Impact of more realistic and earlier practice exams on student metacognition, study behaviors, and exam performance

Muxin Zhang,[1] Jason Morphew,[2] and Tim Stelzer[1]

[1]*Department of Physics, University of Illinois Urbana-Champaign, Urbana, Illinois 61801, USA*
[2]*School of Engineering Education, Purdue University, West Lafayette, Indiana 47907, USA*

(Received 27 June 2022; accepted 22 March 2023; published 27 April 2023)

Preparing for high-stakes exams in introductory physics courses is generally a self-regulated activity. Compared to other exam reviewing strategies, doing practice exams has been shown to help students recognize gaps in their knowledge, encourage active practicing, and produce long-term retention. However, many students, particularly students who are struggling with the course material, are not guided by research-based study strategies and do not use practice exams effectively. Using data collected from a fully online course in Spring 2021, this study examines two interventions aimed at improving student self-regulated studying behaviors and enhancing student metacognition during exam preparation. We found that a modified format of online practice exams with one attempt per question and delayed feedback, increases the accuracy of feedback about student readiness for exams but does not change the accuracy of their predicted exam scores or studying behaviors. Additionally, an added mock exam one week before the actual exam impacts students' intentions for studying but does not impact actual study behaviors or facilitate metacognition. These results suggest that interventions designed to improve exam preparation likely need to include explicit instruction on study strategies and student beliefs about learning.

DOI: 10.1103/PhysRevPhysEducRes.19.010130

I. INTRODUCTION

Exams are an important and widely used method of assessment in introductory college physics courses because of their reliability, validity, and efficiency for large-scale courses [1]. At the University of Illinois, students pursuing degrees in the College of Engineering are required to complete several introductory physics courses, in which hour exams and final exam scores make up about 50% of their total course grades. This means that despite getting good "effort grades" in other components of the course such as online assignments, lab projects, and participation in lectures and discussions, students can receive low or failing course grades solely due to low performance on the exams. This discrepancy between effort grades and exam scores is important to investigate [2], as grades in introductory courses are among the strongest predictors of student persistence for science, technology, engineering, and mathematics majors [3–7].

For these reasons, providing tools that help students prepare for exams more effectively is an important task for instructors and course designers. Particularly, research shows that practice exams are a highly effective learning tool and can significantly improve student performance on their actual exams [8]. At the University of Illinois, we have been implementing online practice exams with solution videos for our large introductory-level physics courses for the past decade. The online system is set up such that, before every exam, students gain access to exam problems that have been given on exams in previous years. They can submit answers, receive immediate feedback, and watch solution videos to the problems. A previous clinical study demonstrated that doing practice problems paired with solution videos improved student performance [9]. In addition, our end-of-semester surveys consistently show that students value online practice exams more than almost any other component of the course. In Fall 2018, 72.4% of the students reported that practice exams were "essential" or "very important" in helping them understand the course material.

Despite the demonstrated benefits of practice exams and the high value students place on them, both anecdotal evidence and analyses of students' usage of online practice exams suggest that many students are not engaging with practice exams in ways that effectively prepare them for actual exams. This may be particularly true for those students who are struggling with the course material. In a preliminary study [10], we found that students do most of their exam practice less than 48 hours prior to the actual exam. This cramming behavior likely limits the effectiveness of their practice, not allowing enough time for improvement despite

working through multiple practice exams. We also found that, because students were able to update their answers on the online practice exam system, they would eventually achieve near-perfect practice exam scores online without fully understanding the material. This inaccurate feedback from the online system may potentially misguide students during their exam preparation, giving them an "illusion of understanding" [11–13].

In this paper, we present findings from two *in situ* interventions implemented in a large-scale introductory calculus-based mechanics course with online instruction to address the observed issues and improve the effectiveness of online practice exams. The first intervention examined the effect of providing students with a new practice exam format to give them more accurate feedback and more realistic practice than the original format. The second intervention provided incentives for students to avoid cramming by providing students the opportunity to take a "mock exam" a week before the actual exam. We then collected practice exam usage data, survey data, and scores on the course exams to answer the following research questions:

1. How do different formats of online practice exams affect students' judgment of their proficiency, exam preparation behaviors, and exam performances?
2. How does introducing a mock exam one week before the actual exam impact students' exam preparation behaviors?

II. THEORETICAL FRAMEWORK

Among the various learning activities that students do for a college physics course, exam preparation is an activity that relies heavily on effective self-regulated learning [14,15], where students have control over their own studying process and how they utilize the resources provided. While instructors can offer tools and materials to guide students learning, such as review sessions and practice exams, most exam studying occurs outside of the classroom. As a result, students can approach exam preparation with different preconceptions about effective exam preparation practices, leading them to adopt various studying strategies, some of which are more likely to result in greater exam performance than others.

A useful model of self-regulated learning for investigating the interaction between metacognition, academic success, and other individual factors is the four-phase model originally proposed by Winne and Hadwin [15,16]. In the first phase, the learner analyzes and forms a perception of a task (e.g., studying for an exam). In the second phase, the learner uses their perception of the task and their epistemology (e.g., beliefs about knowledge and learning) to generate goals for their studying. In the third phase, the learner enacts a study plan to achieve the goals set during planning. During this enactment, the learner monitors their learning against their goals. Based on their monitoring, their epistemology, and their knowledge of study strategies, the learner can continue to study using the same strategies, change their study strategies, discontinue studying, or go back and modify their goals or task definition. After studying, during the fourth phase, the learner, if motivated, might adapt their strategies or beliefs about learning in response to feedback.

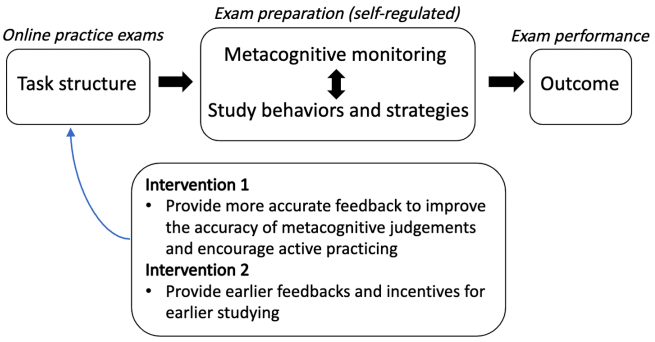


FIG. 1. Summary of the theoretical framework and rationales for interventions.

In this section, we review the literature surrounding two key aspects of self-regulated learning related to our interventions: (i) Metacognition, which is students' judgment of their knowledge and awareness of their learning process and (ii) study strategies, which is students' approach to exam preparation and how they interact with the learning tools provided. These two aspects of self-regulated learning are entangled with each other and together impact the effectiveness and outcome of exam preparation. Additionally, we review existing theories of how the structure of practice exams can support these aspects of self-regulated learning and describe how these theories informed the design of our interventions (see Fig. 1).

A. Metacognition

Self-regulated learning is guided by metacognitive monitoring and control processes [15,17]. There is a dynamic, reciprocal, and iterative relationship between metacognitive monitoring of performance, beliefs about learning and studying, and studying decisions enacted by students [14,18]. When engaged in self-regulated learning tasks, such as preparing for exams, students need to monitor their current level of knowledge or understanding, then compare their perceived understanding to their goal for the exam. By monitoring the discrepancy between a self-assessed current state and an internal model representing the desired state, a student makes decisions about whether to continue studying, change study strategies, or stop studying. These decisions are also impacted by students' beliefs about effective learning and studying practices [19–24].

Because metacognitive control decisions depend on students' monitoring of their learning, the effectiveness of the self-regulated learning process depends on the accuracy of that monitoring. The accuracy of an

individual's metacognitive judgments is often related to one's domain knowledge or proficiency [25,26]. While student metacognitive judgments tend to correlate with their performance in many settings [18,27], studies investigating the accuracy of metacognitive judgments usually find that students overestimate their own performance on exams, with low-performing students being overconfident by as much as 2–3 letter grades [11–13]. This asymmetry in the accuracy of learners' metacognitive judgments is thought to occur because the expertise and skills needed to make accurate metacognitive judgments of performance are the same types of expertise and skills needed to produce good performance on a task [28]. From this perspective, low-performing students suffer from a "double curse" of being both unskilled and unaware of their lack of skill [11]. Because success at self-regulated learning is positively associated with academic performance [29], all students—and low-preforming students in particular—could benefit from interventions that target improving the accuracy of metacognitive monitoring.

B. Study strategies

In addition to accurate metacognitive monitoring, students need to know about effective study strategies, know how to enact these strategies, and understand the types of tasks for which each strategy is most effective in order to improve their study behaviors. The accuracy with which students monitor and evaluate their learning is positively related to planning and enacting study strategies [30]. However, students who accurately monitor their learning can struggle with knowing how to adapt their study strategies. For example, lower-performing students may know that they are unprepared for an upcoming exam but be unaware of how to modify their studying to engage with the material more effectively.

Students tend to prefer using passive methods when studying for exams, such as rereading and reviewing notes [31,32]. When students do engage in problem solving, they often utilize methods that focus on memorizing formulas or attempting to match the surface features to other problems that they have solved [33–35]. Lower-performing students also tend to take means-ends approaches to solve problems, such as working backward from a goal state by reducing the difference between the initial state and the goal state [36,37]. This approach often leads students to use unproductive strategies, such as equation hunting, where students simply search for equations that contain the to-be-solved-for variables [34,38].

Besides using passive studying methods, students tend to focus the majority of the studying one to two days before an exam [31,32,39], a trend that was also apparent in our pilot data [10]. While cramming can facilitate short-term performance and lead to high student confidence [40,41], it has a detrimental effect on long-term retention [42,43]. Additionally, engaging in rehearsal strategies, such as reviewing notes or rewatching lectures, can create false perceptions of mastery [44–46]. For long-term retention, testing (e.g., studying using practice tests) is a particularly effective method, especially when practice testing is spaced over time (e.g., distributed practice and the spacing effect) [47,48].

C. Practice exam as a learning tool that facilitates self-regulated learning

Although students are the main agents that monitor and control their studying, instructors can provide learning tools that are specifically designed to support self-regulated learning. In the case of exam preparation, practice exams are helpful tools not only because they provide an effective form of testing that promotes active studying strategies [49] but also because they act as a type of formative assessment [50] that offers students valuable feedback about their current proficiency, which can guide students' metacognitive monitoring and future studying.

Compared to other exam studying strategies, practice exams are a form of testing that encourages active retrieval and help students recognize gaps in their knowledge (e.g., the testing effect). The testing effect has been shown in both clinical studies [51,52] and secondary and university classrooms [53–56] where students engaging in testing achieve better long-term retention than students engaging in passive studying. In addition, engaging in active problem solving or practice testing is shown to benefit learning for questions similar to those that were tested [57], for analogical problem solving [58], and for inferential and application questions [44,59]. The learning benefits of testing appear to enhance learning for items correctly answered, as well as items that were not correctly solved as long as students are provided with personalized feedback and the ability to restudy the tested material [60–63]. Some studies have even found that engaging in testing can enhance performance for new material that had not been tested (i.e., test-potentiated learning) [64–67]. This test-potentiated learning effect suggests that providing students with feedback and incentives to revisit material following testing can enhance future learning.

Whether these potential learning benefits of practice exams are realized depends on how the tests are implemented, including their similarity to the actual exams and the format of feedback given to students during testing [8,9]. In some cases, using practice exams can lead to inaccurate judgment of proficiency because the feedback students receive from the practice contributes to an illusion of understanding [68]. The effectiveness of practice exams also depends on students' prior knowledge and how they engaged with the practice exams. For example, Balch found that, for two groups of students who were given access to the same practice test, the group that did the test before viewing solutions performed better than the group that only viewed solutions [69].

Therefore, practice exams can be an important tool to facilitate self-regulated learning during exam preparation, but the structure of practice exams and how students interact with the tasks are essential to the tool being beneficial. These factors motivated the design of our interventions in this study.

