to improve performance for underachieving students by combining an existing framework for the process of learning (the study cycle) with concrete tools (outlines and concept maps) that have been shown to encourage deep understanding. To assess the effectiveness of our efforts, we asked 1) how effective our voluntary recruitment model was at enrolling the target cohort, 2) how the course impacted performance on lecture exams, 3) how the course impacted study habits and techniques, and 4) whether there are particular study habits or techniques that are associated with large improvements on exam scores. Voluntary recruitment attracted only 11–17% of our target cohort. While focal students improved on lecture exams relative to their peers who did not enroll, gains were relatively modest, and not all students improved. Further, although students across both semesters of our study reported improved study habits (based on pre and post surveys) and on outlines and concept maps (based on retrospectively scored assignments), gains were more dramatic in the Fall semester. Multivariate models revealed that, while changes in study habits and in the quality of outlines and concept maps were weakly associated with change in performance on lecture exams, relationships were only significant in the Fall semester and were sometimes counterintuitive. Although benefits of the course were offset somewhat by the inefficiency of voluntary recruitment, we demonstrate the effectiveness our course, which is inexpensive to implement and has advantage of providing pedagogical experience to future educators.

INTRODUCTION

For the United States to remain a global leader in science and technology, our educational system needs to increase the number of graduates in science, technology, engineering, and mathematics (STEM) fields by ~100,000 per year over the next decade (President's Council of Advisors on Science and Technology [PCAST], 2012). Among the most effective and economically feasible methods to overcome this shortage is to increase the retention rate of students in STEM majors (PCAST, 2012). Attrition in STEM majors is largely a result of poor performance in introductory classes (Chen, 2013). Many of these performance issues can be attributed to underpreparation for college-level classes, as many students have underdeveloped reasoning skills and poor study habits (Tomanek and Montplaisir, 2004). Additionally, lacking a sense of community in large introductory courses can increase attrition rates (Tinto, 1987;

Hoyle and Crawford, 1994), which may have an especially large impact on women and minorities, who collectively represent 70% of college students but only receive 45% of undergraduate STEM degrees (Chen, 2013).

This bottleneck, where underprepared first-year students are particularly at risk for leaving STEM majors due to poor performance, is critical to address if we hope to overcome the shortage of STEM graduates in the United States. With this goal in mind, universities have tried many methods of remediation, including encouraging student-led out-of-class study groups, restructuring introductory courses to focus on deep learning strategies, supplemental courses, single-session study habits lectures, and tutoring, each with varying degrees of success (Fullilove and Treisman, 1990; Tekian and Hruska, 2004; DeVoe *et al.*, 2007; Deslauriers *et al.*, 2011; Haak *et al.*, 2011; Rybczynski and Schussler, 2011; Buchwitz *et al.*, 2012).These interventions tend to fall into one of two categories: inexpensive to implement with limited effectiveness (DeVoe *et al.*, 2007; Rybczynski and Schussler, 2011) or expensive to implement with some measure of success at increasing participant performance (Richardson and Birge, 2000; Tekian and Hruska, 2004; Preszler, 2006). Supplemental courses tend to fall into the latter category, producing significant gains in student performance in introductory science courses but requiring a large time commitment from faculty for development and implementation of the course, as well as a substantial financial investment from the sponsoring department (Richardson and Birge, 2000; Preszler, 2006).

To mitigate these costs to the host department, we developed a supplemental course that is both designed and taught by graduate students (Sanjeevi, Callahan, and Fernandes, personal communication). We recognize that, for many institutions, graduate teaching assistants (TAs) may not be available to implement a course like this. However, we also believe this course could be effectively implemented by an advanced undergraduate student (T.D.H., personal observations). Indeed, undergraduate students have successfully led supplemental courses in a variety of STEM disciplines through the peer-led team-learning model. This learning model employs an experienced undergraduate student as a facilitator for a small group of students and produces marked gains in participant performance, retention, and attitude in associated classes (e.g., Lyle and Robinson, 2003; Hockings *et al.*, 2008; Loui and Robbins, 2008; Horwitz *et al.*, 2009; Hughes, 2011). Thus, our course, taught by either undergraduate or graduate students, costs less than traditional supplemental courses (requiring no faculty labor) and has the added benefit of developing the pedagogical skills of future educators (Huang *et al.*, 2013).

The strategy emphasized in our course was an adaptation of the study cycle (Cook *et al.*, 2013), a metacognition-based approach designed to promote learning through a straightforward series of steps that can be applied iteratively after every lecture. In particular, the study cycle stipulates that students preview materials before lecture, attend and attentively take notes, review notes and concepts before the following lecture, study in multiple short (yet intensive) bouts, and then assess their knowledge to identify areas in need of improvement. While this process provides an extremely attractive framework and has been previously show to boost student performance (Cook *et al.*, 2013), many students seem to be unfamiliar with how to accomplish various stages of the process. For example,

Cook *et al.* (2013) reported that, despite emphasis on the need for intensive study sessions employing active-learning strategies, only 24% of students reported using intensive study sessions, and self-assessment was absent from their list of commonly employed strategies. In the present study, we taught students to use outlines during the review phase of the study cycle and then to construct concept maps (without their notes) as a method to study and also to assess their understanding. Thus, we provided not only an effective process by which to approach studying but also a set of concrete tools by which to accomplish the more challenging steps. We hoped to teach students to avoid ineffective study skills, which often lead to only a superficial understanding of core concepts (McDermott *et al.*, 1994; Tomanek and Montplaisir, 2004). There is substantial evidence that outlines and concept maps significantly enhance student understanding, critical thinking, and exam performance (Novak, 1990; Okebukola, 1990; Daley *et al.*, 1999; Nesbit and Adesope, 2006).

Work based on previous iterations of our supplemental course suggests that it improves performance on lecture exams, but that gains in performance are somewhat modest (often averaging less than 10% on each exam compared with the exam scores of nonparticipants), though this is consistent with the results from faculty-led supplemental courses elsewhere (Sanjeevi *et al.*, personal communication; Richardson and Birge, 2000; Preszler, 2006). Further, we have observed that relatively few students volunteer to enroll and that not all of these participants improve in the associated introductory course (T.D.H., personal observations). However, it is not clear whether this is because the learning strategies we emphasize are ineffective for some students or because students actively resist enrolling or implementing the learning strategies we provide (Åkerlind and Trevitt, 1999). It is also unclear which elements of the supplemental course are most effective at improving participant performance in the lecture course, and whether there are benefits to enrollment in addition to increased exam scores, which are well-recognized to be imperfect metrics of student learning (Scouller, 1998; Rust, 2002; Gibbs *et al.*, 2005). Thus, we conducted the present study to examine 1) how effective our voluntary recruitment model was at enrolling the target cohort, 2) how the course impacted performance on lecture exams, 3) how the course impacted study habits and techniques emphasized in the intervention, and 4) whether there are particular study habits or techniques that are associated with large improvements on exam scores. This information will allow instructors in this supplemental course and in similar intervention efforts elsewhere to focus on the most effective methods of remediation. Ultimately, successfully implementing these interventions may have a large impact on the ongoing effort to increase student retention in STEM fields.

METHODS

Recruitment

Graduate TAs recruited students for the supplemental course (BIO 104) 1 week after the first exam in their introductory biology courses (BIO 115 and BIO 116 for Fall 2014 and Spring 2015, respectively). Recruitment occurred approximately 7 weeks into the semester and was accomplished by visiting lecture sections and giving a short presentation on the approach taken in the supplemental course. In particular, TAs encouraged

students earning ≤ 77% on exam 1 (a "C" on the grading scale) to enroll in the 8-week pass/fail sprint course for the remainder of the semester. The goal was to recruit 100% of this cohort. However, enrollment was entirely voluntary; TAs accepted any student who indicated interest via email. As such, the analyses presented here represent results from a self-selected group (students who chose to enroll in BIO 104), as opposed to a randomly selected cohort, and should be viewed in the context of this important distinction. However, given that voluntary enrollment models have been used at a variety of institutions (Belzer *et al.*, 2003; Preszler, 2006; Bail *et al.*, 2008; Kibble, 2009) and that there are potential ethical concerns with using students as experimental subjects to test the efficacy of new approaches, there is a need to use available data to glean as many useful lessons as possible.

