Peer vs. Self-Grading of Practice Exams: Which Is Better?

Mallory A. Jackson, Alina Tran, Mary Pat Wenderoth, and Jennifer H. Doherty*
Department of Biology, University of Washington, Seattle, WA 98105

ABSTRACT

Practice exams are a type of deliberate practice that have been shown to improve student course performance. Deliberate practice differs from other types of practice, because it is targeted, mentally challenging, can be repeated, and requires feedback. Providing frequent instructor feedback to students, particularly in large classes, can be prohibitive. A possible solution is to have students grade practice exams using an instructor-generated rubric, receiving points only for completion. Students can either grade their own or a peer's work. We investigated whether peer or self-grading had a differential impact on completion of practice exam assignments, performance on practice exams or course exams, or student grading accuracy. We also investigated whether student characteristics mattered. We found that 90% of students took all practice exams or only missed one and that there was no difference on practice or course exam performance between the peer and self-graders. However, in the peer-grading treatment, students with lower incoming grade point averages and students identified as economically or educationally disadvantaged were less accurate and more lenient graders than other students. As there is no clear benefit of peer grading over self-grading, we suggest that either format can solve the challenge instructors face in giving frequent personalized feedback to many students.

INTRODUCTION

One way for students to gain expertise in a discipline is through deliberate practice (Ericsson *et al.*, 1993; Ericsson and Charness, 1994). Deliberate practice differs from ordinary practice in that deliberate practice is designed specifically to improve a targeted performance, provides continuous feedback, has a high mental demand, and can be repeated (Colvin, 2008). Deliberate practice targets gaps in knowledge, difficult concepts, or other aspects of an activity that are limiting performance. Feedback, often from an external source with expertise in the area, helps to provide objective assessment of how well performance during practice meets established criteria. Deliberate practice has been hypothesized to explain the beneficial impact of highly structured active-learning courses (Haak *et al.*, 2011; Eddy and Hogan, 2014).

One type of deliberate practice used in college courses is the practice exam, a homework assignment in which students take practice or old exam questions. These focused assignments are aligned with the exam format (e.g., multiple choice, open-ended), give students a chance to test themselves on course material using deliberate practice (Fakcharoenphol and Stelzer, 2014), and have been shown to improve course performance (Cheng *et al.*, 2004; Freeman *et al.*, 2007, 2011; Trussell and Dietz, 2013). In some instances, even one practice exam has been shown to improve exam performance (Balch, 1998).

Practice exams meet many of the criteria for deliberate practice, as they are targeted performance (similar to exam questions), require mental effort (as should exam questions), and can be given often. However, for faculty who use open-ended exam and practice exam formats, the main challenge to implementing this type of deliberate practice lies in providing feedback to the student (Adachi *et al.*, 2017). If practice exams contain many questions and are given weekly, the instructor

Brian Sato, *Monitoring Editor*
Submitted Apr 5, 2018; Revised Jun 18, 2018; Accepted Jul 3, 2018
CBE Life Sci Educ September 1, 2018 17:ar44
DOI:10.1187/cbe.18-04-0052
*Address correspondence to: Jennifer H. Doherty (doherty2@uw.edu).

workload associated with providing feedback on each student's performance on each question quickly becomes prohibitive, particularly in large classes.

A possible solution for providing frequent feedback is to have the students grade the practice exam using an instructor-generated rubric. Students can either be assigned to grade their own (i.e., self-graded) or another student's work (i.e., peer graded). There are two potential benefits of students grading practice exams using the instructor's rubric. 1) With a decreased feedback workload, instructors can assign more practice exam questions, thus providing the students with more deliberate practice opportunities (Sadler and Good, 2006; Topping, 2009; Simkin and Stiver, 2016). 2) Student grading has the potential to make students active evaluators rather than passive receivers of grades. Becoming an active evaluator has the possibility of motivating students to learn (Stefani, 1992; Mahlberg, 2015) and helping students elucidate current understanding as compared with the expected level of mastery (Tanner, 2012), thus promoting student's ability to self-regulate their learning (Eccles and Wigfield, 2002). Additionally, peer and self-grading better reflect the changing role of assessment in higher education, which is moving from one in which students are passive recipients of grades to one in which assessments are viewed as valuable tools students can use to monitor and guide them to resources that will strengthen their learning (Dochy *et al.*, 1999). Peer and self-grading are seen as valuable skills students can bring with them to the workplace, skills that will support them in becoming lifelong learners (Sambell *et al.*, 1997).