D. Rationales of interventions

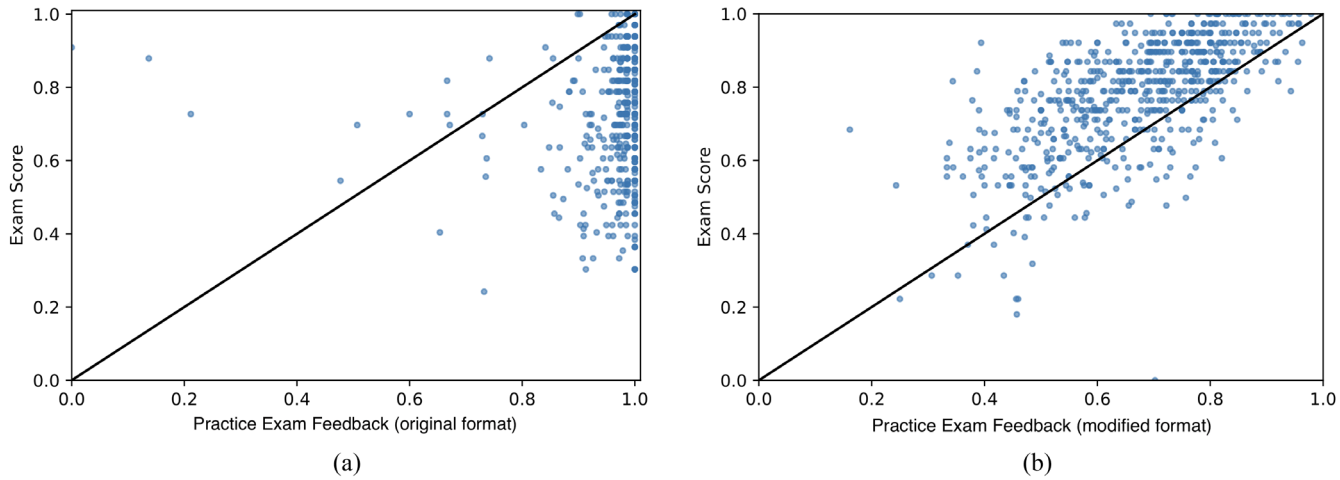
Building on theories of self-regulated learning, we implemented two interventions to support students' exam preparation that aim to help students fully utilize the benefit of practice exams (see Fig. 1). The first intervention modified the format of how practice exams are presented to students online such that it provides students with more accurate feedback on their proficiency (see Fig. 2) and provides a task structure that incentivizes active practicing rather than passive studying. We explain details about the original and modified formats in Sec. IV. We expect this change to improve students' metacognitive judgment of their proficiency during exam preparation, which can guide them to engage in more effective exam preparation behaviors, for example, doing more practice exam problems if their current feedback is not aligned with their goals on the exam.

In the second intervention, we address the issue of cramming by providing incentives for students to do a "mock" exam and receive feedback a week ahead of the actual exam. By providing this early testing opportunity, we hope to improve students' metacognitive judgments early, giving them an extended period of time to practice and reach the desired level of proficiency. We expect students who did this early practice to engage in less cramming behaviors and more evenly distribute practice problems across the week ahead of the exam, hopefully resulting in improved performance on the actual exam.

III. STUDY CONTEXT

This study is situated in a large calculus-based introductory mechanics course (Physics 211) required for students pursuing a bachelor's degree in the College of Engineering at the University of Illinois Urbana-Champaign. The course is designed to help students understand fundamental concepts and continue to refine their problem-solving skills through a variety of learning activities each week, including multimedia prelectures, preflights [70], lectures with Peer Instruction [71], online homework, labs, and group problem solving in the discussion section. The exams are set up so that students take three "hour exams" throughout the semester and a final exam at the end. These exams are meant to assess their understanding of the material. At the end of the course, students were graded on their performance on the exams in addition to their participation in other components of the course, with the exams accounting for 50% of their course grade.

Students are given access to four online practice exams at least one week before each hour exam. The format of these exams has evolved over time, but the current implementation includes both a pdf file of the exam questions and a version of the exam coded into their online homework system. The online version includes the option to grade the questions and provides students access to a video solution of the problem. The practice exams are optional, but their use for exam preparation was encouraged by the professors and students self-reported that the practice exams were very helpful in learning the material.



(a) (b)

FIG. 2. Relationship between students' actual exam score percentages (y axis) and the practice exam feedback they receive (x axis). (a) In the original practice exam format, students can correct their answers after seeing solutions, resulting in feedback displaying near-perfect practice exam scores, which is not reflective of actual proficiency. (b) In the modified practice exam format, students can only input answers once, so the displayed scores better reflect their proficiency (i.e., correlate stronger with exam scores). Figure 2 is reprinted from a Physics Education Research Conference Proceeding paper [10], which contains additional details.

We implemented two separate interventions in Physics 211 during the spring semester of 2021, with 1108 students registered in the course. Because of Covid-19, the course was fully online during this semester. Lectures, discussions, and labs were given synchronously over Zoom, with extraordinary efforts made to engage students. However, it has been robustly found that students and instructors believe that remote instruction is less effective than in-person instruction [72–74], findings that have also been found at Illinois. The exams were done synchronously, but online with teaching assistants (TAs) proctoring students using Zoom.

In the next sections of this paper, we describe the methods and findings from each intervention separately and then summarize the results from both interventions at the end.

IV. METHODS (INTERVENTION 1)

A. Description of intervention 1: Modifying the format of practice exams

In order to compare the effects of two different practice exam formats, we evenly divided the class into two random groups of students and provided them with different formats of practice exams for hour exam 2 and 3 (see Fig. 3). The original format, which we will call "multiple-attempt," was set up such that when students approached each question, they could see a "submit" and a "help" button under the statement of each question. They could submit an answer and receive immediate feedback on whether their answer was right or wrong. They could change their answer as many times as they want and receive feedback each time. Meanwhile, they could click the help button any time during the practice and view a solution video. Students could choose to watch the solution video before they submitted any answer and could change their answer after they watched the video. When using the original format, students often change their answers multiple times until they get nearly perfect practice scores [see Fig. 2(a)].

For the modified format, which we refer to as "single-attempt," we adjusted the format that these practice exams were delivered such that, when students were working on the problems, they had an experience closer to what they would have in a real exam. Each practice exam was divided into "clusters." Students could view one cluster at a time, which contained about 1–4 questions related to a single situation. Students could submit answers to these questions, but they would not get immediate feedback on whether their answers were right or wrong. Only after they had submitted answers to all the questions in that cluster, could they click "Submit Cluster" and get feedback on correctness and access the solution videos. Unlike in the multiple-attempt format, students could not see any feedback or solutions before they had submitted the entire cluster, so they are more likely to actively attempt the whole cluster of questions before reading solutions, similar to what they do on a real exam. Students also could no longer change their answers after they had submitted the entire cluster and had seen solutions, so their final scores on the practice exam in this format would better reflect the scores they would receive on the actual exam [see Fig. 2(b)].

B. Data collection and analysis

We collected data related to students' exam preparation behaviors through the online practice exam system, including how many practice exam questions they attempted, when the questions were attempted, and their use of solution videos. Additionally, we implemented surveys to measure student metacognition, asking students to predict the scores they would receive on the actual exam. The surveys were due on the same day as the hour exams, after students had already done some practice exam questions, thus measuring their metacognitive judgment at approximately the end of their exam preparation.

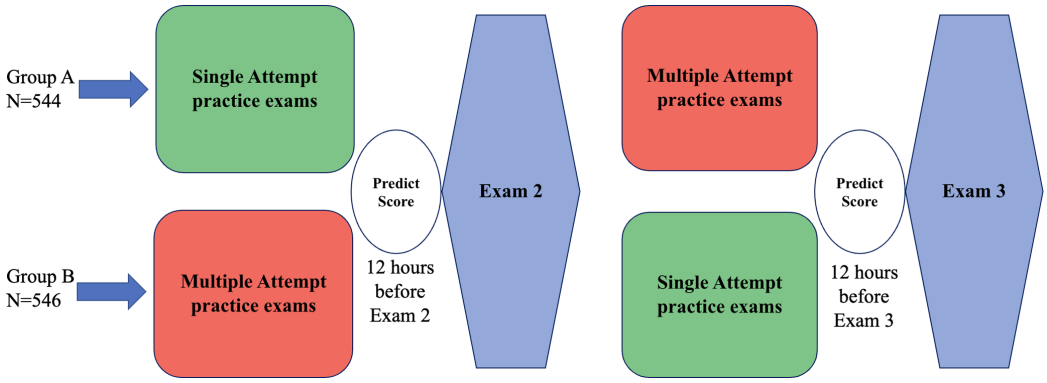


FIG. 3. Timeline and implementation of intervention 1. Practice exams became available to both groups of students one week ahead of the actual exams. For hour exam 2, student group A received the single-attempt format, while group B received the multiple-attempt format. We switched their formats in hour exam 3 so that each group experienced both formats. "Predict score" means that students completed the prediction survey where they made predictions of the scores that they would earn on the exam.

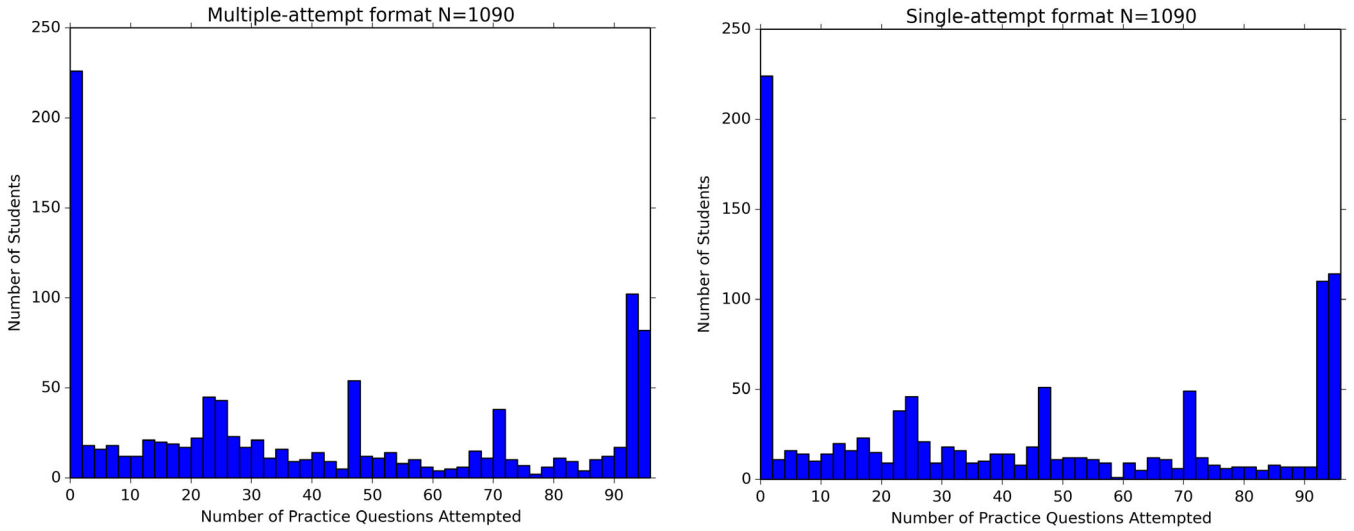


FIG. 4. A comparison of the distributions of the quantity of practice for the two format groups, combining data collected across both hour exam 2 and hour exam 3. In (a), a total of 1090 students received the multiple-attempt format (old version), and on average each student did 41 practice exam questions. In (b), a total of 1090 students received the single-attempt format (modified version), and on average each student did 43 questions.

