The effect of extended student hours on performance of students in an interdisciplinary, introductory undergraduate ecology course

Adrian Treves 1, Nicholas J. Balster 2
University of Wisconsin-Madison
Madison WI
1 Nelson Institute for Environmental Studies, 30A Science Hall, 550 North Park Street, University of Wisconsin, Madison WI 537016, USA, +1-608-890-450, atreves@wisc.edu
2 Department of Soil Science, 341 King Hall, 1475 Observatory Dr., University of Wisconsin, Madison WI 537016, USA, (608) 263-5719, njbalster@wisc.edu
[3] **Corresponding author**: atreves@wisc.eduwisc.edu ORCID 0000-0002-3052-4708
[4] **Acknowledgments.**. NSF provided support under S-STEM award #1643946
[5]**Human subjects protection:** This study was deemed exempt by University of Wisconsin-Madison Institutional Review Board SDS/ER 2018-0264 under federal regulation 45 CFR 46.102(l).

Abstract
Large-enrollment, undergraduate college courses often use plenary reviews before exams. Alternatives such as no review, trivia games, or practice exams have been evaluated. We present a before-and-after comparison of a novel intervention to improve exam performance in an interdisciplinary, introductory ecology course enrolling 150–220 non-majors. We evaluated summative exam performance of 397 participants and non-participants across 3 exams after some students reviewed in 'extended student hours' of sequential student-led meetings with the instructor for >20 minutes per group of <8 students, compared to those using practice exams only. Using a repeated measures, within-subject Hills-Armitage ANOVA and grouped comparisons to detect main, dose, order, and carry-over effects, we found that 4 of 7 treatment groups averaged 73–78% before intervention and improved 7–14% over practice exam participants, whereas the other 3 treatment groups that averaged 83–88% beforehand did not change after intervention, without significant order effects or carry-over effects. We found the positive, dose effect was 1<2=3. We present an approach to minimizing self-selection bias. It is unclear if the content or the format of extended student hours explained the effects. The effect size was similar to reports for trivia game reviews. Extended student hours seem to aid in formative assessment before exams.

Introduction

Exam review has been shown to improve exam performance compared to no review (King, 2010). A few methods of exam review have been subject to rigorous evaluation, including traditional (what we here call plenary review sessions in class), practice exam reviews, and trivia game reviews before exams, to name a few (Hackathorn et al., 2012). Yet, at the time of this writing, only 12 studies cite the latter study, so we echo their assertion that there remains a paucity of strong evidence supporting different methods of review to increase student exam performance.

Dissatisfaction with traditional (plenary) exam reviews arose from students as well as instructors, although perceived effectiveness of exam reviews does not always match measured effectiveness (Hackathorn et al., 2012). Among the criticisms, many plenary reviews become didactic exchanges in lecture format, a modality shown to promote passive, superficial learning

(Chickering and Gamson, 1987; Penner, 1984). Moreover, review sessions tend to backfill information missed during regular classroom instruction and can vary from teacher-centered summaries to active learning exercises (e.g., problem-based). The more student-centered the exam reviews have been the more they show increases in durable learning, reduce test anxiety, and increase academic success for a variety of students (Felder, 2002; Garhenhire, 1996), while providing formative assessment for an instructor's evaluation of content mastery (DiCarlo 2009; Qureshi et al., 2012). However, many studies are confounded by possible self-selection by already high-achieving students inclined to attend review sessions (Hackathorn et al., 2012; Jenson and Moore, 2009; King, 2010). Self-selection bias may be very hard to overcome in real-classroom settings where instructors wish to evaluate a method without coercive or exclusive randomized, controlled treatments. Here, we address both the paucity of evidence about review sessions with a quasi-experimental evaluation of a method we believe is novel and we offer a statistical method for measuring and partially neutralizing student self-selection bias.