Approach in the Sprint Course

Graduate TAs offered two sections of the course for both Fall 2014 and Spring 2015 semesters. The Fall sections were taught by a single TA, who returned in the Spring to teach one section and was joined by another TA, who taught the remaining section. Overall, BIO 104 used a metacognitive approach aimed at using content from the lecture course to help students learn how to more effectively engage with the subject matter in introductory biology. In other words, we focused on teaching study habits and skills using lecture content as common ground for discussions, as opposed to focusing on reviewing content per se. The course focused on two overlapping strategies aimed at improving student learning, which we refer to as study habits and study techniques. Our approach was an adaptation the study cycle, sensu Cook *et al.* (2013). With regard to study habits, we encouraged students to study early and often, in short but intensive bouts, and to regularly review course material. We also required students to design a study schedule (i.e., to create a set schedule with concrete dates and times dedicated to studying biology) and encouraged them to come up with a system to hold themselves accountable for these habits (e.g., by rewarding themselves for successful completion of study sessions). To encourage this behavior, TAs also provided an electronic study log (an Excel spreadsheet template) that allowed students to keep records on time spent studying, to calculate study hours on a weekly basis, and to view plots of their study habits over time.

The course introduced study techniques to students under the context of Bloom's taxonomy (Bloom, 1956), which organizes levels of cognition into a hierarchy, ranging from simple recall at the base of the pyramid to application, synthesis, and evaluation at the higher levels of the pyramid. Instructors stressed that, while students' previous academic experiences may not have required deep learning, college-level science courses require deep knowledge and lecture exams reflect this expectation through administration of challenging application and synthesis questions. TAs then introduced a study system emphasizing the use of outlines and concept maps as tools to engage the course materials. We stressed that, for this approach to promote success, students needed to apply it regularly. We strongly recommended that students make an outline after every lecture and then make a concept map from memory, revisiting their outlines as necessary to fill in gaps and identify areas in which further study was needed.

Outlines stressed the identification of the major and minor concepts from each lecture and encouraged students to organize (or reorganize) course content hierarchically. We asked students to include all relevant details, including definitions of all terms, descriptions of why each topic is important in broader contexts (i.e., in relation to previous lectures or topics), and examples when appropriate. TAs also strongly encouraged students to include diagrams, tables, and figures in their outlines. In the context of the study cycle (Cook *et al.*, 2013), outlines were meant to review recently introduced materials but also to serve as detailed study guides for study sessions immediately before exams. However, instructor-led, pre-exam study sessions were not a part of the sprint course. Concept maps emphasized asking questions to make connections between topics learned in lecture (e.g., for a biological process, where does the process happen, why does it happen, and what if it does not happen?). We suggested that students make concept maps from memory after completing outlines and revisit outlines as necessary to fill in weak areas. Thus, in the context of the study cycle (Cook *et al.*, 2013), concept maps were meant to encourage studying and self-assessment. TAs taught students to place major concepts, definitions, and examples in boxes or bubbles (nodes), and to use arrows to connect related topics to one another. TAs collected one outline and one concept map from each student each week, made photocopies for later use in analyses, and then allowed students to ask content-based questions about the material covered in their outlines and concept maps. This exercise provided limited guided review but primarily reiterated how intensive studying brings new questions to light and how students can use this approach to identify areas where they are struggling (i.e., to assess their knowledge, sensu Cook *et al.*, 2013).

Concept maps and outlines were evaluated and returned by the following week. Both outlines and concept maps were graded on 10-point scales consisting of two major categories: content (0–5 points) and organization (0–5 points). Content assessed whether outlines and concept maps included the breadth of material covered in each lecture, including all pertinent definitions, examples, and graphs when needed. Content also served to evaluate students on their ability to ask appropriate questions to fully develop their ideas. For outlines, organization evaluated students' ability to narrow down ideas hierarchically from major concepts to specific details. In concept maps, organization assessed students' ability to neatly and efficiently link ideas together in a meaningful way. TAs provided recommendations for improvement on all graded assignments.

Data Collection

Students were informed of the data collection and gave consent via forms approved by the Miami University Research Ethics and Integrity Office (IRB Exemption #01234e to T.D.H. and J.J.F.). For data analysis, we scored concept maps and outlines in triplicate using the guidelines listed earlier. TAs made three copies of the first and last outline and concept map for each student and randomly assigned these to a set of three graders (graduate TAs who agreed to help as a part of a graduate seminar on pedagogy), such that no grader received the same map or outline twice. To foster consistency among graders, we held a 1-hour training session in which graders were taught how to penalize various commonly encountered problems and were

allowed to compare comments on a standardized set of outlines and maps. When we were missing either the first or last outline or map from any individual student, we took the second assignment or the second-to-last assignment as substitutes. If this was not possible, then the student was dropped from the analysis. We also bolstered internal consistency among graders via a data-cleaning step. Briefly, we examined replicate scores for all students and endpoints and highlighted values that were ≥25% different from other replicates; for observations falling outside this range, a single, haphazardly selected author revisited the assignment in question and rescored the endpoint. We scored content and organization on the same scale as described earlier, because students received feedback throughout the semester in this format and used these metrics to gauge their own improvement. We also scored other metrics that we thought might be associated with the complexity and quality of concept maps, including number of nodes, number of arrows, number of linking words, number of terminal nodes (nodes that do not flow into another node), and quality of linking words (ordinal scale of 0–3, where 0 indicates absence of linking words altogether and ascending scores represent increasing quality of linking words). These are also features that have been examined in previous attempts at assessing the quality of concept maps (Luckie *et al.*, 2011).

We used data from pre and post surveys administered to students who enrolled in the supplemental course to assess whether and how they improved on the study habits we emphasized. Full pre and post surveys (included as Supplemental 2 in the Supplemental Material) included questions aimed at assessing how many hours per week students spent studying, how far in advance students began studying before their most recent exam, how often they made outlines, how often they made concept maps, how often they reorganized notes from class, and how regularly they reviewed materials. We also collected lecture exam scores for both students who ultimately enrolled in BIO 104 and their peers who did not enroll.

Analyses

Analyses were conducted separately for Fall and Spring data sets because TAs, students, and lecture materials all varied from semester to semester. For this reason, we also elected not to explicitly include time (Fall vs. Spring semester) as a factor in our analyses, but rather used parallel statistical approaches and compared results across semesters qualitatively.

To examine the effectiveness of voluntary recruitment, we calculated the number of students in the lecture course falling into the target cohort (≤77% on exam 1), the proportion of these students who enrolled in the sprint course, and the number and proportion of students who enrolled in the sprint course but were not a part of the target cohort (i.e., enrolled despite doing relatively well on the first exam). To elucidate whether students who enrolled in BIO 104 (the focal group) improved on lecture exams relative to their peers who did not enroll (the control group), we used the multivariate analysis of variance (MANOVA) approach for repeated measures (O'Brien and Kaiser, 1985). For elucidating when the groups diverged, repeated-measures ANOVAs were followed by planned, linear contrasts for each pair of subsequent exams. Effects in the model were exam (the repeated measure), treatment (focal vs. control), lecture section (section A specified as reference level), and

the interaction of lecture section and treatment; the response variable was exam score (% correct). To decrease the probability that observed changes in exam scores were simply artifacts of initial, pre-enrollment differences in exam scores, we also calculated normalized change in exam score (c, sensu Marx and Cummings, 2007). Normalized change calculates the ratio of actual gains or losses in exam scores to the maximum possible gain or loss, based on the pre score. It ranges from –1 to 1, with 0 indicating equality in exam scores, positive values indicating improvement from the pre-exam, and negative values indicating decreased scores. For each student, we calculated normalized change between exam 1 and each subsequent exam. We then used repeated-measures ANOVA to test whether treatment (focal vs. control), lecture section, or their interaction impacted normalized change over time.