For intervention 1, a total of 1090 students completed both hour exam 2 and hour exam 3. In our data analysis, we focused on comparing the study behaviors, metacognition, and exam performance of two groups: students who received single-attempt practice exam format and students who received multiple-attempt format. Except for the comparison of the quantity of practice in Fig. 4, we analyzed the data from hour exam 2 and hour exam 3 separately. There was a difference in the exam averages and individual students differed in when they began studying for each exam, making a combined statistical analysis difficult. However, we interpret the analyses of exams 2 and 3 as replication studies and interpret any findings that are not consistent across both exams as potential type I errors.

V. FINDINGS (INTERVENTION 1)

A. The effect of practice exam format on exam preparation behaviors

Comparing the two format groups, we found no significant difference in the way students interacted with the online practice exam problems. Figure 4 shows that, students who received the multiple-attempt format, either in hour exam 2 or in hour exam 3, attempted similar number of practice questions as students who received the single-attempt format, with a similar distribution. To compare the number of practice exam questions completed by both groups, we examined homogeneity of variance (HOV) and normality of the distributions. Levine's tests of HOV indicated that homogeneity of variance could be assumed, however Shapiro-Wilk tests of normality indicated non-normal distributions (see Fig. 4). Because t tests are robust to deviations from normality, independent t tests were conducted to compare the number of practice exam questions completed before each hour exam between the groups. The results show that there was no evidence for a difference in the number of practice exam questions students completed in either the single-attempt or multiple-attempt formats [hour exam 2: $t(1088) = 1.89$, $p = 0.06$, $d = 0.11$, hour exam 3: $t(1088) = 0.42$, $p = 0.67$, $d = 0.02$].

A factor that may explain this null finding is that students' cramming behaviors limit the number of practice questions they do. To analyze whether there were any differences between the conditions based on when an individual began using the practice exams, we conducted two analysis of covariance (ANCOVAs), with the number of practice exam questions as the response variable, practice test format as the independent variable, and time of first practice test as the covariate, were conducted. Analysis of the q-q plots indicated that the residuals were relatively normally distributed, Levene's tests indicated that homogeneity of variance could be assumed, and examination of the interactions between the independent variable and the covariate indicated that homogeneity of regression could be assumed for both exams. For exam 2, students who began studying earlier attempted more practice exam problems, $F(1, 1087) = 886.59$, $p < 0.001$, $\eta^2 = 0.45$, and students who received single-attempt feedback attempted more practice exam problems on average controlling for the time of the study, $F(1, 1087) = 3.89$, $p = 0.04$. However, this effect was small both practically (four additional problems on average) and statistically ($\eta^2 = 0.002$). For exam 3, students who began studying earlier attempted more practice exam problems, $F(1, 1087) = 724.88$, $p < 0.001$, $\eta^2 = 0.40$,

however, for this exam, students who received single-attempt feedback did not attempt more practice exam problems, $F(1, 1087) = 3.15$, $p = 0.08$, $\eta^2 = 0.002$.

Therefore, despite being designed to encourage effective exam preparation behaviors, the single-attempt format practice exams did not have a significant effect on the quantity of practice questions that students did. With the single-attempt format, students have to work on a cluster of practice questions without any help initially, so one might expect some resistance from students to attempt practice questions. We did find that students who begin studying earlier attempt more problems on average, but prior research has shown that students often do not engage with the practice exams until the day or two before the course exam [10]. We designed Intervention 2 to address this cramming behavior.

B. The effect of practice exam format on the metacognitive bias

Since the single-attempt practice exam format provides more accurate feedback for students than the multiple-attempt format [10] (see also Fig. 2), we expect students in the single-attempt format group to demonstrate more accurate metacognitive monitoring in the prediction survey than students in the multiple-attempt format group. To examine this, we calculated metacognitive bias using the signed difference between predicted exam scores and actual exam scores. Although there are many other ways to measure the accuracy of student metacognitive monitoring [75], this metacognitive bias value captures how close a prediction is to the actual performance, as well as whether a student is overconfident (positive metacognitive bias) or underconfident (negative metacognitive bias).

We removed 8 students who made predictions greater than 100 for their exam scores, leaving 1082 students who completed exams 2 and 3. For exam 2, 1000 students provided a valid prediction (i.e., between 0 and 100), and for exam 3, 988 students provided a valid prediction. For hour exam 2, students generally overestimated their exam scores by 15 points, which is much higher bias overall than hour exam 3 (see Table I). This is likely due to the different exam difficulty levels between hour exam 2 and 3, with hour exam 3 having a higher class average. Indeed, for both cases, when we ask students to predict their exam scores, we observe similar prediction distributions with means around 80%.

Because the purpose of the practice exams is to give students formative feedback regarding their preparedness, we would expect that students who attempt at least 75% of a practice exam, or about 20 practice questions, may demonstrate more accurate metacognitive monitoring. Therefore, to examine whether attempting practice exam problems before making a prediction impacted student metacognition or exam performance, we divided students into three categories based on their use of the practice exams. Students who did practice exams early (i.e., attempted 20 or more practice questions before they took the prediction survey), those who did practice exams late (i.e., if they attempted fewer than 20 practice questions before they took the prediction survey), and those who did not do any practice exam questions at all. This categorization is crucial because students who did not do any practice exams, or students who attempted less than 20 questions before they did the prediction survey, would not have received valid proficiency feedback before making their prediction. In other words, for the late and no practice group, there is no reason for us to expect that the format of the practice exam would impact their metacognitive bias. We replicated the analyses using 1, 10, and 30 practice test questions attempted and found no differences (see the Appendix).

To examine the effect of practice exam format and the timing of practice exam attempts, we conducted two 2×3 (format $\times$ time) two-way analysis of variance (ANOVAs) with metacognitive bias as the response variable, practice test format and time of first practice as the independent variables. Analysis of the q-q plots and distributions of the residuals indicated that the residuals were relatively normally distributed, Levene's tests indicated that homogeneity of variance could be assumed for both exams. See Table II for the detailed results.

TABLE I. Means and standard errors by condition and time to attempt the 20th question. Note that meta bias is the signed difference between the exam prediction and the exam score for a given student. Single and multiple refer to the single-attempt and multiple-attempt practice exam condition.

	Early		Late		No practice test	
	Single M (SE)	Multiple M (SE)	Single M (SE)	Multiple M (SE)	Single M (SE)	Multiple M (SE)
			Exam 2			
Meta bias	12.7 (1.5)	16.9 (1.7)	15.8 (1.1)	15.5 (1.2)	11.6 (2.0)	16.1 (1.9)
Exam score	71.3 (1.4)	67.9 (1.4)	69.9 (1.1)	69.6 (1.1)	70.0 (2.0)	69.2 (1.9)
			Exam 3			
Meta bias	−0.4 (1.4)	−1.6 (1.4)	3.4 (1.3)	2.2 (1.1)	5.1 (2.4)	10.0 (2.7)
Exam score	81.5 (1.2)	85.5 (1.1)	77.6 (1.0)	78.9 (0.9)	75.2 (2.0)	72.6 (2.3)

TABLE II. Analysis of variance of metacognitive bias for exams 2 and 3. Note that for exam 2, practice exam condition was only significant for the analysis using 20 questions. This main effect was not significant for all other analyses (see the Appendix).

Exam 2					
Source	d.o.f.	Type III SS	MS	F	p
Practice exam condition	1	1512.90	1512.90	4.28	0.04
Timing	2	486.11	243.06	0.69	0.50
Practice exam condition × timing	2	1285.22	642.61	1.82	0.16
Error	994	351 066.43	353.19		
Exam 3					
Source	d.o.f.	Type III SS	MS	F	p
Practice exam condition	1	129.55	129.55	0.34	0.56
Timing	2	6869.00	3434.50	8.93	<0.001
Practice exam condition × timing	2	1312.05	656.02	1.71	0.18
Error	982	37 7791.37	384.72		

For exam 2, we did not find a significant interaction, thus we examined the main effects. The timing of the practice test did not have a significant effect on metacognitive bias. However, a significant difference was found between practice exam conditions, with students who received the single-attempt practice exam format having a lower metacognitive bias. However, given the low effect size and the presence of a difference between conditions for students who did not attempt any practice exam questions (see Fig. 5), we believe this result is not robust and is likely a type I error.

For exam 3, we did not find any significant interaction, thus we examined the main effects. The timing of the practice exams had a significant effect on metacognitive bias. However, a difference between the exam conditions was not found. The difference in metacognitive bias by the timing of the practice exam was as predicted with those who began studying earlier demonstrating lower metacognitive bias (Fig. 5).

In summary, unlike our expectations, the more accurate feedback provided in the single-attempt practice exam format did not affect the metacognitive bias that students had when they were asked to predict their exam performance. It is possible that the prediction data we collected did not reflect students' actual judgment of their proficiency because of their optimism on exam day. However, this result is consistent with our finding comparing their exam preparation behaviors where the two format groups did a similar number of practice questions and tended to use the practice tests in the days immediately preceding an exam.

C. The effect of practice exam format on exam performance

To examine the effect of practice exam format and practice test timing, two 2×3 (format × time) two-way ANOVAs were conducted with exam score as the response variable, practice test format and time of first practice test as

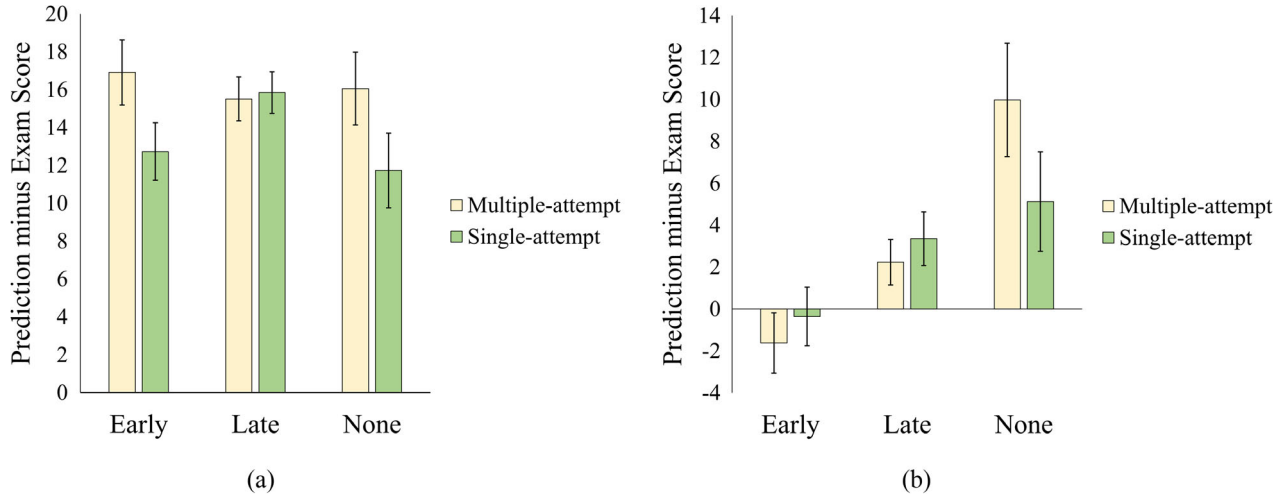


FIG. 5. Mean metacognitive bias for (a) hour exam 2 and (b) hour exam 3. Note: Error bars represent 1 standard error of the mean.

TABLE III. Analysis of covariance of exam score for exams 2 and 3.

Exam 2					
Source	d.o.f.	Type III SS	MS	F	p
Practice exam condition	1	695.15	695.15	2.23	0.14
Timing	2	278.54	139.27	0.45	0.64
Condition × timing	2	763.72	381.86	1.22	0.29
Error	1084	338 239.63	312.03		
Exam 3					
Source	d.o.f.	Type III SS	MS	F	p
Practice exam condition	1	88.96	88.96	0.30	0.58
Timing	2	10 174.43	5087.21	17.34	<0.001
Condition × timing	2	1414.69	707.35	2.41	0.09
Error	1084	318 027.83	293.38		

the independent variables. Analysis of the q-q plots and distributions of the residuals indicated that the residuals were relatively normally distributed, and Levene's tests indicated that homogeneity of variance could be assumed for both exams. See Table III for the detailed results.