Peer and self-grading of course work has been correlated with improving course performance. Dochy *et al.* (1999), in a review of 63 studies that used peer and self-grading in higher education, found that the vast majority of students saw benefit in peer assessment, that students who self-assess while learning score higher on tests than those without self-assessment, and that self-assessment leads to more self-reflection and responsibility for one's own learning. Additionally, including written peer-graded exercises in a nonmajors physiology course produced a significant improvement in exam scores (Pelaez, 2002). Students in an introductory biology course with assigned weekly, timed, peer-graded assessments showed an improved course performance compared with prior offerings of the course that lacked this form of formative assessment (Freeman *et al.*, 2007). Similar positive results of peer assessment on student learning, as compared with instructor feedback, have been found in a matched randomized crossover experiment in a college statistics course (Sun *et al.*, 2015). Therefore, self-grading may provide students with the opportunity to reflect on their current state of knowledge and possibly improve their metacognition, while peer grading may provide students with models of more sophisticated answers that they could emulate.

When using peer- or self-grading methods, instructors and students are often concerned about grading accuracy (Liu and Carless, 2006). Researchers have indeed documented that students are less accurate and easier graders (Freeman and Parks, 2010; Panadero *et al.*, 2013; Simkin, 2015). Given that students are novice learners, grading, even with a detailed rubric, can be a challenge, as they have a tendency to be very literal and unaware of alternative phrasing of an answer. Furthermore, the Dunning-Kruger effect predicts that underperforming students often overestimate their performance, as they lack sufficient metacognitive skills, and this leads to cognitive bias (Kruger and Dunning, 1999). Therefore, by overscoring their own or a peer's work, low-performing students may further overestimate their understanding of the material, which could negatively impact their preparation for the actual exam.

Given that students have a tendency to overestimate their performance when peer or self-grading, these practice exam scores would inflate their course grades. An alternative option is to have students complete the peer or self-grading for participation points only, also known as effort-based grading (Schinske and Tanner, 2014). However, awarding points for mere completion of the assignment may disincentivize students from completing the assessment or fully engaging their intellectual effort.

The current study examines the following questions: 1) Do students complete all or most of the weekly practice exams even though they only receive points for completion and are not graded by the instructor? 2) Does peer or self-grading have a differential impact on student performance on practice exam questions or on course exams? 3) Does peer or self-grading have a differential impact on student grading accuracy? Given that student characteristics, such as college ability and demographics (e.g., gender, race/ethnicity, and socioeconomic status), have been shown to impact course performance, we investigated whether these student characteristics impacted our results (Freeman *et al.*, 2011; Haak *et al.*, 2011; Eddy and Hogan, 2014).

METHODS

Participants

Undergraduate students ($n = 550$) enrolled in one section of the final quarter of the three-quarter introductory biology sequence at the University of Washington (UW), a large R1 university, took part in the study. The course, BIOL 220: Introductory Biology III, is an introduction to animal and plant physiology and is taught by one of the authors (J.H.D.). The course has a laboratory portion with ~24 students per lab. Laboratory sections are taught by teaching assistants (TAs) who each teach an even number of sections, two, four, or six. Laboratory sections were randomly assigned to either the peer- or self-grading treatment group, stratified by TA. Students remained in either the peer- or self-grading group the entire quarter.

Student demographic information was obtained from the registrar and included gender (63% female), incoming grade point average (GPA) at start of the term, participation in the UW Educational Opportunity Program (EOP; i.e., students identified as economically or educationally disadvantaged; 14%), whether the student was from a race/ethnicity that is underrepresented in science (underrepresented minority [URM]; i.e., African American, Hispanic, Native American, or Pacific Islander; 7%), and whether the student's parents did not graduate from college (i.e., first-generation college students; 14%).

Practice Exam Implementation

Practice exams were administered weekly online in our course management system (the course lasted a total of 9 weeks). They consisted of two assignments, the answering assignment and the grading assignment. These assignments were completed on consecutive days to ensure that all students completed the

answering assignment before peer graders were randomly assigned answers. Each answering assignment consisted of three exam questions from prior offerings of the course. Each exam question was presented individually, and students typed their answers into a response box before moving on to the next question. The students were given 24 hours to complete the answering assignment online. Students were told to complete this assignment on their own and without using books or notes. For example, during week 2, students received the following question plus two others:

> (8 points) Predict what would happen to skeletal muscle contraction if there are twice as many Ca^{++} pumps on the SR. Defend your answer using Mass Balance reasoning.