Individual and small-group meetings with students (e.g., office hours) can provide a more personalized learning environment where the instructor can focus on specific misunderstandings and customized interventions. Indeed, these student-faculty interactions have long been considered a cornerstone to conventional pedagogy (McCabe and Pavela, 2004) and a critical element of effective teaching (Webb 2005), as they have been shown to improve academic achievement quantitatively in a variety of disciplines (Pascarella and Terenzini, 2005). For example, Guerrero and Rod (2013) examined the academic performance of 406 undergraduates over a four-year period in seven political science courses and found a strong correlation between attending office hours and course grades. Even when done remotely or on-line, the frequency of synchronous office hours with individuals and groups of students correlated with multiple measures of exam performance and academic achievement (Lavooy and Newlin, 2008; Li and Pitts, 2009). However, while most instructors integrate office hours into their course syllabi, both faculty and students share negative perceptions of their use (Guerrero and Rod, 2013) with consistent attendance being rare. Thus, recommendations including explicit mention of outside help in a course syllabus (Perrine et al. 1995) to pedagogical intervention to increase attendance have been proposed (Urban-Lurain and Weinshank, 2000). Replacing review sessions with extended office hours may increase exam performance, as we would expect student participation and engagement in the material to increase prior to an exam. This reasoning centers on an assumption of intrinsic motivation by students to improve their exam score (Tavakol et al. 2009), but more deeply rests on building substantive student-instructor interactions during the office hours. Moreover, such interactions may especially help students who are struggling in the course and who would be otherwise reluctant or intimidated to seek help (Karabenick and Knapp 1988). To our knowledge, inviting student-centered exam review through supplementation of some sort to traditional office hours (described in this study as extended student hours) specifically designed to improve exam performance has not yet been examined.

Small-group review sessions might optimally reduce some perceived disadvantages of one-on-one office hours (student intimidation, appearing to interrupt the professor, entering a new learning environment, etc.), encourage student participation by having peers around them, while also permitting more efficient formative assessment of several students simultaneously for the instructors of large enrolment classes. The motivation for evaluating the effect of extended student hours and a break from the more traditional approach was our consistent observations of

unsatisfactory, plenary review sessions (50-75 minute class periods) in classes ranging from 50 to 200 undergraduates. During these unsatisfactory sessions, we could not determine whether comprehension improved during or after the plenary review session and would typically receive fewer than five questions that rarely kindled student-instructor interactions to assess the student's mastery of the content. We often felt time was wasted repeating lecture material rather than helping students assess their learning. Although mid-course surveys revealed that students viewed these plenary reviews as moderately helpful, Author nonetheless discontinued the practice after 2015.

Concomitant with these observations of the reviews prior to exams, a similar lackluster trend with traditional office hours (one-on-one meetings) persisted, consistent with the peer-reviewed literature and with reports by the University of Wisconsin-Madison Teaching Academy's UCLASS group following focus groups with students held from 2015-2019 (https://teachingacademy.wisc.edu/uw-teach-2/u-class/ last accessed 27 March 2020). Fewer than 10 percent of students took the opportunity to attend office hours and those that did largely appeared shy or nervous when participating. An occasional student was forthcoming enough to help AUTHOR assess their understanding and thereby, assess classroom instruction. The typical student visit was 15 minutes or less. Similarly, mid-course surveys revealed that students evaluated office hours as moderately helpful with 20–25% reporting attendance at office hours at any time during the semester. Unlike the plenary reviews for the exams, we continued to hold office hours (renamed student hours), but in 2016, Author implemented a hybrid approach of extended student hours. Here we describe a case-control evaluation of these extended student hours. We focus on quasi-experimental (before-and-after comparison of impact, BACI without random assignment) evaluation of this pedagogical intervention as a way to improve student performance on summative assessments (3 midcourse exams) in a 15-week ecology lecture course enrolling 150–220 non-majors at university level. We hypothesized that extended student hours would improve the subsequent exam performance of participating students (measured within-subjects) more than not participating students (also measured within-subjects), against the alternative of no improvement.

Materials and methods

In 2016, Author designed and implemented a hybrid approach to reviewing for 3 midcourse exams over two consecutive autumn semesters (15 weeks long) of the same course (Botany/Zoology/Environmental Studies 260 Introductory Ecology) with 190 and 207 undergraduate, non-major students respectively. AITHOR intended to retain the advantages and diminish the disadvantages of the traditional plenary review, and maintain traditional office hours, and practice exams supplemented by the new extended student hours method.