For exam 2, we did not find a significant interaction, thus we examined the main effects. The timing of the practice test did not have a significant effect on exam score nor did the practice exam condition. For exam 3, we did not find a significant interaction, thus we examined the main effects. A difference between the practice exam conditions was not found, however, the timing of the practice exams had a significant effect on exam scores. *Post hoc* Tukey's tests indicated that the difference in exam 3 score by timing of the practice exam was as predicted with those who began studying earlier earning higher exam scores, as shown in Fig. 6(b). This could be due to how practice exams are more effective when students use them earlier to prepare for exams. This result could also be explained by how students with higher proficiency tend to choose to start practicing earlier. Interestingly, the effect of the timing of practice exams on exam scores was not found for exam 2. It is not clear why engaging with the practice exams earlier did not impact scores on exam 2, but did on exam 3. In both cases, students who started earlier completed about two practice exams before the prediction survey and an additional one or two after the prediction survey, while those who began the practice exams after the prediction survey completed two to three exams.

To examine the effect of the number of practice exam questions attempted, we conducted two multivariate linear regressions with exam score as the response variable and the number of practice exam questions attempted and practice exam condition as the independent variables. Analysis of the q-q plots and distributions of the residuals indicated that the residuals were relatively normally distributed. For both exam 2 and exam 3, students who attempted more problems tended to earn higher exam

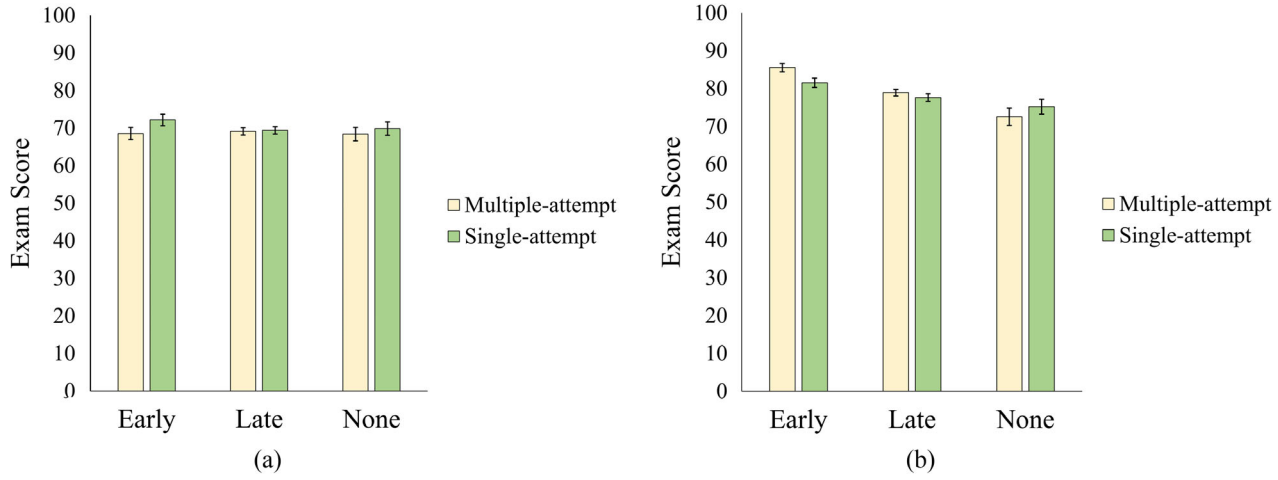


FIG. 6. Comparison of mean exam scores for (a) hour exam 2 and (b) hour exam 3. Note: Error bars represent 1 standard error of the mean.

TABLE IV. Simple linear regression of exam score for exams 2 and 3.

Exam 2				
Variable	*B*	*SE B*	*t*	*p*
Intercept	67.73	0.98	69.14	<0.001
Number of practice exam problems	0.03	0.01	2.17	0.03
Single-attempt practice exam	1.17	1.07	1.09	0.27
Exam 3				
Source	*B*	*SE B*	*t*	*p*
Intercept	72.33	0.93	77.5	<0.001
Number of practice exam Problems	0.14	0.01	9.63	<0.001
Single-attempt practice exam	−1.13	1.01	1.12	0.26

scores, but the format of practice exams had no significant effect. See Table IV for detailed results.

In conclusion, we failed to find a significant exam score difference between students who received multiple-attempt and single-attempt format for either exam. This is perhaps not surprising given the similarity in their exam preparation behaviors and metacognitive bias. In other words, while the new format of the practice exams provided more accurate feedback, students tended to engage with the different practice exam formats in very similar ways.

VI. METHODS (INTERVENTION 2)

A. Description of intervention 2: Adding a mock exam

The second intervention targeted students' preparation for hour exam 1 and was applied to the entire class. One week before the actual exam, students were offered the chance to take a mock exam online, which included 13 multiple-choice questions similar to the ones that would be on the exam. Although optional, students were encouraged to take the mock exam to become familiar with the new environment. In addition, if students scored higher on their mock exam than the actual exam, the mock exam would count toward 25% of their exam 1 score. The potential bonus scoring incentivized participation in the mock exam because a low score would not negatively impact exam grade, but a high score could potentially increase exam grade. The online practice exams (in single-attempt format) were made available to students one week before the mock exam (two weeks before the actual exam). Students who chose to take the mock exam received feedback on how well they did the next day.

The mock exam provides students with a realistic exam-taking experience and an evaluation of their proficiency one week ahead of the actual exam so that students can have better metacognitive monitoring and plan their exam preparation activities accordingly. For example, students who did not do as well as they wanted in the mock exam can plan to start studying earlier.

B. Data collection and analysis

For the second intervention, we gave a survey after students had just received their mock exam scores, five days before when they took the actual exam, measuring students' metacognitive judgment at the early stage of their exam preparation. To understand the connection between students' metacognition and studying plan, we also asked them how many practice exams they plan to do in the next five days on the same survey. We also collected students' mock exam scores and exam scores.

To understand the effect of the mock exam, we first compared exam preparation behaviors between a semester with a mock exam (Spring 2021) and a semester when a mock exam was not offered (Spring 2019). This comparison may be affected by the fact that the course was in-person in Spring 2019, which we provide further explanations for in the findings section. We did an additional analysis with students' prediction survey answers and their mock exam performance to further understand our results. We included all 1112 students who completed hour exam 1 in Spring 2021 in the analysis. Since the mock exam was optional, 847 students completed the mock exam, and out of those, 801 completed the prediction survey.

VII. FINDINGS (INTERVENTION 2)

A. The effects of adding a mock exam on exam prep behaviors

In terms of exam preparation behaviors, we found that adding a mock exam did not change the number of practice questions students chose to do. Comparing data collected in Spring 2021, when we implemented the mock exam, to data collected from the same course in Spring 2019, a semester when the mock exam was not offered to students, we see that the distribution of the total number of practice questions attempted was almost identical (see Fig. 7).

In Spring 2019 (without mock exam), there were a total of 1077 students who did 41 practice questions on average. In Spring 2021 (with mock exam), there were a total of 1112 students who did 40 practice questions on average. In both plots in Fig. 7, the left peak is the percentage of students that did not use any practice exams, and the right peak is the percentage of students that attempted all available practice exam questions. Four practice exams were available to students, each containing about 23 questions, so it is natural for some students to stop practicing at the end of one full practice exam, which explains the smaller peaks in the middle. It is worth noting that, in Spring 2021, students did extra practice questions one week before the actual exam (when they took the mock exam), which we did not include in the graph.

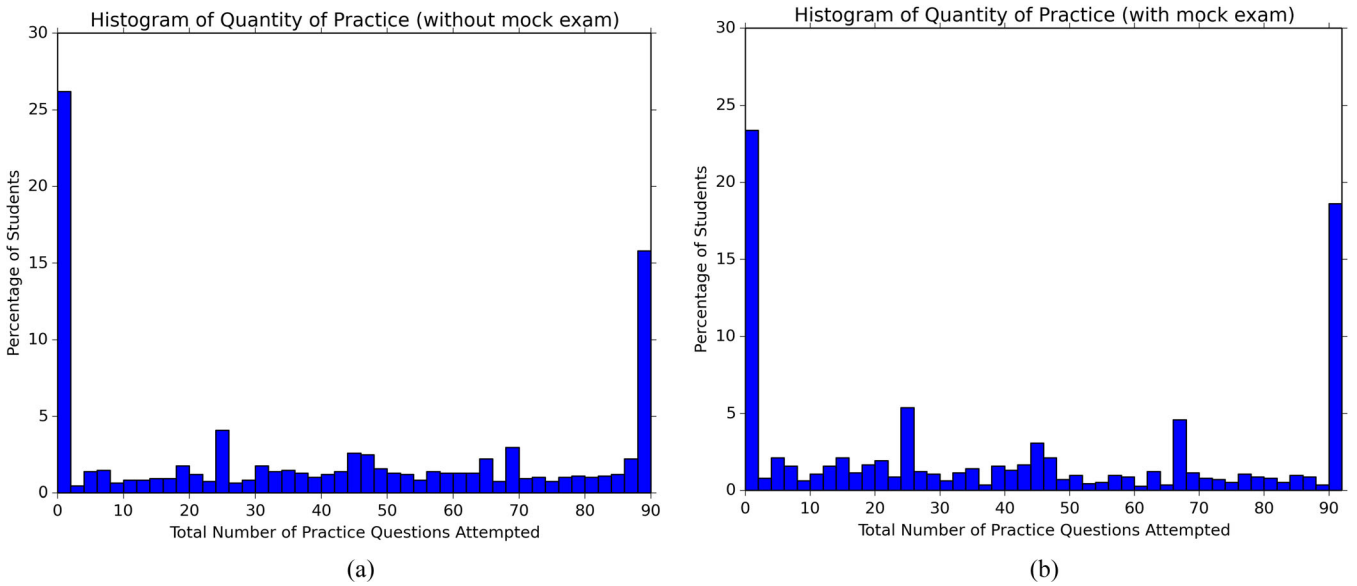


FIG. 7. Distribution of the number of practice exam questions attempted by students, without the mock exam (a) and with the mock exam (b). The height of each bar represents the percentage of students whose number of attempted problems falls in that bin on the x axis. In Spring 21 (b), students had access to one more practice question than in Spring 19 (a), so the maximum number of questions attempted was 90 instead of 89.

We also found that adding a mock exam did not lead to a qualitative change in the timing of students' practice exam use. Although the intervention was designed to incentivize earlier and more evenly distributed practice over the week before the exam, we saw similar cramming behavior. With and without intervention, most students started their practice within 48 hours of the actual exam, as shown in Fig. 8.

We notice a slightly worse cramming behavior in Spring 2021 compared to Spring 2019. This may be due to the course being fully online in Spring 2021, whereas in Spring 2019, the course had fully in-person instruction. This may also be explained by there being a homework due at 8 am on the day of the exam in Spring 2021, so students might have been working on that homework instead of doing practice exams, which could explain the extra practice on the day of the exam.

We did find that, in Spring 2021, there was a small increase in practice exam use about one week before the exam compared to Spring 2019. This change came from a small portion of students using the practice exams to study for the mock exam, leading to a bump in practice about one week before exam time in Fig. 8(b). However, this effect went away after the mock exam was over, so the mock exam did not impact students' practice timing pattern overall.