Change in weekly study hours was analyzed using a paired *t* test. All other variables from surveys were scored on ordinal scales (1–4). To facilitate interpretation of reported changes in these variables, we defined using strategies after every class (4 on the scale, meaning "used the strategy after every class") as "acceptable," and all other levels as "unacceptable." We felt that this was appropriate, because the necessity of applying the strategies after every class was heavily emphasized during every BIO 104 class meeting. After partitioning pre and post scores for study habits, we tested for changes using McNemar's tests. Unfortunately, some students failed to fill out either pre or post surveys or left questions blank, so we were forced to drop these students from the analysis, and present here analyses based on the subsets of students for which we had sufficient data. To examine whether and how students improved on outlines and concept maps, we used average scores from three replicate graders to calculate change for each metric of outline and map quality, which we defined as: change = (score on final assignment – score on first assignment). Thus, positive values indicate increases in scores (i.e., improvement), while negative values indicate decreases in scores (i.e., diminishing quality). To test the statistical significance of observed changes, we used paired *t* tests.

Finally, we used multivariate regression models to explore whether and how changes in self-reported study habits (from pre and post surveys) and changes in the quality of outlines and concept maps (from retrospectively scored documents) scaled with changes in exam scores. The aim of these analyses was to use all available data to see whether changes in individual variables or subsets of variables were associated with an increase in performance on lecture exams. During the data-exploration phase of the study, we tried a variety of regression approaches, including principal components analysis (to address potential multicollinearity by combining variables into principal components), logistic regression with dummy coding for improvement or no improvement for all independent variables (to deal with potential nonlinearity and analyze improvement as a binary response), and random forest regression. None of these approaches fitted the data better than multiple regression with backward selection, which we elected to use in our analysis. To ensure the absence of severe multicollinearity among independent variables, we examined variance inflation factors (VIFs) and did not interpret models in which any individual factor had a VIF >10 (Hair *et al.*, 1995); this was never the case in final models.

TABLE 1. Summarized recruitment data from Fall 2014 and Spring 2015

Semester (lecture)	Lecture enrollment	Students in target cohort	Number enrolled in BIO 104	% of target cohort enrolled in BIO 104	Number enrolled in BIO 104 despite being outside target cohort
Fall 2014 (BIO 115)	511	307	54	16.94	2
Spring 2015 (BIO 116)	550	267	44	11.30	10

Unfortunately, because surveys were voluntary, not all students completed them. Rather than dropping survey data from these analyses altogether, we present two separate models for each semester, or a total of four models. For each semester, one model included both survey data and outline/concept map data as predictors, while the other included only outline and map variables as predictors and contained a complete sample of BIO 104 participants. In all cases, we selected initial (full) models to maximize predictive power by examining all possible combinations of predictors and selecting the model that maximized adjusted R^2. We then performed backward, stepwise selection to sequentially remove terms with the smallest contribution to the model and stopped when all terms were significant at $\alpha = 0.10$. The one exception to this method was for the model using all of the data from Spring 2015, for which we had sufficient survey data from only 17 of the 44 focal students. To avoid starting with a full model based on variables with excessive missing observations, we instead used forward, stepwise selection for this model, adding terms sequentially until no additional variables met the criterion of $\alpha = 0.15$.

RESULTS

Fall 2014

There were 511 students in the Fall 2014 (BIO 115) data set, representing sample sizes of 54 and 457 in the focal and control groups, respectively (Table 1). While 307 students fell into the target cohort (i.e., scored $\leq 77\%$ on exam 1), only 52 enrolled in the sprint course (Table 1). These students were joined by two students who joined the sprint course despite scoring $>77\%$ on exam 1 (Table 1). This means that only 16.94% of the target cohort enrolled, but also that only 3.07% of students who did enroll were not from the target cohort (Table 1).

Repeated-measures ANOVA revealed that exam scores changed over time, exam scores over time differed among lecture sections, and scores over time differed among focal and control students, but there was no interaction between treatment (focal vs. control) and lecture section over time (Table 2 and Figure 1). Focal students began with lower scores than their peers who did not enroll (exam 1), but this difference disappeared after enrollment in BIO 104; planned contrasts revealed that focal students improved relative to control students between exam 1 and exam 2 ($F(1, 504) = 30.51$, $p < 0.0001$; Figure 1) but that relative performance did not change between exams 2 and 3 ($F(1, 504) = 0.09$, $p = 0.7604$; Figure 1) or between exams 3 and 4 ($F(1, 504) = 0.90$, $p = 0.3444$; Figure. 1). We found that normalized change in exam score (c) was variable through time and varied depending on lecture section, but that there was no effect of treatment on normalized exam scores over time, nor was there any evidence of a three-way interaction (Table 2 and Figure 1B). However, between-subjects tests revealed that, across all pairs of exams, normalized change in scores was higher for focal students than control students ($F(1, 504) = 18.69$, $p = 0.0001$, Figure 1B),

TABLE 2. Repeated-measures ANOVA table for exam scores and normalized change in exam scores over time (exam), as well effects of treatments (focal vs. control), lecture section, and their interaction over time

Response	Source of variation	F	Numerator df	Denominator df	p^a
Exam score, Fall 2014	Exam	93.19	3	502	**<0.0001**
	Exam × lecture section	14.34	6	1004	**<0.0001**
	Exam × treatment	12.76	3	502	**<0.0001**
	Three-way interaction	0.52	6	1004	0.7886
Exam score, Spring 2015	Exam	9.77	2	534	**<0.0001**
	Exam × lecture section	16.15	4	1068	**<0.0001**
	Exam × treatment	5.62	2	534	**0.0038**
	Three-way interaction	0.43	4	1068	0.7889
Normalized change (c), Fall 2014	Exam	105.99	2	503	**<0.0001**
	Exam × lecture section	14.04	4	1006	**<0.0001**
	Exam × treatment	0.33	2	503	0.7197
	Three-way interaction	0.78	4	1006	0.5384
Normalized change (c), Spring 2015	Exam	0.47	1	535	0.4911
	Exam × lecture section	21.27	2	535	**<0.0001**
	Exam × treatment	0.34	1	535	0.5604
	Three-way interaction	0.56	2	535	0.5704

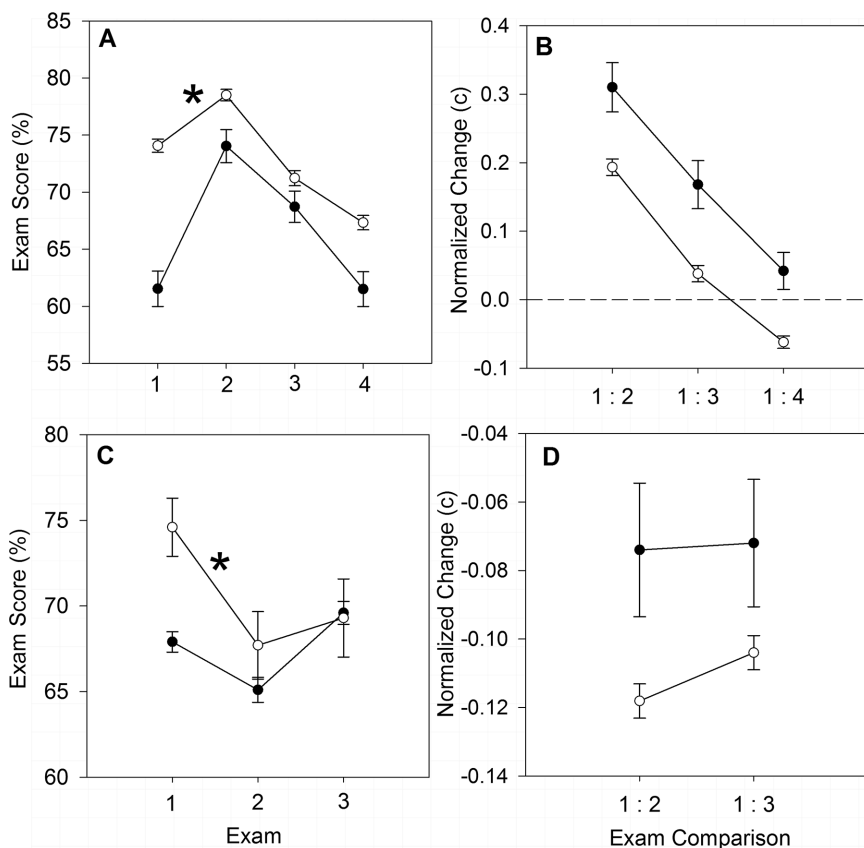[a]Bolded p values are significant at $\alpha = 0.05$.