Because we did not conduct a randomized trial, we discuss potential confounding variables that might bias the results due to self-selection bias and treatment bias. Therefore, we present multiple analyses of the effects of the intervention along with a discussion of one confounding variable that we cannot disentangle from the effect of the intervention. We offer this evaluation of a pedagogical intervention in hopes of stimulating discussion and motivating a future, gold-standard randomized, controlled experiment (Ioannidis, 2005).

Beginning during the Fall semester of 2016 and repeated in Fall 2017, Author convened extended student hours 2 or 5 days before each of six exams (three per semester). Exams were not cumulative, occurred in regular class sessions and consisted of 33 multiple-choice questions

scored automatically using Scantron® technology. Students ranged in rank from first-year to fifth-year undergraduates and came from diverse majors because the course fit a biological sciences distribution requirement. The teaching style was lecture with optional weekly discussion sections. AUTHOR was the sole instructor with two graduate student teaching assistants who did not attend extended student hours. Summative evaluations of the course by students were higher than average for AUTHOR's unit (4.1–4.3 out of 5 every year).

Intervention design: First, the name 'student hours' was intended to convey the time was for students and was not interrupting the instructor's other work, i.e., hypothetically more welcoming than 'office hours', as reported by the University of Wisconsin-Madison Teaching Academy's UCLASS group following focus groups with students held from 2015-2019 (https://teachingacademy.wisc.edu/uw-teach-2/u-class/ last accessed 27 March 2020). The rest of the name 'extended' referred to the time allotted, in which the AUTHOR allocated 3-4.5 hours to the effort.

Beyond naming, the design of 'extended student hours' allowed six to seven students into the instructor's office simultaneously in a space where they could all take notes and see each other and the instructor simultaneously while sitting around a large table with 6–7 comfortable chairs. For each group of students, the instructor allotted 20 minutes and students were invited to choose the 20-minute slot that fit their individual schedules within the total time allocated. The system was first-come, first-served using a Doodle® meeting planner so students were effectively reserving their seat. Throughout the session, the instructor invited students who were interested to remain beyond the 20 minutes, but they had to give up their seat if new students had arrived. Students who reserved but did not attend were not uncommon, which allowed the student in a prior session to double or even triple their attendance time if seats were available for newcomers. The instructor recorded attendance for the purpose of this analysis.

Two aspects of the design of extended student hours produce random error or conservative error more likely to make treatment and control similar. When a major, substantive source of confusion emerged in two or more sessions, AUTHOR inferred the teaching or course content had been unclear for many and thereafter volunteered the clarification for all subsequent sessions and posted a clarifying announcement on the course learning system for all students to benefit. Therefore, the design potentially could benefit non-participants, which is a conservative source of error reducing the likelihood of detecting a treatment effect.

Also, the design incorporated an inevitable treatment bias. Ideally, treatments are uniform and standardized across subjects (Ioannidis, 2005; Treves et al. 2019). That was impossible because extended student hours allowed students to choose their time of day and the questions they might ask. Moreover, the compositions of groups of students were haphazard and the instructor's responses were customized to the student and their questions or could even differ from group to group for the same question. We suspect the output was not systematic but random error, but we cannot rule out the possibility of treatment bias. Yet the effect of treatment bias as described above would be conservative, by blurring the difference in effect of treatment and control.

Case-control design: We did not randomly assign students to treatment or control (gold standard), but rather employed the silver standard of case-control (before-and-after comparison

of intervention also called quasi-experimental), in which subjects were compared to themselves before the intervention. Our primary response variable was to calculate exam t+1 – exam t performances among participants in the treatment between those two exams and for non-participants we calculated the same for any two exams between which they did not participate in extended student hours. A student might therefore contribute scores as both treated and control group participants at different times in the semester. All students contributed 3 scores to the analysis. Students could participate in 1, 2, or 3 interventions during each semester. Therefore, the same students might appear in 1-3 treatment conditions plus the complement of control conditions, at different times.