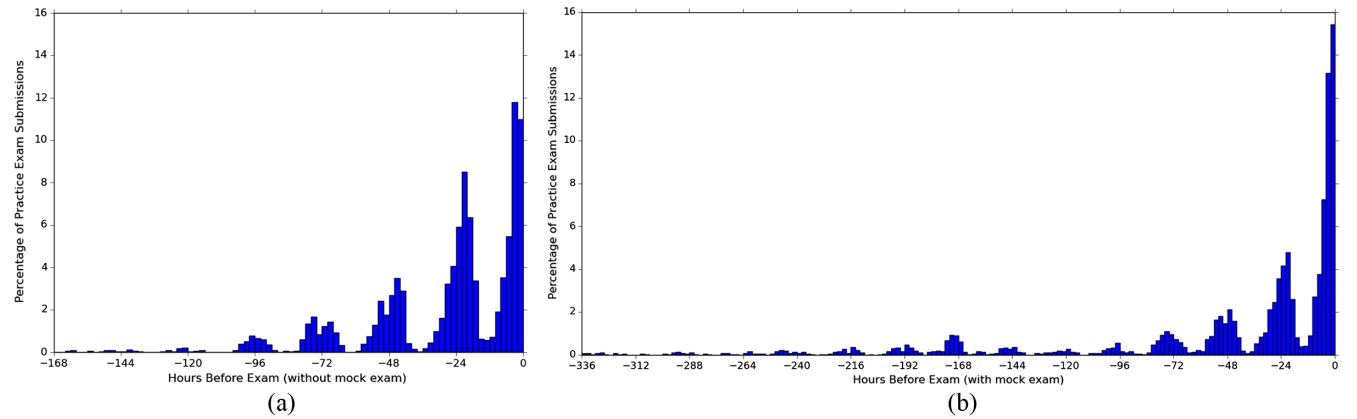


FIG. 8. Timing of online practice activities. The height of each bar represents the percentage of online submissions that falls in the time bin on the x axis. In Spring 2019 (a), there were a total of 47 686 submissions (whenever students click on anything in the online practice exam system). In Spring 2021, with mock exam intervention (b), there were a total of 47 947 submissions. In Spring 2021, we made the practice exams available to students 2 weeks (336 h) ahead of the exam so that they could practice for the mock exam.

TABLE V. Descriptive statistics for predicted learning by expected practice test use.

		Predicted learning		
Practice tests predicted	N	Median	Mean	SE
0	11	15.6	15.75	5.79
1	68	20.7	21.85	2.07
2	224	25.4	25.67	1.35
3	200	29.1	29.03	1.54
4	281	29.5	29.66	1.21

B. Effects of mock exam on students' metacognition

Analysis of the prediction survey data reveals several possible reasons that the mock exam did not affect students' overall exam preparation behavior. First, we found that the number of practice exams students predicted they would do is only weakly correlated with predicted improvement, $r = 0.13$, $p < 0.01$, which was calculated by subtracting their mock exam score from their predicted exam score. To examine the difference in predicted learning by predicted practice test usage, an ANOVA was conducted. ANOVA was used instead of regression because predicted practice test usage was not continuous (i.e., only whole numbers were options on the survey) and linearity could not be assumed. Examination of the q-q plots indicated normality could be assumed, and Levene's test for homogeneity of variance indicated homogeneity can be assumed, $F(4, 779) = 1.21$, $p = 0.30$. The ANOVA indicated differences in predicted learning, $F(4, 779) = 3.74$, $p = 0.30$, $\eta^2 = 0.02$ (see Table V). *Post hoc* Tukey tests indicate that those who predicted they would complete four practice tests expected significantly more learning than those who predicted they would complete one practice test. However, since only a small number of students predicted that they were going to do zero or one practice test, this effect does not show up in the weak correlation shown in Fig. 9(a).

This means that, overall, students who predicted more improvement did not plan on doing a significantly greater number of practice exam problems [see Fig. 9(a)]. We can interpret this finding in two ways: (i) Students do not believe that doing more practice problems will lead to much improvement in exam scores. Although students indicated that practice exams are a very helpful component of the course for them in the course survey, valuing practice exams could mean different things. Students may see it as a way of checking what kind of questions are going to be on the exam or a test of their familiarity with the material, and not necessarily see it as a tool that facilitates learning during exam preparation. (ii) Students are not using mock exams scores as a measurement of their proficiency at the moment, so the predicted score subtracting mock exam score does not reflect the amount of improvement students have in mind for themselves.

Second, we found that students who predicted that they would do more practice questions actually did more practice questions. Because normality and homogeneity of variance could not be assumed, a Kruskal-Wallis test was conducted to examine the relationship between predicted practice test use and actual practice test use. The results indicate that those who predict greater practice test use actually attempt more practice test problems, $\chi^2(4) = 176.91$, $p < 0.0001$.

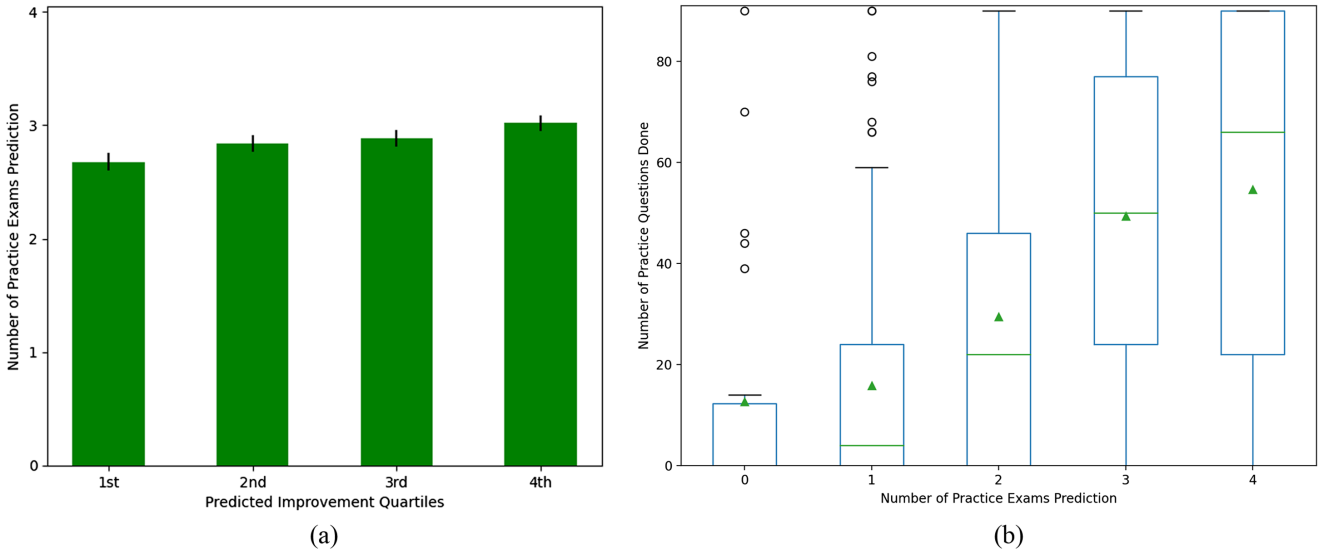


FIG. 9. Prediction survey findings. (a) Students' prediction of how many practice exams they will do given their predicted improvement, which is calculated by subtracting mock exam scores from their prediction of exam scores. Note: Error bars represent one standard error of the mean. (b) Box plot of the quantity of actual practice given predicted practice (triangles are mean values, lines in the boxes are median values). Note: Since each practice exam contains about 22 questions when a student predicts that they will do four practice exams, we interpret that it is equivalent to about 88 practice questions.

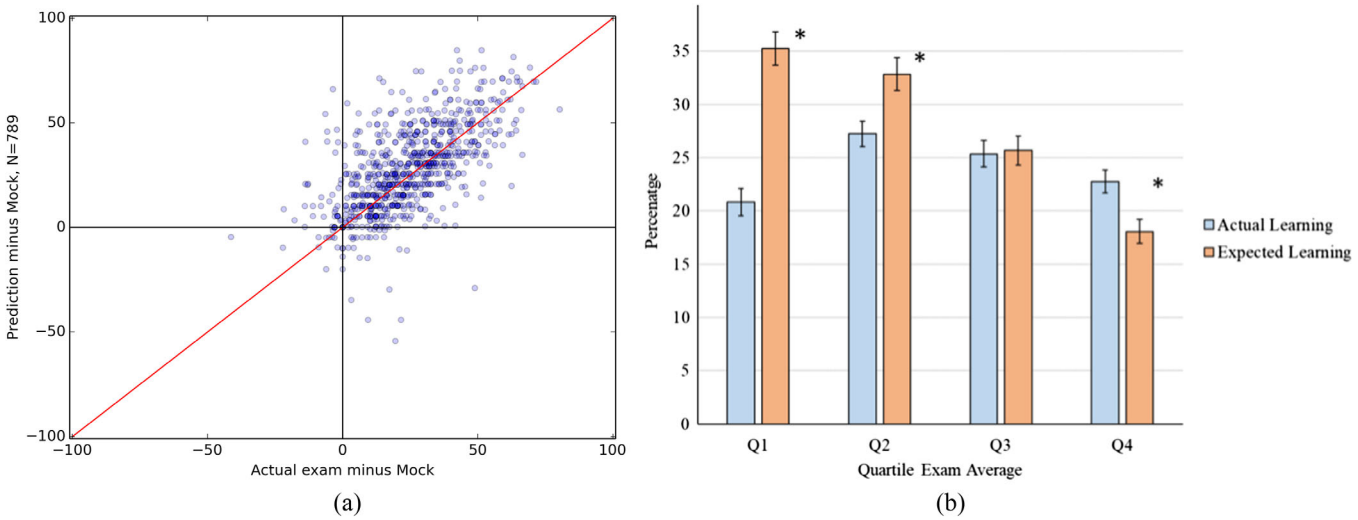


FIG. 10. (a) A scatterplot of students' predicted learning (predicted exam score subtracting mock exam score) given their actual learning (actual exam score subtracting mock exam score). (b) Comparison between students' expected learning and actual learning for different proficiency quartiles.

Post hoc Dwass, Steel, and Critchlow-Flinger tests indicate that those who predict three or four practice tests actually attempt more practice test problems than those who predict zero, one, or two. However, students attempted slightly fewer practice exam questions than they predicted that they would [see Fig. 9(b)], showing a mismatch between their actual study behavior and their initial intention a week before the actual exam.

Third, we found that students' predicted learning correlated with their actual learning overall, $r = 0.63$, $p < 0.0001$, where the actual learning is calculated using their actual exam scores subtracting their mock exam scores [Fig. 10(a)]. To examine the magnitude of the relationship, a linear regression was conducted with predicted learning as the response variable and actual learning as the predictor variable. Examination of the q-q plots and kernel density plots of the residuals indicated a normal distribution of the residuals, and a White test was nonsignificant, $\chi^2(2) = 1.98$, $p = 0.37$, indicating homoscedasticity. The results indicate that for every 10% of learning that was expected, about 7% of learning was observed, $\beta = 0.74$, $F(1, 782) = 515.31$, $p < 0.001$, $\eta^2 = 0.40$.

Given the extensive literature that lower-performing students are less accurate in predicting performance, we examined the relationship between the student's exam scores and learning (both predicted and observed). Two linear regressions were conducted with predicted learning and actual learning as the response variables and exam score as the predictor variable. Examination of the q-q plots and kernel density plots of the residuals indicated normal distributions of the residuals. A White test was nonsignificant, $\chi^2(2) = 0.58$, $p = 0.75$, for actual learning indicating homoscedasticity. However, the White test was significant, $\chi^2(2) = 9.45$, $p = 0.01$, for predicted learning indicating heteroscedasticity. Examination of the residuals for predicted learning indicated that the heteroscedasticity is the result of a few extremely low scores, but this does not appear to impact the regression parameters. The regressions indicated that higher-performing students experienced more learning, $\beta = 0.26$, $F(1, 782) = 42.10$, $p < 0.001$, $\eta^2 = 0.05$, with a moderate effect size. However, in contrast, lower-performing students predicted that they would learn more, $\beta = -0.45$, $F(1, 782) = 97.27$, $p < 0.001$, $\eta^2 = 0.11$, with a large effect size.