FIGURE 1. Exam scores (A and C) and normalized change in scores between exam 1 and subsequent exams (B and D) for focal students (closed circles) and control students (open circles) in Fall 2014 (A and B) and Spring 2014 (C and D). Asterisks indicate significant planned contrasts for the effects of treatment between each pair of subsequent exams (i.e., indicate when treatment × exam interactions occurred). Error bars represent ± 1 SE.

normalized change differed by lecture section ($F_{(2, 504)} = 9.19$, $p < 0.0001$), but the effect of treatment did not vary by lecture section ($F_{(2, 504)} = 0.47$, $p = 0.6229$).

Students made improvements in some, but not all, study habits emphasized in the sprint course. Focal students reported an increase of 2.46 ± 0.46 (mean ± SE) study hours per week, which is more than would be expected by chance (paired *t* test, $t_{36} = 5.39$, $p < 0.0001$). Students also reported increases in outline usage, but relatively large numbers of students failed to improve study habits to levels defined as acceptable. Outline use changed substantially from the beginning to the end of the sprint course (McNemar's $\chi^2 = 7.11$, $p = 0.0077$; Table 3A), as did the regularity of note reorganization (McNemar's $\chi^2 = 6.67$, $p = 0.0098$; Table 3B) and concept map usage (McNemar's $\chi^2 = 5.82$, $p = 0.0159$; Table 3C). However, regularity of reviewing did not change significantly from the beginning to the end of the sprint course (McNemar's $\chi^2 = 2.72$, $p = 0.0990$; Table 3D). The trend was very similar across all three study habits for which we detected changes; while some students changed from unacceptable to acceptable patterns of outline use, note reorganization, and concept map use, two to three times as many students were still not using these strategies in an acceptable manner by the end of the course, despite heavy emphasis (Table 3).

TABLE 3. Two-by-two contingency tables, showing the number of students using study habits in an acceptable (Yes) or unacceptable (No) manner from the beginning (week 1; rows), to the end (week 7; columns) of the sprint course[a]

Fall 2014				Spring 2015			
A. Outline usage		Week 7		**E. Outline usage**		Week 7	
	Acceptable time	No	Yes		Acceptable time	No	Yes
Week 1	No	33	9	**Week 1**	No	14	6
	Yes	0	1		Yes	0	0
B. Note reorganization		Week 7		F. Note reorganization		Week 7	
	Acceptable time	No	Yes		Acceptable time	No	Yes
Week 1	No	27	13	**Week 1**	No	11	3
	Yes	2	0		Yes	3	3
C. Map usage		Week 7		G. Map usage		Week 7	
	Acceptable time	No	Yes		Acceptable time	No	Yes
Week 1	No	32	10	**Week 1**	No	17	3
	Yes	1	0		Yes	0	0
D. Regular reviewing		Week 7		**H. Regular reviewing**		Week 7	
	Acceptable time	No	Yes		Acceptable time	No	Yes
Week 1	No	16	13	**Week 1**	No	4	10
	Yes	5	9		Yes	1	4

[a]Data are from pre and post surveys from Fall 2014 (A–D, left) and Spring 2015 (E–H, right). Bolded items showed significant changes over the course of the semester (McNemar's tests, α = 0.05).
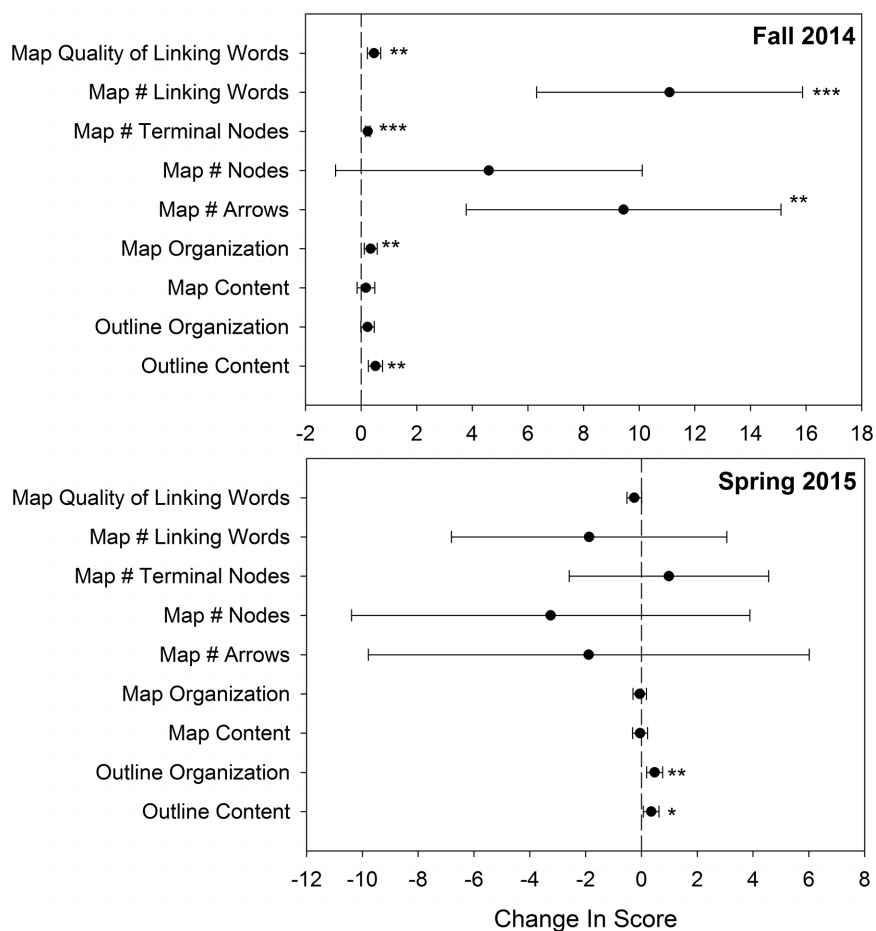
FIGURE 2. Change in average scores from three replicate graders for scored components of outlines and concept maps. The vertical, dotted line indicates no net change. Error bars indicate 95% confidence intervals, while asterisks indicate significance from paired *t* tests at *p* = 0.05 (*), 0.01 (**), and 0.0001 (***).

variance in change in exam scores and included the intercept, lecture section C, map organization, map content, number of nodes in concept maps, and number of arrows in concept maps (Table 4A). With the exception of number of nodes, all variables in this final model were significant at $\alpha = 0.05$ (Table 4A). However, somewhat counterintuitively, slopes for lecture section, map content, and number of arrows in concept maps were negative (Table 4A). This means that students who were in lecture C tended to improve less on exams relative to the reference level (section A), while among all lecture sections, decreases in map content scores and in number of nodes were associated with improvement on exams.