Silver standard tests such as this case-control provide approximately half of the strength of inference about the effect of interventions because of the effect of time as a confounding variable all else being equal (Treves et al. 2016, 2019). However, individual variations and self-selection bias might lower the strength of inference further. Individual variation is likely to play a large part in silver-standard experiments. For example, students self-selected to participate, so they may have been higher- or lower-performing students than non-participants at the outset, or those more motivated prior to the exam. A grouped comparison (average treatment effect versus average effect for non-participating control students) would not produce strong inference because self-selection would produce measurement bias (systematic error in favor of the treatment effect). Therefore, we relied on within-subjects before-and-after measures and secondly, we were able to estimate self-selection bias by comparing the scores on exam 1 of 290 students who never participated in the treatment to 32 students that participated at some time but did not do so before exam 1 (i.e., the latter were late adopters that only later became self-selected). This provides a minimum estimate of self-selection bias, because those who participated before exam 1 were both self-selected and early adopters. In sum, we have three categories of participants: non-participant control students who only had the benefit of practice exams, students who participated before the first exam (early adopter, self-selected), and students who only participated after exam 1 scores had been recorded (late adopter, self-selected).

Participants also had access to practice exams. We discuss the consequences of having two categories of participants for our results and for the design of pedagogical interventions in the future, because randomized trials are more difficult to implement than silver-standard before-and-after comparison within-subjects when instructors attempt interventions in real classrooms where randomized, controlled trials may be perceived as coercive or exclusive and therefore unfair.

An obvious limitation of our study is not having a true placebo, although the control condition was a practice exam posted for the entire class. Student awareness of receiving the treatment might affect performance independently of the content. Because it was obvious to a student if they participated or not, the self-selected students might also convince themselves of an effect even if the pedagogical content was neutral, e.g., some other aspect of the treatment, such as making time, participating, or discussing with one's instructor. Typically, one avoids misleading conclusions about treatment effects by using a realistic placebo (all but the therapy) or by 'blinding' subjects to treatment (e.g., extended student hours in which no content is discussed, i.e., just spending 20 minutes chatting with a group of students). We did not employ such controls, so we cannot rule out unintended effects on our treated students. Indeed, the content of

the intervention might be less important than the time spent interacting between student and instructor. However, non-content effects of extended student hours (i.e., greater confidence, positive interactions with peers or instructor) might be expected to carry-over to subsequent exams or have a carry-over effect beyond the imminent exam. We would not expect the treatment to have a persistent effect beyond the exam it immediately preceded because the nature of the intervention was to address questions of course content for the upcoming non-cumulative exam only, not address study habits or other longer-lasting ways of improving performance. Similarly, one expects student performance to improve over a semester as they learn the instructor's style, the content solidifies into knowledge, and perhaps as students settle into other classes simultaneously. Our design allowed us to detect such temporal dynamics, dose, or carry-over effects.

Because students could choose to participate in any or all of three treatments (Table 1), we documented a mix of participants. Participants were recorded as treated 1-3 times and those with fewer than 3 treatments might have participated early or late in the semester. Therefore, we could estimate any differences between those who participated a similar number of times (dose) but started at different exams (order effects), and we could compare a student who participated once and then stopped participating to detect carry-over effects from the early treatment. Therefore, we employed the Hills-Armitage procedure for analysis of unbalanced, cross-over design, which preserved the order of treatments by handling every permutation of treatment and control differently (e.g., AAB was different from ABA where A= control and B=treatment).
Table 1 here

Statistical analyses: Table 2 presents coding for the Hills-Armitage cross-over designs (subjects sometimes appeared as treatment and sometimes as control) following (Díaz-Uriarte, 2002). This approach employs t tests (assuming unequal variance) within-subjects and handles the period (which exam) as a factor, which allows detection of order effects. Too implement the Hills-Armitage approach, we created treatment groups (e.g., AAA, ABA, with the order of participation coded by position of the letters, Table 2).
Table 2 here