To more easily visualize the discrepancy between student expectations and actual performance, we divided the students into quartiles based on their exam scores and plotted the mean predicted learning and actual learning observed in each quartile [Fig. 10(b)]. Four paired t tests were conducted to examine differences between individuals' predicted learning and their actual learning. Bonferroni corrections were applied such that $\alpha = 0.013$ for these paired t tests. The results indicated that students in the lower two quartiles significantly overestimated how much they would improve over the week, $t(154) = 14.56$, $p < 0.0001$, and $t(138) = 5.28$, $p < 0.0001$, respectively, while the higher two quartiles did not exhibit an overconfidence in the predicted amount of learning. Again, this could be due to students not seeing their mock exam score as an accurate assessment of their proficiency.

In summary, our additional analysis shows that (i) students may not be using the feedback they received from the mock exam to set appropriate goals or lay out plans for their studying, (ii) they may not necessarily associate doing more practice exam problems with improved performance, and (iii) there is a mismatch between their intended studying and their actual studying. All of these factors combined explain the finding that the mock exam

intervention did not have an effect on how students interacted with the practice exams, and in particular, did not reduce the cramming behavior as we expected.

C. Testing the theoretical model

To examine the theoretical model used to ground this study (Fig. 1), a path analysis was conducted using the students who both took the mock exam and made predictions (see Fig. 11). A path analysis is a statistical technique that is a special case of structural equation modeling, used to evaluate causal models. Path analyses allow researchers to decompose correlations into direct and indirect effects. Direct effects represent the direct effect that one variable has on another as predicted in theoretical models and are represented by the arrows (as seen in Fig. 11). Indirect effects represent the indirect effect that one variable has on another through their relationship with one or more other variables. These indirect effects can be visualized by following the arrows, or path, from one variable to another by going through other variables. Pairwise correlations are the sum of the direct effects and the indirect effect, which can be found by multiplying the standardized path coefficients along the indirect paths. A fully specified model contains directional paths between all variables and final models are compared to the fully specified models to establish that there is no evidence for a difference in model fit. Rather than finding the best fitting model, path analyses aim to evaluate whether the data are consistent with a prespecified theoretical model. It should be noted that because path analyses are based on the decomposition of correlations, the hypothesized causal relationships between variables are not validated by the path analysis. Rather conclusions from path analyses can only establish that the data are consistent or inconsistent with a given theoretical model.

Because students were only incentivized but not required to participate in the mock exam, we first examined differences between students who completed the mock exam and those who did not. We found that 847 (76.2%) students chose to take the mock exam, while 265 (23.8%) students chose to not complete the mock exam. Students who chose to take the mock exam not only had significant higher actual exam scores ($81.8\% \pm 0.51$) than students who did not take the mock exam ($74.3\% \pm 1.03$), $t(1110) = 7.03$, $p < 0.001$ but also students who took the mock exam also did almost twice as many practice exam problems (44.6 problems $\pm$ 1.19), $t(1110) = 7.95$, $p < 0.0001$, and completed their first practice exam earlier (53.4h $\pm$ 2.4), $t(1110) = 7.15$, $p < 0.0001$, than students who chose to not participate in the mock exam (25.8 problems $\pm$ 1.80,20.4 h $\pm$ 2.5). It should be noted that students self-selected whether to complete the mock exam, so causal conclusions should be avoided when interpreting these results.

Of the 847 students who completed the mock exam, 801 made predictions about both their exam score and their exam preparation. Seventeen of these students made predictions of 0% or above 100%. These students were removed from the dataset leaving a sample size of 784 for the path analysis. Path analyses were conducted using proc Calis in SAS version 9.4. Model fit was assessed using the benchmarks proposed by Hu and Bentler, standardized root mean square residual (SRMR) < 0.08, root mean square error of approximation (RMSEA) < 0.06, comparative fit index (CFI) > 0.95, and normed fit index (NFI) > 0.95 [76].

We examined how the relationship between a student's score on the mock exam and the actual exam was mediated by their metacognitive monitoring (as measured by their prediction) and metacognitive control (as measured by the expected and actual number of practice exam problems completed). A fully specified version of the theoretical model shown in Fig. 11 was tested and the nonsignificant paths were removed. Chi-square tests were nonsignificant indicating that the model shown does not fit the data worse

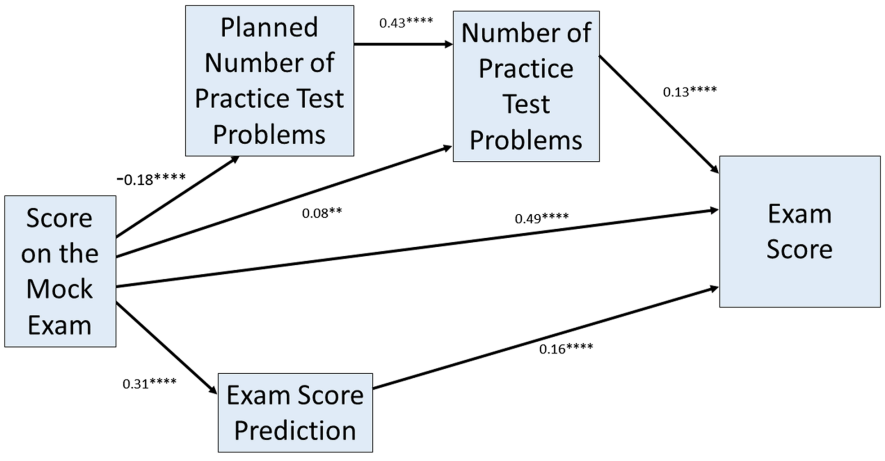


FIG. 11. Path analysis of mediating variables that connect students' mock exam scores and their actual exam scores. Note: $* < 0.05$, $** \leq 0.01$, $*** \leq 0.001$, $**** \leq 0.0001$.

than the fully specified model, $\chi^2(3) = 0.44$, $p = 0.93$. In addition, the fit indices indicate that the theoretical model fit the observed data (SRMR = 0.0055, RMSEA = 0.0000, CFI = 1.000, NFI = 0.9992). The path coefficients shown in Fig. 11 are standardized coefficients allowing for the strengths of the paths to be directly compared to each other.

The path analysis indicates that, consistent with our theoretical model, the score earned on the mock exam was related to students' metacognitive monitoring, practice exam use, and performance on the exam. Unsurprisingly, students who scored higher on the mock exam were more confident and did better on the actual exam one week later. As expected, there was a negative relationship between the mock exam score and planning on doing more practice exams. In other words, those who scored lower on the mock exam planned on doing more practice exam problems than those who score higher on the mock exam. However, there was a weak positive relationship between the score on the mock exam and the actual number of practice exam problems completed. In other words, those who scored higher on the mock exam actually did more practice exam problems.

There was also a positive relationship between the number of practice problems attempted and scores on the exam, which seems to indicate a positive benefit for completing practice exams as expected. In addition to the significant direct effects shown, there were significant indirect effects of the mock exam score on the number of practice test problems completed ($p < 0.001$) and on the exam score ($p < 0.001$). These indirect effects provide evidence that supports the theoretical association between metacognition (monitoring and control) and exam performance. One expected correlation that was surprisingly not significant in our data was the relationship between metacognitive monitoring as measured by predictions and one aspect of metacognitive control as measured by the number of practice problems completed. The correlation between the exam score prediction and the number of practice problems completed was not significant. This may be due to the impact of metacognitive beliefs about the effectiveness of testing as a learning strategy. In other words, students who view practice exams as measuring rather than facilitating learning are likely to demonstrate metacognitive control in other ways. Alternatively, the lack of correlation could be due to the presence of the mock exam score in our model or the general overconfidence that is typically found in exam predictions.

Because we found that the introduction of an optional mock exam did not have a significant effect on students' behaviors when compared with a previous semester without a mock exam (see Sec. VII A), we believe that the differences between students who took the mock exam and those who did not may be due to self-selection. This means that students who already tend to do more exam preparation may have been more willing to participate in the mock exam and thus scored higher on the exam. In other words, these students may have performed better even if a mock exam were not given in the course. This finding shows that it is important to consider that optional interventions may not be reaching students who are most in need of the intervention.

VIII. DISCUSSION

We report findings from two interventions designed to support students' exam preparation. The first intervention modified online practice exams such that the way it is delivered to students provided more accurate feedback for students while they are practicing. We found that this intervention did not have a significant effect on students' metacognitive judgment, how much practice exam questions they attempted, or their actual exam performance. The second intervention introduced a mock exam to mitigate the cramming behavior that was impeding the benefit of practice exams. We did not find any significant effect of this intervention on the timing or quantity of practice exam use, or how students planned for their studying over the week before the exam. Further analysis showed that this intervention may not have reached students who struggle with the course.

We interpret this null result in two ways. First, the interventions we implemented were small-scale adjustments to a much larger course that already includes multiple forms of formative assessment, such as online homework and in-class clicker questions, where students receive feedback on their understanding of course material. Students also take weekly low-stakes quizzes that, in total, make up 10% of their final course grade. Therefore, the extra feedback we provided by modifying the format of practice exams or adding a mock exam may not have been enough to generate a measurable difference.

Second, this null result signals complications in our initial theoretical model of self-regulated learning (Fig. 1). The rationales of our interventions were built upon a chain of events in the theoretical model of self-regulated learning: we expected more accurate feedback on task performance to lead to more accurate metacognitive judgment, more accurate metacognitive judgment to lead to more productive exam preparation behaviors, and better exam preparation behaviors to lead to better exam performance. However, the findings in this study show that there are other important factors in this process each step of the way. Particularly, many factors impact students' metacognitive judgments other than the feedback we provide them. This is supported by existing literature showing that many metacognitive monitoring judgments are also driven by the desire for positive outcomes and misconceptions about the normative difficulty of the tasks as well as misconceptions about their own performance [77–79]. As such, students tend to underutilize past exam performance when making predictions about performance on future exams [80]. In

addition, low-performing individuals maintain their unwarranted overconfidence even after receiving feedback concerning their performance and relative skill [81,82].

To conclude, this study shows that effective interventions to support exam preparation require more than adjustments of task structure or added incentives for earlier practice, such as those implemented in this study. The benefits of a single intervention may not be carried out in the complicated process of self-regulated learning and the context of a large course structure. Because the study took place in an online course during the COVID-19 pandemic, it is difficult to draw strong causal conclusions for the reason the null results were observed. It may be that students were less engaged with the interventions due to the mental stress of the online course. Alternatively, the positive effects of the interventions may have been masked by the negative impacts of online instruction. It is important to note that both interventions were conducted during the same semester, so the earlier intervention (intervention 2) may have led to smaller format group differences in the later intervention (intervention 1). However, as the earlier intervention was the same for all students, we would not expect the earlier intervention to impact the findings from the later intervention.

Although the null results may be partially due to the course being online, an effective intervention likely needs to directly address students' metacognition and control of learning [83–85], existing study habits, beliefs about learning, and additional factors that affect how students make goals and study plans [86,87]. This study also suggests it maybe be helpful for future research to further explore the mechanism of self-regulated learning with both quantitative and qualitative data to identify variables that may not have been considered in existing theories and literature, such as using interviews and surveys to better understand how metacognition interacts with students' actual study behaviors.

ACKNOWLEDGMENTS

We would like to thank Sean Golinsky, Joseph Kuang, Alex Nickl, and Jenny Campbell for their important contributions during the early stages of data analysis for this project. We would also like to thank Eric Kuo for his valuable feedback throughout the project. This material is based upon work supported by the National Science Foundation under Grant No. DUE 2021099. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

APPENDIX: REPLICATED ANALYSES FOR INTERVENTION 1

While analyzing data collected from Intervention 1, we divided students into 3 groups based on the timing of their

TABLE VI. Means and standard errors by condition and time to attempt the 1st question.