For the data set including only outline and concept map variables (but maximizing sample size), multivariate regression indicated a weak but significant association between change in outline and map scores and change in exam scores (Table 4B). The full model after optimization of adjusted R^2 included the intercept and eight predictor variables and explained a significant portion of variance in change in exam scores (Supplemental Table S1). The best model after backward selection indicated a significant association with change in exam scores and explained 21.85% of the variance (Table 4B). With the exception of the intercept, all predictors in the final model were significant at $\alpha = 0.05$ (Table 4B). However, as we found in models including survey data, some predictors carried negative slopes. Decreases in outline content scores were associated with increases in exam scores, and decreases in the number of linking words were associated with increases in exam scores (Table 4B). Conversely, increases in number of nodes and number of terminal nodes were associated with exam score improvement, which was consistent with our predictions (Table 4B).

Spring 2015

There were 550 students in the Spring 2015 (BIO 116) data set, representing sample sizes of 44 and 506 in the focal and control groups, respectively (Table 1). While 267 students fell into the target cohort (i.e., scored ≤ 77% on exam 1), only 34 enrolled in the sprint course (Table 1). These students were joined by a group of 10 students who joined the sprint course despite scoring > 77% on the first exam (Table 1). This means that only 11.3% of the target cohort enrolled, and that 22.7% of students who did enroll were not from the target cohort (Table 1).

Repeated-measures ANOVA revealed that exam scores changed over time, that exam scores over time differed among lecture sections, that scores over time differed among focal and control students, but that there was no interaction between treatment (focal vs. control) and lecture section over time

Students improved on seven of the nine metrics we examined for the quality of outlines and concept maps (Figure 2). Student outlines improved for content scores ($t_{53} = 3.96$, $p = 0.0002$; Figure 2), but showed only marginal improvement on organization ($t_{53} = 1.87$, $p = 0.0670$; Figure 2). In contrast, for concept maps, students improved on organization ($t_{53} = 2.89$, $p = 0.0056$; Figure 2), but not on content ($t_{53} = 1.05$, $p = 0.2972$; Figure 2). With the exception of number of nodes, which did not change from the beginning to the end of the intervention ($t_{53} = 1.63$, $p = 0.1087$), students improved on all other metrics of concept map quality that were analyzed (number of arrows: $t_{53} = 3.27$, $p = 0.0019$; number of terminal nodes: $t_{53} = 5.55$, $p < 0.0001$; number of linking words: $t_{53} = 4.55$, $p < 0.0001$; quality of linking words: $t_{53} = 3.82$, $p = 0.0004$; Figure 2).

For the data set including both study habits and outline and concept map variables, multivariate regression identified a significant association between the best subset of predictors and change in exam scores (Table 4A). The full model after optimization of adjusted R^2 values included an intercept and nine predictor variables and predicted a significant portion of the variance in change in exam scores (Supplemental Table S1). After backward selection, the best model explained 40.08% of the

TABLE 4. Summary of best models for Fall 2014 when survey data were included (A) or excluded (B)[a]

Variable	df	β	SE β	Standardized β	t	p	VIF
A. All data, best model							
Intercept	1	3.35	1.99	0.00	1.69	0.0994	0
Lecture section (C)	1	**−8.00**	**3.11**	**−0.41**	**−0.26**	**0.0143**	**1.65**
Map organization	1	**5.99**	**2.22**	**0.54**	**2.70**	**0.0102**	**2.50**
Map content	1	**−3.85**	**1.71**	**−0.50**	**−2.25**	**0.0303**	**3.18**
Map number of nodes	1	0.22	0.12	0.47	1.76	0.0863	4.61
Map number of arrows	1	**−0.20**	**0.10**	**0.46**	**−2.06**	**0.0463**	**3.18**
Global model statistics: $F(5, 38) = 5.08$, $p = 0.0012$, $R^2 = 0.4008$, adjusted $R^2 = 0.3220$							
B. Survey data excluded, best model							
Intercept	1	−0.15	1.68	0.00	−0.09	0.9287	0
Outline content	1	**−3.19**	**1.35**	**−0.31**	**−2.36**	**0.0222**	**1.05**
Map number of nodes	1	**0.42**	**0.15**	**0.87**	**2.76**	**0.0081**	**6.25**
Map number of linking words	1	**−0.48**	**0.21**	**−0.85**	**−2.32**	**0.0247**	**8.55**
Map number of terminal nodes	1	**24.30**	**8.33**	**0.76**	**2.92**	**0.0053**	**4.34**
Global model statistics: $F(4, 49) = 3.42$, $p = 0.0151$, $R^2 = 0.2185$, adjusted $R^2 = 0.1547$							

[a]Predictors that were significant at $p \leq 0.05$ are bolded. Overall model statistics are provided below each model.

(Table 2 and Figure 1). The observed trend was similar to the trend observed in Fall 2014, in that focal students were initially underperforming relative to peers who did not enroll, but this difference disappeared following enrollment in BIO 104; planned contrasts revealed that focal students improved relative to their peers who did not enroll between exams 1 and 2 ($F(1, 535) = 6.78$, $p = 0.0095$; Figure 1) but that the trajectories for focal and control students between exams 2 and 3 were indistinguishable ($F(1, 535) = 0.10$, $p = 0.7523$; Figure 1). We found that normalized change in exam score (c) was consistent through time, varied depending on lecture section, but that there was no effect of treatment on normalized exam scores over time, nor was there a three-way interaction (Table 2 and Figure 1D). However, between-subjects tests revealed that, across all pairs of exams, normalized change in scores was higher for focal students than control students ($F(1, 535) = 5.49$, $p = 0.0152$, Figure 1D), normalized change differed by lecture section ($F(2, 535) = 4.22$, $p = 0.0195$), and the effect of treatment did not vary by lecture section ($F(1, 535) = 0.13$, $p = 0.8762$).

Students made improvements in some but not all study habits emphasized in the sprint course. Focal students reported an increase of 2.42 ± 0.67 (mean $\pm$ SE) study hours per week by the end of the semester, which was more than would be expected by chance alone ($t_{18} = 3.62$, $p = 0.0200$). Students also reported increased outline usage, with six students reporting using outlines after every lecture, despite not having done so at the beginning of the semester (McNemar's $\chi^2 = 6.0$, $p = 0.015$; Table 3E); however, this represents only 30% of the students for whom we had survey data. We also detected an increase in self-reported regularity of reviewing, with 10 students who began the course without regular review reporting reviewing after every class by the end of the semester (McNemar's $\chi^2 = 7.36$, $p = 0.0059$; Table 3H). However, as was the case with outline usage, only a subset of students reported making the desired changes, despite this being a central component of our strategy. We did not detect substantial changes in note reorganization (McNemar's $\chi^2 = 0$, $p = 0.6563$; Table 3F) or in use of concept maps (McNemar's $\chi^2 = 9.0$, $p = 0.1250$; Table 3G).

Analysis of outline and map variables for Spring 2015 revealed that students improved on outlines but not on concept maps (Figure 2), which was a direct contrast to the pattern observed among Fall BIO 104 students. Student outlines improved over the duration of the sprint course both in terms of organization ($t_{43} = 3.15$, $p = 0.0029$; Figure 2) and content ($t_{43} = 4.17$, $p < 0.0001$; Figure 2). Conversely, we did not observe changes in scores on concept maps in any metric analyzed (Figure 2), including content ($t_{43} = 0.72$, $p = 0.4747$), organization ($t_{43} = 0.56$, $p = 0.5785$), number of arrows ($t_{38} = 1.15$, $p = 0.2584$), number of nodes ($t_{42} = 0.99$, $p = 0.3270$), number of terminal nodes ($t_{41} = 1.80$, $p = 0.0788$), number of linking words ($t_{41} = −0.11$, $p = 0.9112$), or in quality of linking words (t_{41}, $p = 0.2438$).