Results and Discussion

Anecdotal qualitative information: The 20-minute duration of extended student hours, although still too brief for some students, allowed about 6 conversations about various topics, in which students could probe and seek clarification. The instructor did not lecture but waited for questions and used them to engage individual students in a discussion or follow up, such as a short explanation with a question of the AUTHOR's own to assess learning or understanding. To reply fully to each question posed and integrate formative assessment of each student's comprehension, the instructor would often probe the comprehension behind a question before answering and would refer students back to the appropriate course content whenever possible. First-year students seemed especially well represented and likely to stay past their allotted times although the instructor did not record such data. Once a student stayed for the entire set of sessions (ultimately sitting on the floor for hours after their reserved slot had elapsed). Shy students appeared to benefit from the relative 'safety' of a group of peers and followed the lead of bolder students' questions. Some students never asked a question, yet the instructor saw they took notes and attended to their peer's questions.

Sample for quantitative analysis*:* Of 397 students who might have participated in the treatment, 73% never did so (control) and 27% did at least once (treatment), with 15% electing one dose, 7% electing two doses, and 5% electing three doses (Tables 1, 2). The years (2016 and 2017) did not differ in average or variability of summed exam scores (mean difference 0.4%, SD difference 0.3%), so we pooled the data for different years below. Exam performance across all students did not change appreciably over the course of the semester with grand averages of 79%, 80%, and 78% for exams 1, 2, and 3 respectively.

Treatment effects within-subjects: The Hills-Armitage test within-subjects for changes in exam scores by treatment group (Table 2), revealed a strong treatment effect (exact F=8.2, df=2, p=0.003). Inspection of the results revealed that 4 of 7 treatment conditions in Table 2 improved 7–14% over the prior exam (these 4 conditions averaged 73-78% on exam 1); but 3 of the 7 treatment conditions did not change appreciably after averaging 83–88% on exam 1.

We also estimated within-subjects carry-over effects of treatment beyond the upcoming exam. We tested if participants before exam 1 (BAA or BAB) improved on exam 2 more than controls (AAA) or participants before exam 2 (BBA or ABA) improved on exam 3 more than controls (AAA). There was no detectable within-subject carry-over effect (exam 2: F=0.42, p=0.52 and exam 3: F=0.36, p=0.56 respectively), so we infer whatever the treatment is doing, it has a short-term effect on the next exam only.

Self-selection bias and carry-over effects for early adopters: Were participants a priori different from controls in their tendency to change exam performance? Scores on exam 1 of 290 students who never participated (average score = 79%, SD 10.8%) and scores of 32 students that participated later but did not do so before exam 1 (82%, SD 9.6%) were close to statistically significantly different (comparison of group means assuming unequal variance F test=0.44, t ratio= 1.6, p=0.057 one-tailed because the hypothesis was identified a priori). Therefore, we infer a 3% difference as a minimum estimate of self-selection bias, independent of treatment effect, when comparing the *changes* in exam performance of participants to those of non-participants.

Examining group averages for actual exam scores, early adopters or participants before exam 1 averaged 5.6% higher than non-participants on exam 1. Late adopters averaged 4.9% higher on exam 2, and 5.5% higher on exam 3 than non-participant controls. These are not treatment effect (because they are grouped comparisons of single exam scores not within-subjects measures of change in exam scores), therefore we infer that after subtracting our minimum estimate of self-selection bias above (3%), there remained a slight difference in exam scores between participants and controls regardless of whether the participants were early or late adopters. Early adopters did not seem to differ from late adopters. The average of summed exam scores of early adopters was 0.7% higher than the average for participants before exam 2, and 0.2% higher than the average for participants before exam 3. These were not statistically significant ($F<2.2$, $p>0.09$ in both cases). Therefore, we infer that early and late adopters were similar in exam performance after treatment without carry-over effects on cumulative exam scores.

When we considered the whole semester and all exam scores (i.e., not a within-subjects measure of change), we found a small but significant dose effect on the sum of exam scores but not on the average exam score, which suggested participation twice or thrice was better than once (+3-4% increase in the sum of all three exam scores) but participating three times was no better than participating twice (+0.5%, p=0.81).