	Early		Late		No. of practice test	
	Single M (SE)	Multiple M (SE)	Single M (SE)	Multiple M (SE)	Single M (SE)	Multiple M (SE)
			Exam 2			
Meta bias	13.5 (1.3)	16.5 (1.6)	15.5 (1.2)	15.6 (1.2)	11.6 (2.0)	16.1 (1.9)
Exam score	71.6 (1.4)	68.4 (1.5)	69.7 (1.0)	69.2 (1.0)	70.0 (2.0)	69.2 (1.9)
			Exam 3			
Meta bias	1.6 (1.3)	−0.5 (1.4)	2.8 (1.5)	2.2 (1.2)	5.1 (2.4)	10.0 (2.7)
Exam score	80.7 (1.3)	84.5 (1.1)	77.7 (1.0)	79.0 (0.9)	75.2 (2.0)	72.6 (2.3)

TABLE VII. Analysis of covariance of exam score for exams 2 and 3 using time to attempt the 1st question.

Exam 2					
Source	d.o.f.	Type III SS	MS	F	p
Practice exam condition	1	559.75	559.75	1.79	0.18
Timing	2	8.87	4.44	0.01	0.98
Practice exam condition × timing	2	674.39	337.20	1.08	0.34
Error	1084	338 618.70	312.38		
Exam 3					
Source	d.o.f.	Type III SS	MS	F	p
Practice exam condition	1	70.04	70.04	0.24	0.63
Timing	2	7808.32	3904.16	13.22	<0.001
Practice exam condition × timing	2	1697.38	848.69	2.87	0.06
Error	1084	320 194.47	295.39		

TABLE VIII. Analysis of variance of metacognitive bias for exams 2 and 3 using time to attempt the 1st question.

Exam 2					
Source	d.o.f.	Type III SS	MS	F	p
Practice exam condition	1	1234.44	1234.44	3.49	0.06
Timing	2	394.17	197.08	0.56	0.57
Practice exam condition × timing	2	832.95	416.48	1.18	0.31
Error	994	351 599.29	353.72		
Exam 3					
Source	d.o.f.	Type III SS	MS	F	p
Practice exam condition	1	92.10	92.10	0.24	0.63
Timing	2	5003.45	2501.72	6.47	0.002
Practice exam condition × timing	2	1512.07	756.04	1.96	0.14
Error	982	379 506.05	386.46		

TABLE IX. Means and standard errors by condition and time to attempt the 10th question.

	Early		Late		No. of practice test	
	Single M (SE)	Multiple M (SE)	Single M (SE)	Multiple M (SE)	Single M (SE)	Multiple M (SE)
			Exam 2			
Meta bias	13.9 (1.8)	14.6 (2.0)	15.0 (1.0)	16.3 (1.1)	11.6 (2.0)	16.1 (1.9)
Exam score	70.7 (1.9)	69.6 (2.0)	70.3 (0.9)	68.7 (0.9)	70.0 (2.0)	69.2 (1.9)
			Exam 3			
Meta bias	−0.7 (1.7)	−1.9 (1.8)	3.1 (1.2)	1.8 (1.0)	5.1 (2.4)	10.0 (2.7)
Exam score	82.3 (1.5)	86.2 (1.4)	77.8 (0.9)	79.5 (0.8)	75.2 (2.0)	72.6 (2.3)

practice. We used "attempting 20 or more questions" as the criterion for the "early" group and presented descriptive statistics and ANOVA results in Table I–III. Here, we present summaries of descriptive statistics and ANOVA results from replicated analyses using 1 practice question (Table VI–VIII), 10 practice questions (Table IX–XI), and 30 practice questions (Table XII–XIV) as the criterion. This is to show that the significance of our statistical findings is not affected by us choosing 20 as the criterion.

TABLE X. Analysis of covariance of exam score for exams 2 and 3 using time to attempt the 10th question.

Exam 2					
Source	d.o.f.	Type III SS	MS	F	p
Practice exam condition	1	579.62	579.62	1.85	0.17
Timing	2	47.59	23.80	0.08	0.93
Practice exam condition × timing	2	504.02	252.01	0.81	0.45
Error	1084	338749.93	312.50		
Exam 3					
Source	d.o.f.	Type III SS	MS	F	p
Practice exam condition	1	78.77	78.77	0.27	0.61
Timing	2	8783.49	4391.75	14.90	<0.001
Practice exam condition × Timing	2	1435.21	717.61	2.44	0.09
Error	1084	319 452.15	294.70		

TABLE XI. Analysis of variance of metacognitive bias for exams 2 and 3 using time to attempt the 10th question.

Exam 2					
Source	d.o.f.	Type III SS	MS	*F*	*p*
Practice exam condition	1	1237.69	1237.69	3.50	0.06
Timing	2	565.77	282.88	0.80	0.45
Practice exam condition × timing	2	746.57	373.28	1.06	0.35
Error	994	351 523.72	353.65		
Exam 3					
Source	d.o.f.	Type III SS	MS	*F*	*p*
Practice exam condition	1	85.81	85.81	0.22	0.64
Timing	2	5834.46	2917.23	7.56	<0.001
Practice exam condition × timing	2	1386.14	693.07	1.80	0.17
Error	982	378 782.35	385.73		

TABLE XII. Means and standard errors by condition and time to attempt the 30th question.

	Early		Late		No. of practice test	
	Single M (SE)	Multiple M (SE)	Single M (SE)	Multiple M (SE)	Single M (SE)	Multiple M (SE)
			Exam 2			
Meta bias	12.7 (1.5)	16.9 (1.7)	15.8 (1.1)	15.5 (1.2)	16.1 (1.9)	11.6 (2.07)
Exam score	72.5 (1.6)	68.6 (1.7)	69.4 (1.0)	69.1 (1.0)	70.0 (2.0)	69.2 (1.9)
			Exam 3			
Meta bias	−0.4 (1.4)	−1.6 (1.4)	3.4 (1.3)	2.2 (1.1)	5.1 (2.4)	10.0 (2.7)
Exam score	81.5 (1.2)	85.5 (1.1)	77.6 (1.0)	78.9 (0.9)	75.3 (2.0)	72.6 (2.3)

TABLE XIII. Analysis of covariance of exam score for exams 2 and 3 using time to attempt the 30th question.

Exam 2					
Source	d.o.f.	Type III SS	MS	*F*	*p*
Condition	1	307.27	307.27	0.98	0.32
Timing	2	49.90	24.95	0.08	0.92
Condition × timing	2	12.95	6.47	0.02	0.98
Error	1084	339 239.31	312.95		
Exam 3					
Source	d.o.f.	Type III SS	MS	*F*	*p*
Condition	1	105.49	105.49	0.36	0.55
Timing	2	9598.59	4799.30	16.32	<0.001
Condition × timing	2	1318.86	659.43	2.24	0.11
Error	1084	318 835.50	294.13		

TABLE XIV. Analysis of variance of metacognitive bias for exams 2 and 3 using time to attempt the 30th question.

Exam 2					
Source	d.o.f.	Type III SS	MS	F	p
Practice exam condition	1	711.75	711.75	2.01	0.16
Timing	2	608.34	304.17	0.86	0.42
Practice exam condition × timing	2	435.36	217.68	0.62	0.54
Error	994	353 557.25	353.92		
Exam 3					
Source	d.o.f.	Type III SS	MS	F	p
Practice exam condition	1	102.86	102.86	0.27	0.61
Timing	2	6280.42	3140.21	8.15	<0.001
Practice exam condition × timing	2	1351.15	675.57	1.75	0.17
Error	982	378 363.86	385.30		

[1] M. Scott, T. Stelzer, and G. Gladding, Evaluating multiple-choice exams in large introductory physics courses, Phys. Rev. ST Phys. Educ. Res. **2**, 020102 (2006).

[2] A. B. Simmons and A. F. Heckler, Grades, grade component weighting, and demographic disparities in introductory physics, Phys. Rev. Phys. Educ. Res. **16**, 020125 (2020).

[3] G. R. Flanders, The effect of gateway course completion on freshman college student retention, J. Coll. Student Retention **19**, 2 (2017).

[4] T. Perez, J. G. Cromley, and A. Kaplan, The role of identity development, values, and costs in College STEM Retention, J. Educ. Psychol. **106**, 315 (2014).

[5] J. G. Cromley, T. Perez, and A. Kaplan, Undergraduate STEM achievement and retention: Cognitive, motivational, and institutional factors and solutions, Policy Insights from the Behavioral and Brain Sciences **3**, 4 (2016).

[6] T. Dai and J. G. Cromley, Changes in implicit theories of ability in biology and dropout from STEM majors: A latent growth curve approach, Contemp. Educ. Psychol. **39**, 233 (2014).

[7] B. King, Changing college majors: Does it happen more in STEM and do grades matter?, J. Coll. Sci. Teach. **44**, 44 (2015), https://www.jstor.org/stable/43631938.

[8] J. A. Kulik, C.-L. C. Kulik, and R. L. Bangert, Effects of practice on aptitude and achievement test scores, Am. Educ. Res. J. **21**, 435 (1984).

[9] W. Fakcharoenphol, E. Potter, and T. Stelzer, What students learn when studying physics practice exam problems, Phys. Rev. ST Phys. Educ. Res. **7**, 010107 (2011).

[10] M. Zhang, A. Engel, T. Stelzer, and J. W. Morphew, Effect of online practice exams on student performance, *presented at PER Conf. 2019, Provo, UT*, 10.1119/perc.2019.pr.Zhang.

[11] J. (1) Kruger and D. (2) Dunning, Unskilled, and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-asessments, J. Pers. Soc. Psychol. **77**, 1121 (1999).

[12] J. W. Morphew, Changes in metacognitive monitoring accuracy in an introductory physics course, Metacogn. Learn. **16**, 89 (2021).

[13] N. S. Rebello, How accurately can students estimate their performance on an exam and how does this relate to their actual performance on the exam?, AIP Conf. Proc. **1413**, 315 (2012).

[14] J. A. Greene and R. Azevedo, A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions, Rev. Educ. Res. **77**, 334 (2007).

[15] P. H. Winne and A. F. Hadwin, Studying as self-regulated learning, in *Metacognition in Educational Theory and Practice* (Lawrence Erlbaum Associates Publishers, Mahwah, NJ, 1998), pp. 277–304.

[16] P. H. Winne and A. F. Hadwin, The weave of motivation and self-regulated learning, in *Motivation and Self-Regulated Learning: Theory, Research, and Applications* (Lawrence Erlbaum Associates Publishers, Mahwah, NJ, 2008), pp. 297–314.

[17] B. J. Zimmerman, Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects, Am. Educ. Res. J. **45**, 166 (2008).

[18] T. O. Nelson and L. Narens, Why investigate metacognition?, in *Metacognition: Knowing about Knowing* (The MIT Press, Cambridge, MA, 1994), pp. 1–25.

[19] J. Dunlosky and K. A. Rawson, Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention, Learn. Instr. **22**, 271 (2012).