For the analysis including all data, multivariate regression identified a significant relationship between the best subset of predictors and change in exam scores (Table 5). The best model after forward, stepwise selection explained 65.14% of the variance in change in exam scores and included effects of the intercept, number of arrows on concept maps, number of linking words on concept maps, and regularity of reviewing; all of the predictors in this final model were significantly associated with change in exam score (Table 5). However, while increases in the number of linking words on concept maps were associated with increases on lecture exam performance, negative coefficients for number of arrows and for regularity of reviewing indicated that decreases in number of arrows and in the regularity of review were associated with increased exam scores, which was counter to our predictions. An important caveat is that this model was based on responses from only 17 of the 44 total focal students and thus represents only 38.6% of the focal group at large. For the analysis including only data from outlines and concept maps (and thus maximizing sample size), multivariate regressions revealed that both the full (most predictive) model after optimization of adjusted R^2 and the best model after backward elimination failed to predict change in exam scores and explained only a small proportion of its variance (Supplemental Table S2).

TABLE 5. Summary of best model Spring 2015 data when survey responses were included (all data)[a]

Variable	df	β	SE β	Standardized β	t	p	VIF
All data, best model							
Intercept	1	0.100	0.024	0.000	4.260	0.0009	0
Map number of arrows	1	−0.004	0.001	−1.149	−3.850	0.0020	3.32
Map number of linking words	1	0.004	0.002	0.705	2.530	0.0253	2.90
Regular reviewing	1	−0.125	0.030	−0.798	−4.190	0.0011	1.35
Global model statistics: $F(3, 16) = 8.10$, $p = 0.0027$, $R^2 = 0.6514$, adjusted $R^2 = 0.5170$							

[a]Predictors that were significant at $p \leq 0.05$ are bolded. Global model statistics are provided at the bottom of the table.

DISCUSSION

Many colleges and universities are exploring options to decrease attrition rates in introductory STEM courses. Evidence shows that supplemental courses can significantly enhance high-risk students' performance and increase their chance of success and ultimate retention in STEM; however, many common methods of remediation are either costly to implement or have limited effectiveness (Maloof and White, 2005; DeVoe et al., 2007; Rybczynski and Schussler, 2011). In this study, we investigated the impacts of a supplemental course offered at Miami University. A key feature of this supplemental course is that it was taught by graduate student TAs and thus was inexpensive for the university to implement. Further, it was designed to emphasize outlines and concept maps as concrete tools with which to accomplish the more demanding steps of the study cycle (i.e., reviewing, studying, and self-assessment). Our analyses of this course were designed to determine which individual remediation strategies were most effective at enhancing student performance in the associated introductory biology courses. While the students who participated in our supplemental courses were a self-selected cohort of the larger body of students who received ≤77% on their first lecture exam, we assert that it is still valuable to investigate how supplemental instruction influences self-selected groups, given that many institutions are relying on or have relied on voluntary supplemental instruction (e.g., single-session study habits lectures, tutoring, review sessions, and voluntary recitations) to address poor performance in introductory courses (Belzer et al., 2003; Preszler, 2006; Bail et al., 2008; Kibble, 2009).

In the broadest sense, the supplemental course was successful. Participant exam scores began significantly below the class average and rose to become statistically indistinguishable from nonparticipants' scores on subsequent exams. These results are remarkably consistent with previous findings in this course, spanning eight semesters (Sanjeevi et al., personal communication). Interestingly, the gains in participant performance are realized by the first exam administered after enrollment in the supplemental course, which is the second exam in the introductory biology course overall. In other words, participant scores become statistically indistinguishable from nonparticipant scores within only the first few weeks of the supplemental course, but there are no additional gains on subsequent exams. Analysis of normalized change in scores, which takes into account how much students' scores change relative to the maximum possible change, reiterated this interpretation. Focal students had larger normalized gains as early as exam 2 and then maintained this difference. Thus, it is unlikely that observed effects are simply an artifact of focal students starting the semester with more room for improvement.

While we do not have the data to measure potential differences in motivation between our self-selected cohort and their peers in the present study, other authors have made estimates. For example, Arendale (1997) and Ramirez (1997) established a motivational control, consisting of students who expressed interest in supplemental instruction but were unable to enroll in supplemental courses. Arendale (1997) used the percent of students earning "A's" or "B's" as their metric of performance and found that ~45.1% of the difference in percentages between the control and focal group was attributable to differences in motivation. Similarly, Ramirez (1997) used final course grades as their metric of performance and found that ~59% of the difference between focal and control students was attributable to motivational differences. If we assume that 60% of our treatment effect is attributable to motivational differences and examine treatment differences in change in scores from exam 1 to exam 2 (when the treatment effect was most pronounced), the difference as compared with the control would be reduced from 8.15% to 4.89% in Fall 2014 and from 4.90% to 2.94% in Spring 2015. Thus, while we do not have the data to quantitatively test for the impact of motivation, it does appear that there would be at least some residual benefit for our students when we take into account motivational effects estimated elsewhere. Furthermore, Sanjeevi et al. (personal communication) used surveys to compare changes in attention and persistence for focal and control students in a previous iteration of BIO 104 and found that focal students gained more over the course of semester in terms of both attention and persistence. If our students improved more in terms of motivation than their peers who did not enroll, this might be one mechanism that can explain how our course improves student performance on exams.

It is well recognized that exam grades are not an all-encompassing metric of student performance (Scouller, 1998; Rust, 2002; Gibbs et al., 2005) and may not capture all learning outcomes of BIO 104. As such, we investigated the impact the supplemental course had on self-reported study habits through pre and post surveys, as well as changes in quality of outlines and concept maps. The supplemental course instructors emphasized the study cycle (Cook et al., 2013) and frequent reviewing and reorganization of class notes by creating concept maps and outlines. Changes in study habits were strikingly different across the two semesters in our study. In the Fall of 2014, students reported increased usage of outlines and maps and increased regularity of note reorganization, while in the Spring, improvements were only reported for outline usage and regularity of reviewing. In agreement with the observation that few students reported increases in map usage during the Spring of 2015, focal students in the Spring cohort did not improve on

any metric of concept maps analyzed, with the exception of "map organization." This is in stark contrast to the trend for outline and concept map variables in the Fall cohort, in which seven of nine metrics were improved. Together, these observations suggest, perhaps intuitively, that students must regularly use outlines and concept maps to improve. Further, it is intriguing that the general trend was for Fall students to improve more dramatically than the Spring cohort on a wider breadth of both study habits and metrics associated with outlines and maps. The vast majority of biology majors at our institution take BIO 115 in their first Fall semester, followed by BIO 116 in their second semester. Thus, our data reiterate the importance of bridge-type courses, offered during the summer or early in the first semester, as this is a period when students are experiencing new academic challenges and searching for ways to study more effectively (Martin and Arendale, 1992; Tinto, 2001).

Although we detected statistically significant improvements in study habits across both semesters (Table 3), a striking majority of students did not shift into what we considered to be acceptable patterns of behavior. For example, while outline usage increased across both cohorts, only 23% and 30% of focal students were using outlines after every lecture for the Fall and Spring cohorts, respectively. Outlines were a centerpiece of the study cycle (Cook *et al.*, 2013) as we presented it to students. We heavily emphasized making an outline after every class as a way to review, study, and self-assess, but it is not clear why such a small proportion of students were able or willing to apply these stages of the study cycle. Because we used a voluntary recruitment strategy, we assumed that motivation was high among focal students and that simply introducing the strategies to students and providing limited feedback (e.g., returning comments on one outline and one concept map per week) would be sufficient to spur the necessary behavioral changes. It may be that making the requirements in the sprint course more rigorous, perhaps by increasing the number of graded outlines and concept maps or by changing the course from "Pass/Fail" to a more traditional "ABC" grading system would provide the incentives necessary to change behaviors. However, this approach would require substantially more investment from instructors and also run the risk of devaluing what we consider to be two central pillars of successful scholarship: strong, internal motivation to succeed and the desire to learn as opposed to earn passing scores. The answers to this issue might be best solved on a case-by-case basis, depending on the institution's priorities. For example, focusing on increasing student performance and retention in the major seems to favor altering the course structure to provide more feedback and greater incentive to adopt the emphasized strategies, while trying to provide the tools for success to motivated students might favor the supplemental course design we present here.