We contribute to rigorous evidence about exam reviews, which remains sparse 8–10 years after a review of the topic and call for more study (Hackathorn et al., 2012; King, 2010).

Limitations: Because our control condition was not a placebo we could not discern if the content of extended student hours had an effect or simply participating in extended student hours had the effect on the upcoming exam. We recommend a gold-standard experiment with random-assignment and a placebo control rather than our pseudo-control of a practice exam open to all students, for which use was not monitored. The improvements in scores of lower-performing students justify such an investment. Before-and-after comparisons of ongoing pedagogical interventions do not have the strength of inference of randomized, controlled experiments but nonetheless may diminish the effects of self-selection bias by comparing time series within-subjects as students join or do not join in the pedagogical intervention over the course of a semester. Another limitation is that only one instructor ,AUTHOR, participated in this initial study.

An unplanned benefit of our study was the comparison we could make between a novel intervention (extended student hours) against a control condition (practice exams) to which all students had access. Prior work has shown that practice exams are better than no review, but that practice exams are outperformed by trivia game exams, summarized in (Hackathorn et al., 2012). The latter authors cited one study that found trivia games improved over no exam review by 8–15%, which is very similar to our findings here. Therefore, we predict that direct comparisons of trivia games and extended student hours for exam review would yield similar effects on exam performance.

We began this intervention after dissatisfactions with both plenary exam review sessions and office hours as tools to help students improve on exams. Instructors who adopt extended student hours may also benefit through efficiencies of accessing student learning per unit time, additional opportunity for formative assessment, and encouraging less-confident students to communicate their learning during review sessions (Table 3). In Table 3, we array qualitative observations and impressions of the advantages and disadvantages of the three techniques. Attention to how instructors and students interact over content and how students demonstrate their understanding to instructors in formative assessments is a growing area of interest among pedagogical researchers (Hackathorn et al., 2012; Karpicke & Blunt, 2011). We call for additional quasi-experimental and randomized, controlled experimental evaluations of pedagogical techniques. We also recommend broader dissemination to teachers outside of the educational research community who may not read specialized journals.

Table 3 here

We present this example from a STEM course covering introductory ecology because it reached a large-enrollment, interdisciplinary class of non-majors in Botany and Zoology with majors in Environmental Studies, Environmental Sciences, and a handful of other non-STEM majors,

spanning first-year to fifth-year students. Some of these students will be future opinion leaders, government officials, or activists, so instructors who can improve summative exam performance by investing more time in formative assessment with less-motivated students may have a lasting impact on perceptions of science and ecosystem change.

Summary

The data from this study show that extended student hours significantly increased exam scores (7–14% increase) compared to practice exams and did so among students in the lower two-thirds of the grade distribution (average first exam score of <82%) but not the upper third. We found a slight dose effect (2 treatments as good as 3 but both better than one dose), no carry-over effects beyond the imminent exam, no order effects, and self-selection bias accounting for a 3% difference in exam performance between participants and non-participants before treatment. We conclude extended student hours, defined as 20-minute voluntary review sessions with <8 students before summative mid-course, multiple-choice exams, is among the most effective, known methods for improving performance in undergraduate, large-enrollment, science courses for students in the lower three quartiles of a class.

Literature Cited

DiCarlo, S.E. 2009. Too much content, not enough thinking, and too little FUN! Advanced Physiological Education 33, 257-264.

Díaz-Uriarte, R. 2002. Incorrect analysis of crossover trials in animal behaviour research. Animal Behaviour, 63, 815-822.

Ioannidis, J. P. 2005. Why most published research findings are false. PLOS Medicine 2(8), e124.

Favero, T.G. 2011. Active review sessions can advance student learning. Advanced Physiological Education 35: 247-248.

Felder, R.M. 2002. The effective, efficient professor. Chemical Engineering Education. 36(2): 114-115.

Gardenhire, J.F. 1996. Laney's success model for first year students. ERIC Document Reproduction Service No. ED417772.

Guerrero, M. and Rod, A.B. 2013. Engaging in office hours: A study of student-faculty interaction and academic performance. Journal of Political Science Education 9: 403-416.