[20] J. Metcalfe and B. Finn, Evidence that judgments of learning are causally related to study choice, Psychon. Bull. Rev. **15**, 174 (2008).
[21] L. Mihalca, C. Mengelkamp, and W. Schnotz, Accuracy of metacognitive judgments as a moderator of learner control effectiveness in problem-solving tasks, Metacogn. Learn. **12**, 357 (2017).
[22] K. Desender, A. Boldt, and N. Yeung, Subjective confidence predicts information seeking in decision making, Psychol. Sci. **29**, 761 (2018).
[23] K. Morehead, J. Dunlosky, and N. L. Foster, Do people use category-learning judgments to regulate their learning of natural categories?, Mem. Cogn. **45**, 1253 (2017).
[24] T. C. Toppino, M. H. LaVan, and R. T. Iaconelli, Metacognitive control in self-regulated learning: Conditions affecting the choice of restudying versus retrieval practice, Mem. Cogn. **46**, 1164 (2018).
[25] W. Fakcharoenphol, J. W. Morphew, and J. P. Mestre, Judgments of physics problem difficulty among experts and novices, Phys. Rev. ST Phys. Educ. Res. **11**, 020128 (2015).
[26] *The Nature of Expertise*, edited by M. T. H. Chi, R. Glaser, and M. J. Farr (Psychology Press, New York, 2013).
[27] K. Ohtani and T. Hisasaka, Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance, Metacogn. Learn. **13**, 179 (2018).
[28] T. Schlösser, D. Dunning, K. L. Johnson, and J. Kruger, How unaware are the unskilled? Empirical tests of the "signal extraction" counterexplanation for the Dunning–Kruger effect in self-evaluation of performance, J. Econ. Psychol. **39**, 85 (2013).
[29] A. L. Dent and A. C. Koenka, The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis, Educ. Psychol. Rev. **28**, 425 (2016).
[30] M. Raković, M. L. Bernacki, J. A. Greene, R. D. Plumley, K. Hogan, K. Gates, and A. Panter, Examining the critical role of evaluation and adaptation in self-regulated learning, Contemp. Educ. Psychol. **68**, 102027 (2021).
[31] R. N. Blasiman, J. Dunlosky, and K. A. Rawson, The what, how much, and when of study strategies: Comparing intended versus actual study behaviour, Memory **25**, 784 (2017).
[32] M. K. Hartwig and J. Dunlosky, Study strategies of college students: Are self-testing and scheduling related to achievement?, Psychon. Bull. Rev. **19**, 126 (2012).
[33] T. J. Bing and E. F. Redish, Analyzing problem solving using math in physics: Epistemological framing via warrants, Phys. Rev. ST Phys. Educ. Res. **5**, 020108 (2009).
[34] B. Hegde and B. N. Meera, How do they solve it? An insight into the learner's approach to the mechanism of physics problem solving, Phys. Rev. ST Phys. Educ. Res. **8**, 010109 (2012).
[35] L. N. Walsh, R. G. Howard, and B. Bowe, Phenomenographic study of students' problem solving approaches in physics, Phys. Rev. ST Phys. Educ. Res. **3**, 020108 (2007).
[36] J. H. Larkin, J. McDermott, D. P. Simon, and H. A. Simon, Models of competence in solving physics problems, Cogn. Sci. **4**, 317 (1980).
[37] J. Sweller, Cognitive load during problem solving: Effects on learning, Cogn. Sci. **12**, 257 (1988).
[38] J. Meijer, M. V. J. Veenman, and B. H. A. M. van Hout-Wolters, Metacognitive activities in text-studying and problem-solving: Development of a taxonomy, Educ. Res. Eval. **12**, 209 (2006).
[39] N. Kornell and R. A. Bjork, The promise and perils of self-regulated study, Psychon. Bull. Rev. **14**, 219 (2007).
[40] P. C. Brown, H. L. Roediger III, and M. A. McDaniel, *Make It Stick: The Science of Successful Learning* (Harvard University Press, Cambridge, MA, 2014).
[41] D. Rohrer and K. Taylor, The shuffling of mathematics problems improves learning, Instr. Sci. **35**, 481 (2007).
[42] D. Rohrer, K. Taylor, H. Pashler, J. T. Wixted, and N. J. Cepeda, The effect of overlearning on long-term retention, Appl. Cogn. Psychol. **19**, 361 (2005).
[43] D. Rohrer and K. Taylor, The effects of overlearning and distributed practise on the retention of mathematics knowledge, Appl. Cogn. Psychol. **20**, 1209 (2006).
[44] A. C. Butler, Repeated testing produces superior transfer of learning relative to repeated studying, J. Exp. Psychol. **36**, 1118 (2010).
[45] D. P. Larsen, A. C. Butler, and H. L. Roediger, Repeated testing improves long-term retention relative to repeated study: A randomised controlled trial, Med. Educ. **43**, 1174 (2009).
[46] R. A. Schmidt and R. A. Bjork, New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training, Psychol. Sci. **3**, 207 (1992).
[47] S. K. Carpenter, N. J. Cepeda, D. Rohrer, S. H. K. Kang, and H. Pashler, Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction, Educ. Psychol. Rev. **24**, 369 (2012).
[48] K. A. Rawson, J. Dunlosky, and S. M. Sciartelli, The power of successive relearning: Improving performance on course exams and long-term retention, Educ. Psychol. Rev. **25**, 523 (2013).
[49] H. L. Roediger III, A. L. Putnam, and M. A. Smith, Ten benefits of testing and their applications to educational practice, Psychol. Learn. Motiv. **55**, 1 (2011).
[50] P. Black and D. Wiliam, Assessment and classroom learning, Assessment in Education Principles Policy and Practice **5**, 7 (1998).
[51] C. F. Darley and B. B. Murdock, Effects of prior free recall testing on final recall and recognition, J. Exp. Psychol. **91**, 66 (1971).
[52] H. L. Roediger and J. D. Karpicke, Test-enhanced learning: Taking memory tests improves long-term retention, Psychol. Sci. **17**, 249 (2006).
[53] J. W. Morphew, M. Silva, G. Herman, and M. West, Frequent mastery testing with second-chance exams leads to enhanced student learning in undergraduate engineering, Appl. Cogn. Psychol. **34**, 168 (2020).
[54] B. C. Johnson and M. T. Kiviniemi, The effect of online chapter quizzes on exam performance in an undergraduate social psychology course, Teach. Psychol. **36**, 33 (2009).
[55] M. A. McDaniel, R. C. Thomas, P. K. Agarwal, K. B. McDermott, and H. L. Roediger, Quizzing in middle-school science: Successful transfer performance on classroom exams, Appl. Cogn. Psychol. **27**, 360 (2013).

[56] T. Nip, E. L. Gunter, G. L. Herman, J. W. Morphew, and M. West, Using a computer-based testing facility to improve student learning in a programming languages and compilers course, in *Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (Association for Computing Machinery, New York, NY, 2018), pp. 568–573.

[57] M. A. McDaniel, J. L. Anderson, M. H. Derbish, and N. Morrisette, Testing the testing effect in the classroom, Eur. J. Cogn. Psychol. **19**, 494 (2007).

[58] D. J. Peterson and K. T. Wissman, The testing effect and analogical problem-solving, Memory **26**, 1460 (2018).

[59] R. C. Thomas, C. R. Weywadt, J. L. Anderson, B. Martinez-Papponi, and M. A. McDaniel, Testing Encourages transfer between factual and application questions in an online learning environment, J. Appl. Res. Mem. Cogn. **7**, 252 (2018).

[60] A. C. Butler and H. L. Roediger, Testing improves long-term retention in a simulated classroom setting, Eur. J. Cogn. Psychol. **19**, 514 (2007).

[61] S. H. K. Kang, K. B. McDermott, and H. L. Roediger, Test format and corrective feedback modify the effect of testing on long-term retention, Eur. J. Cogn. Psychol. **19**, 528 (2007).

[62] J. D. Karpicke and H. L. Roediger, Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention, J. Exp. Psychol. Learn. Mem. Cogn. **33**, 704 (2007).

[63] L. Richland, L. S. Kao, and N. Kornell, Can unsuccessful tests enhance learning, in *Proceedings of the Annual Meeting of the Cognitive Science Society* (2008), p. 2338, https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=1dd7f9d95b0ce30a952463d21b20b63f6a396128.

[64] J. C. K. Chan, Long-term effects of testing on the recall of nontested materials, Memory **18**, 49 (2010).

[65] J. L. Little, E. L. Bjork, R. A. Bjork, and G. Angello, Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting, Psychol. Sci. **23**, 1337 (2012).

[66] S. C. Pan and T. C. Rickard, Transfer of test-enhanced learning: Meta-analytic review and synthesis, Psychol. Bull. **144**, 710 (2018).

[67] S. K. Carpenter, L. Mickes, S. Rahman, and C. Fernandez, The effect of instructor fluency on students' perceptions of instructors, confidence in learning, and actual learning, J. Exp. Psychol. Appl. **22**, 161 (2016).

[68] Linda Bol and Douglas J. Hacker, A comparison of the effects of practice tests, and traditional review on performance, and calibration, J. Exp. Educ. **69**, 133 (2001).

[69] W. R. Balch, Practice versus review exams and final exam performance, Teach. Psychol. **25**, 181 (1998).

[70] G. M. Novak, A. Gavrin, E. Patterson, and W. Christian, *Just-in-Time Teaching: Blending Active Learning with Web Technology* (Prentice Hall, Englewood Cliffs, NJ, 1999).

[71] E. Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, Englewood Cliffs, NJ, 1997).

[72] A. Adaramola, A. Godwin, and B. Boudouris, Student outcomes related to academic performance, motivation, and mental health in an online materials and energy balances course during the COVID-19 pandemic, Chem. Eng. Educ. **56**, 36 (2022).

[73] D. Serhan, Transitioning from face-to-face to remote learning: Students' attitudes and perceptions of using Zoom during COVID-19 pandemic, Int. J. Technol. Educ. Sci. **4**, 335 (2020).

[74] K. A. Walker and K. E. Koralesky, Student and instructor perceptions of engagement after the rapid online transition of teaching due to COVID-19, Nat. Sci. Educ. **50**, e20038 (2021).

[75] G. Schraw, A conceptual analysis of five measures of metacognitive monitoring, Metacogn. Learn. **4**, 33 (2009).

[76] L. Hu and P. M. Bentler, Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification, Psychol. Methods **3**, 424 (1998).

[77] J. Ehrlinger, K. Johnson, M. Banner, D. Dunning, and J. Kruger, Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent, Organ. Behav. Hum. Decis. Process. **105**, 98 (2008).

[78] M. J. Serra and K. G. DeMarree, Unskilled and unaware in the classroom: College students' desired grades predict their biased grade predictions, Mem. Cogn. **44**, 1127 (2016).

[79] D. J. Simons, Unskilled and optimistic: Overconfident predictions despite calibrated knowledge of relative skill, Psychon. Bull. Rev. **20**, 601 (2013).

[80] N. L. Foster, C. A. Was, J. Dunlosky, and R. M. Isaacson, Even after thirteen class exams, students are still overconfident: The role of memory for past exam performance in student predictions, Metacogn. Learn. **12**, 1 (2017).

[81] T. M. Miller and L. Geraci, Training metacognition in the classroom: The influence of incentives and feedback on exam predictions, Metacogn. Learn. **6**, 303 (2011).

[82] J. L. Nietfeld, L. Cao, and J. W. Osborne, Metacognitive monitoring accuracy and student performance in the postsecondary classroom, J. Exp. Educ. **74**, 7 (2005), https://www.jstor.org/stable/20157410.

[83] J. D. Stanton, A. J. Sebesta, and J. Dunlosky, Fostering metacognition to support student learning and performance, CBE Life Sci. Educ. **20**, fe3 (2021).

[84] P. R. Husmann and T. C. Smith, Do students know what they think they know? Evaluating the relationships between online practice questions, knowledge monitoring, and course outcomes, Coll. Teach. **70**, 482 (2022).

[85] A. Lionelle, S. Ghosh, S. Ourada, and W. Musser, Increase performance and retention: Teach students how to study, in *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 1* (Association for Computing Machinery, New York, NY, 2022), pp. 349–355.

[86] J. B. Jenifer, C. S. Rozek, S. C. Levine, and S. L. Beilock, Effort(less) exam preparation: Math anxiety predicts the avoidance of effortful study strategies, J. Exp. Psychol. **151**, 2534 (2022).

[87] A. Kirk-Johnson, B. M. Galla, and S. H. Fraundorf, Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice, Cogn. Psychol. **115**, 101237 (2019).