In general, our ability to explain change in exam scores using changes in study habits and techniques was limited. While we did detect statistically significant associations between changes in study habits and techniques and changes in exam scores, these relationships were somewhat weak (in terms of variance explained), and the identities and signs for individual coefficients were inconsistent among semesters. If there were a single variable strongly associated with change in exam scores, we would have expected for it to remain in best models across semesters, for the slope to be substantially larger than other significant slopes, and for the sign to remain consistent. This was not the case here, as variables remaining in best models were largely inconsistent across semesters, and no individual coefficients stood out with particularly large slopes. Thus, our primary conclusion from this set of analyses is that there is no single habit or technique that consistently underlies the impact of our course on exam scores, nor is there a consistent subset of predictors.

We acknowledge that our multiple regression analyses may have lacked the statistical power to isolate subtle impacts of individual regression coefficients, if this were the true underlying pattern. We endeavored to reduce technical sources of error via training of concept map and outline evaluators, using means from triplicate graders and then performing an exhaustive data-cleaning step to remove aberrant scores, but it is likely that outline and concept map data are highly variable by nature. There are many ways to effectively build outlines and concept maps, and while their pedagogical merits have been repeatedly demonstrated (Novak, 1990; Okebukola, 1990; Daley *et al.*, 1999; Nesbit and Adesope, 2006), a major challenge is how to score them and provide consistent feedback (Luckie *et al.*, 2011; Dowd *et al.*, 2015). Given that first-year students come from diverse backgrounds, enter introductory courses with highly variable levels of prior preparation, and vary substantially in maturity (Tinto, 2001), it is perhaps expected that different students would benefit from supplemental instruction in different ways. For example, Haak *et al.* (2011) found that increased structure and active learning disproportionately benefited underrepresented minority (URM) students, which the authors attributed to differences in student backgrounds (i.e., while non-URM students were accustomed to active learning, URM students were not, so there was a disproportionate benefit for the latter). A similar mechanism could have been operating here, but as opposed to a cohort selected based on predefined criteria (e.g., grade point average [GPA] and Scholastic Aptitude Test [SAT] Verbal scores; Haak *et al.*, 2011), our cohort self-selected through voluntary recruitment, potentially increasing the diversity of student backgrounds relative to a recruitment mechanism based on quantitative criteria. To understand how students respond to supplemental instruction in the context of voluntary recruitment, it may be necessary to differentiate among groups of students based on pre-existing metrics indicative of prior experience, such as socioeconomic background, ACT scores, or high school GPA. In the absence of some ability to subdivide students based on their specific needs, it appears unlikely that studies with relatively small focal cohorts will have the statistical power to isolate individual features associated with improvement, particularly if these effects are subtle.

Even though voluntary recruitment approaches are regularly used at diverse institutions (Belzer *et al.*, 2003; Bail *et al.*, 2008; Kibble, 2009), our data cast some doubt on their ability to effectively attract the students who need them. We found that the voluntary enrollment strategy failed to recruit many of the high-risk students (those who received a 77% or lower on the first exam). Only 17% of high-risk students enrolled in our 8-week sprint course, suggesting that other recruitment methods might be necessary to more effectively help a greater number of students. We can only speculate as to the mechanisms preventing efficient recruitment of the target cohort, but it is possible that a sense of stigma surrounding the course

played a role, as has been reported elsewhere (Somers, 1988). Some institutions have addressed this issue by building supplemental instruction into the mandatory structure of the lecture course (e.g., Preszler, 2009). Other institutions have overcome the problem by requiring only at-risk students to participate in a remediation program, a strategy that relies on successfully identifying at-risk students before they start the introductory course, using high school GPA, SAT scores, and socioeconomic status as the primary predictors (Hodges and White, 2001; Hensen and Shelley, 2003; Lotkowski *et al.*, 2004). This can be problematic because it can be difficult to identify the measures that best predict performance in introductory biology courses. One possible remedy for this challenge would be to conduct placement exams before entrance into introductory STEM courses. However, such obligate remediation strategies might prevent students from coming to the conclusion that they need to change their study habits on their own, because they would be receiving help from the outset of the introductory courses. Without having the opportunity to realize their need for remediation, students might not value the strategies that are being taught, which is likely to increase their resistance to the ideas (Arendale, 1994). This idea is corroborated by the instructors of our supplemental course, who report that students often admit that they would not have considered the remediation necessary without first struggling on lecture exam 1 (T.D.H. and K.A., personal observations). Although the use of risk assessment to track students into supplemental courses may be difficult to effectively implement, it could be that such methods are needed to increase the recruitment of high-risk students and ultimately necessary to address the attrition issue.

To conclude, our assessment of the influence of a supplemental course paired with the introductory biology courses at Miami University suggests that remediation can be successful with graduate student instructors, helping to make such remediation sustainable at most institutions due to low cost. Further, we believe that our approach could be successfully implemented using advanced undergraduate students as instructors, making similar solutions available to undergraduate-only departments and institutions. Additionally, our analysis of the effects our course had on participant study habits and techniques revealed that a small portion of participants (23–30%) fully adopted the study habits we emphasized and that creating concept maps and outlines has a modest effect on lecture course exam scores. However, we suspect that the modest correlations between concept map and outline metrics and exam scores are indicative of the different needs and learning styles of individual students. In other words, some students might benefit more from the more visual construction of a concept map, while others might find greater benefit in reorganizing their notes into outlines. These results may also be influenced by the difficulty of effectively grading these assignments and the large portion of students who did not adopt the strategies. Even so, our analyses suggest that presenting a variety of research-supported study techniques may be the best method to help the largest number of struggling students. Finally, our results highlight the challenge of getting students the help they need. The voluntary recruitment strategy for this study caused this effort to reach only ~17% of students who struggled with the first lecture exam. This shortcoming may be addressed by obligatory remediation for at-risk students, though it is not clear what effects this strategy will have on student moti-

vation and buy-in regarding the supplemental course. To gain a clearer idea of how these various methods affect success and long-term retention in STEM majors, future studies should examine retention rates among self-selected participants and nonparticipants in voluntary enrollment courses and compare those results against retention rates from obligatory remediation strategies. Despite the difficulties presented by the diversity of student backgrounds, needs, and motivations, we demonstrate that a graduate student–taught supplemental course that teaches a variety of study skills and techniques can effectively enhance undergraduate performance in introductory biology courses.

REFERENCES

Åkerlind, G. S., & Trevitt, C. A. (1999). Enhancing self-directed learning through educational technology: When students resist the change. *Innovations in Education and Training International 36*(2), (pp. 96–105).

Arendale, D. (1997). Supplemental instruction (SI): Review of research concerning the effectiveness of SI from the University of Missouri-Kansas City and other institutions from across the United States. Paper presented at Proceedings of the 17th and 18th Annual Institutes for Learning Assistance Professionals, University of Arizona (pp. 1–25).

Arendale, D. R. (1994). Understanding the supplemental instruction (SI) model. *New Directions for Teaching and Learning*, *60*(4), 11–22.

Bail, F. T., Zhang, S., & Tachiyama, G. T. (2008). Effects of a self-regulated learning course on the academic performance and graduation rate of college students in an academic support program. *Journal of College Reading and Learning*, *39*(1), 54–73.