Jenson, P.A., and R.J. Moore. 2009. What do help sessions accomplish in introductory science courses? Journal of College Science Teaching 38, 60-64.

Karabenick, S.A., and J.R. Knapp. 1988. Help seeking and the need for academic assistance. Journal of Educational Psychology 80(3): 406-408.

Karpicke, J. D., & Blunt, J. R. 2011. Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping. Science 331(6018), 1-7.

Li, L. and Pitts, J.P. 2009. Does it really matter? Using virtual office hours to enhance student-faculty interaction. Journal of Information Systems Education 20(2): 175-185.

McCabe, D.L. and G. Pavela. 2004. Ten (updated) principles of academic integrity: how faculty can foster student honesty Change. The Magazine of Higher Learning 36(3): 10-15.

Pascarella, E.T. and P.T. Terenzini. 2005. How College Affects Students: A third decade of research. San Francisco, CA: Jossey-Bass.

Platt, J. R. 1964. Strong inference. Science 146, 347–353.

Qureshi, A., Cozine, C., and F. Rizvi. 2012 Combination of didactic lectures and review sessions in endocrinology leads to improvement in student performance as measured by assessments. Advanced Physiological Education 37(1):89-92.
Treves, A., M. Krofel, J. and McManus, 2016. Predator control should not be a shot in the dark. *F*rontiers in Ecology and the Environment. **14**, 380-388.
Treves, A., M. Krofel, O. Ohrens, and L. M. Van Eeden, 2019. Predator control needs a standard of unbiased randomized experiments with cross-over design. Frontiers in Ecology and Evolution **7** 402-413.
Webb, D. A. 2005. Twelve easy steps to becoming an effective teaching assistant. PS: Political Science & Politics 38(4): 757-761.

Table 1. Number of subjects and the 'dose' of extended student hours they elected prior to each exam in an undergraduate, non-major, lecture-based, ecology course. The codes (AAA, BBB) are used in analysis and Table 2.

	Student never participated (control) AAA	Subject participated at least once (treatment)	Subject participated in only 1 exam (dose=1)	Subject participated in 2 exams (dose=2)	Subject participated in all 3 exams (dose=3) BB
N	290	107	59	28	20
First participation before exam 1, BAA		75			
First participation before exam 2, ABA		26			
First participation before exam 3, AAB		6			

Table 2. Treatment conditions and sample sizes of students exposed to the treatment of extended student hours in a before-and-after comparison of impact, where A=non-participant, B=participant, and the position in a trio of such letters indicates when the student participated. Because students could choose to participate in any or all of three treatments (Table 1), we documented a mix of participants. Participants were recorded as treated 1-3 times and those with fewer than 3 treatments might have participated early or late in the semester. Therefore, we could estimate any differences between those who participated a similar number of times (dose) but started at different exams (order effects), and we could compare a student who participated once and then stopped participating to detect carry-over effects from the early treatment. Therefore, we employed the Hills-Armitage procedure for analysis of unbalanced, cross-over design, which preserved the order of treatments by handling every permutation of treatment and control differently (e.g., AAB was different from ABA where A= control and B=treatment).

Treatment before	Codes, N	One exam only, N	Two exams, N	Three exams, N
No exam	AAA, 290	-	-	-
First exam	-	BAA, 37	ABB, 10	-
Second exam	-	ABA, 16	BBA, 12	-
Third exam	-	AAB, 6	BAB, 6	BBB, 20

Table 3. Exam review methods and proposed relative advantages and disadvantages ranked.

Method	Plenary session	Office hours	Extended student hours (this study)
Relative rank	1 = relatively most effective, 3 = relatively less effective, blank = no known relative difference		
Efficiency for reaching most students	1	3	2
Formative assessment possible	3	1	2
Ease of changing teaching style to match learning style of student	3	1	2
Student intimidation to attend	1	3	2
Student intimidating to speak up	3	2	1
Students learn from each other	2	3	1
Instructor learns about individual students	3	1	2
Median rank (mean)	2 (2.3)	2 (2)	2 (1.7)