Belzer, S., Miller, M., & Hoemake, S. (2003). Concepts in biology: A supplemental study skills course designed to improve introductory students' skills for learning biology. *American Biology Teacher*, *65*(1), 30–40.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals*. New York: Longman.

Buchwitz, B. J., Beyer, C. H., Peterson, J. E., Pitre, E., Lalic, N., Sampson, P. D., & Wakimoto, B. T. (2012). Facilitating long-term changes in student approaches to learning science. *CBE—Life Sciences Education*, *11*(3), 273–282.

Chen, X. (2013). *STEM attrition: College students' paths into and out of STEM fields* (Statistical Analysis Report NCES 2014-001). Washington, DC: National Center for Education and Statistics.

Cook, E., Kennedy, E., & McGuire, S. Y. (2013). Effect of teaching metacognitive learning strategies on performance in general chemistry courses. *Journal of Chemical Education*, *90*(8), 961–967.

Daley, B. J., Shaw, C. A., Balistrieri, T., Glasenapp, K., & Piacentine, L. (1999). Concept maps: A strategy to teach and evaluate critical thinking. *Journal of Nursing Education*, *38*(1), 42–47.

Deslauriers, L., Schelew, E., & Wieman, C. (2011). Improved learning in a large-enrollment physics class. *Science*, *332*(6031), 862–864.

DeVoe, P., Niles, C., Andrews, N., Benjamin, A., Blacklock, L., Brainard, A., … Osgood, M. (2007). Lessons learned from a study-group pilot program for medical students perceived to be "at risk." *Medical Teacher*, *29*(2–3), e37–e40.

Dowd, J. E., Duncan, T., & Reynolds, J. A. (2015). Concept maps for improved science reasoning and writing: Complexity isn't everything. *CBE—Life Sciences Education*, *14*(4), ar39.

Fullilove, R. E., & Treisman, P. U. (1990). Mathematics achievement among African American undergraduates at the University of California, Berkeley: An evaluation of the mathematics workshop program. *Journal of Negro Education*, *59*(3), 463–478.

Gibbs, G., Simpson, C., Gravestock, P., & Hills, M. (2005). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education*, *1*, 3–31.

Haak, D. C., HilleRisLambers, J., Pitre, E., & Freeman, S. (2011). Increased structure and active learning reduce the achievement gap in introductory biology. *Science*, *332*(6034), 1213–1216.

Hair, J. F. Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis* (3rd ed.). New York: Macmillan.

Hensen, K. A., & Shelley, M. C. (2003). The impact of supplemental instruction: Results from a large, public, Midwestern university. *Journal of College Student Development*, *44*(2), 250–259.

Hockings, S. C., DeAngelis, K. J., & Frey, R. F. (2008). Peer-led team learning in general chemistry: Implementation and valuation. *Journal of Chemical Education*, *85*(7), 990–996.

Hodges, R., & White, W. G. (2001). Encouraging high-risk student participation in tutoring and supplemental instruction. *Journal of Developmental Education*, *24*(3), 2–7.

Horwitz, S., Rodger, S. H., Biggers, M., Binkley, D., Frantz, C. K., Gundermann, D., ... Sweat, M. (2009). Using peer-led team learning to increase participation and success of under-represented groups in introductory computer science. *ACM SIGCSE Bulletin*, *41*(1), 163–167.

Hoyle, R. H., & Crawford, A. M. (1994). Use of individual-level data to investigate group phenomena issues and strategies. *Small Group Research*, *25*(4), 464–485.

Huang, Y., Strawderman, L., & Usher, J. (2013). A new model for mentoring graduate students: Teach them how to teach. Paper presented at Proceedings of the ASEE Annual Conference & Exposition.

Hughes, K. S. (2011). Peer-assisted learning strategies in human anatomy & physiology. *American Biology Teacher*, *73*(3), 144–147.

Kibble, J. D. (2009). A peer-led supplemental tutorial project for medical physiology: Implementation in a large class. *Advances in Physiology Education*, *33*(2), 111–114.

Lotkowski, V. A., Robbins, S. B., & Noeth, R. J. (2004). *The role of academic and non-academic factors in improving college retention* (ACT policy report). Iowa City, IA: American College Testing.

Loui, M. C., & Robbins, B. A. (2008, October). Work-in-progress-assessment of peer-led team learning in an engineering course for freshmen. *Paper presented at: Frontiers in Education Conference, Saratoga Springs, NY* (pp. F1F 7–8).

Luckie, D., Harrison, S. H., & Ebert-May, D. (2011). Model-based reasoning: Using visual tools to reveal student learning. *Advances in Physiology Education*, *35*(1), 59–67.

Lyle, K. S., & Robinson, W. R. (2003). A statistical evaluation: Peer-led team learning in an organic chemistry course. *Journal of Chemical Education*, *80*(2), 132–134.

Maloof, J., & White, V. K. (2005). Team study training in the college biology laboratory. *Journal of Biological Education*, *39*(3), 120–124.

Martin, D. C., & Arendale, D. R. (1992). *Supplemental instruction: Improving first-year student success in high-risk courses* (The freshman year experience: Monograph series number 7), Columbia: Center for the Study of the Freshman Year Experience, South Carolina University.

Marx, J. D., & Cummings, K. (2007). Normalized change. *American Journal of Physics*, *75*(1), 87–91.

McDermott, L. C., Shaffer, P. S., & Somers, M. D. (1994). Research as a guide for teaching introductory mechanics: An illustration in the context of the Atwood's machine. *American Journal of Physics*, *62*(1), 46–55.

Nesbit, J. C., & Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research*, *76*(3), 413–448.

Novak, J. D. (1990). Concept mapping: A useful tool for science education. *Journal of Research in Science Teaching*, *27*(10), 937–949.

O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, *97*(2), 316–333.

Okebukola, P. A. (1990). Attaining meaningful learning of concepts in genetics and ecology: An examination of the potency of the concept-mapping technique. *Journal of Research in Science Teaching*, *27*(5), 493–504.

President's Council of Advisors on Science and Technology. (2012). *Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics*. Washington, DC: U.S. Government Office of Science and Technology.

Preszler, R. W. (2006). Student- and teacher-centered learning in a supplemental learning biology course. *Bioscene: Journal of College Biology Teaching*, *32*(2), 21–25.

Preszler, R. W. (2009). Replacing lecture with peer-led workshops improves student learning. *CBE—Life Sciences Education*, *8*(3), 182–192.

Ramirez, G. M. (1997). Supplemental instruction. Paper presented at Proceedings of the 13th and 14th annual institutes for learning assistance professionals, University of Arizona (pp. 78–91).

Richardson, D., & Birge, B. (2000). Effects of an applied supplemental course on student performance in elementary physiology. *Advances in Physiology Education*, *24*(1), 56–61.

Rust, C. (2002). The impact of assessment on student learning: How can the research literature practically help to inform the development of departmental assessment strategies and learner-centred assessment practices? *Active Learning in Higher Education*, *3*(2), 145–158.

Rybczynski, S. M., & Schussler, E. E. (2011). Student use of out-of-class study groups in an introductory undergraduate biology course. *CBE—Life Sciences Education*, *10*(1), 74–82.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education*, *35*(4), 453–472.

Somers, R. L. (1988). *Causes of marginal performance by developmental students. Telementoring project study guide number six*. Boone, NC: National Center for Developmental Education.

Tekian, A., & Hruska, L. (2004). A review of medical school records to investigate the effectiveness of enrichment programs for "at risk" students. *Teaching and Learning in Medicine*, *16*(1), 28–33.

Tinto, V. (1987). *Leaving college: Rethinking the causes and cures of student attrition*. Chicago, IL: University of Chicago Press.

Tinto, V. (2001). *Rethinking the first year of college*. Higher education monograph series, Syracuse, NY: Syracuse University.

Tomanek, D., & Montplaisir, L. (2004). Students' studying and approaches to learning in introductory biology. *Cell Biology Education*, *3*(4), 253–262.