On Friday, all students went back online to complete the grading assignment. In the grading assignment students were instructed to use a detailed rubric created by the course instructor for each question to score each question, to enter that score in a box and to add a comment to a comment box. Students in the self-grading group opened both their answering assignment (to see their answers) and the grading assignment to grade their own answers. Students in the peer-grading group were randomly assigned a peer's answers by the course management system and used the same rubric as the self-graders to assign the number of points earned and give comments. Students could, but were not obligated to, look at their peers' comments, and we had no way of tracking if they did or not. New random peers were assigned each week. Practice exam questions ranged in point value from 8 to 30 points. For example, in week 2, students received the following detailed rubric:

> (8 points total) Twice as many Ca^{++} pumps on the SR will double the rate at which Ca^{++} leaves the cytoplasm (2 points). If the rate out of the cytoplasm increases and the rate into the cytoplasm remains the same (2 points), the amount of Ca^{++} in the cytoplasm will not peak as high and will decrease faster (2 points). Therefore the strength and length (duration) of muscle contraction will decrease (2 points).

Students awarded partial credit as indicated on the rubric. Though students awarded a point value for each practice exam question, that score did not contribute to their course grade, as students received full credit (10 points) for both taking and grading the practice exam. Students were not initially given examples of completed grading assignments to train them how to grade. As we wanted students to value the practice exam assignment, the total participation points for practice exams constituted 10% of the course grade.

All practice exam questions and rubrics are included in the Supplemental Material.

Expert Grading of Practice Exams

After the quarter was over, the first question from each week's practice exam was graded by one of two expert graders (authors M.A.J. and A.T.), who were trained to use the rubric by the course instructor (author J.H.D.). Using expert graders allowed us to investigate how well students answer practice exam questions and whether treatment impacted grading accuracy.

We calibrated expert grading to ensure that each expert grader was consistent. The two expert graders graded the first 50 answers for each question and calculated differences between their assigned expert grades. If any of the expert grades had a difference of greater than 15%, the two expert graders and the instructor worked together to refine the rubric and its interpretation. The two expert graders then graded another 50 answers and checked for differences. The expert graders and the instructor discussed the differences and came to consensus on rubric interpretation. This process was repeated until the difference for each answer was less than 15% of points. Then expert graders each graded half of the remaining answers and 10% of the other's half as a reliability check. As a result, the interrater reliability at this final stage was always less than 15% of the points for that question. Given that practice exam questions ranged from 8 to 30 points, we chose 15% as the acceptable discrepancy between graders, because 15% is roughly equivalent to a 1 point discrepancy in an 8 point question.

We defined student grading accuracy as the difference between student grade and expert grade. If the student was more generous than the expert in grading, a positive value was obtained, while a negative value indicated that the student was a harsher grader than the expert. A small difference indicated greater accuracy of student grading.

To incorporate data from across the quarter, we calculated an average grading accuracy for each student. Because students completed varied numbers of practice exams and practice exams questions were worth different numbers of points, we calculated a student's average grading accuracy by taking the difference between the student and expert grades for each question and dividing by the total points available for each question. We then averaged these percent differences over the total number of practice exams the student completed (a maximum of nine). The equation below summarizes how we calculated average grading accuracy.

$$\text{Average grading accuracy} = \frac{\sum_{n}^{9}\left(\frac{\text{Student grade for practice exam}_n - \text{Expert grade for practice exam}_n}{\text{Points possible for practice exam}_n}\right)}{\text{Number of practice exams student completed}} \quad (1)$$

Modeling Procedure

We used generalized multilevel models for our analyses (Gelman and Hill, 2007). For each research question, we started with the most complex model. For all research questions, this included, as fixed effects, all student characteristics (i.e., incoming GPA, gender, EOP status, URM status, first-generation status) and

TABLE 1. Parameter estimates and SEs (in parentheses) for analysis models

Outcome[a]	Treatment (ref: peer)	EOP (ref: non-EOP)	GPA	Gender (ref: male)	PE score (in percent)	EOP:Trt	GPA:Trt	ΔAICc[b]
Number of PEs taken	—	—	—	—	—	—	—	0
PE performance	—	—	22.4 (1.55)	—	—	—	—	172.2
Course exam performance[c]	—	—	79.3 (4.97)	–3.9 (1.52)	1.2 (0.12)	—	—	485.4
Grading accuracy	2.5 (0.96)	–0.8 (0.96)	–16.9 (1.90)	—	—	–3.5 (0.95)	–10 (1.9)	91

ref, reference level; Trt, treatment.
[a]All models included TA as a random effect. PE, practice exam.
[b]ΔAICc is the difference between the best-fit model and the null model, the intercept-only model that included TA as a random effect.
[c]Course exam performance is based on total exam points for all course exams, with a maximum of 550 points.

grading treatment. To determine whether treatment had a differential impact on students with different characteristics, we also included interactions between grading treatment and student characteristics in this model. As students were randomly assigned to treatment by laboratory sections, stratified by TA, we included TA as a random effect in each model. Additional factors were added to some models (details below in Question 2 methods section). To determine the best-fit model for the analysis, we used backward selection. We sequentially removed the parameter with the highest *p* value, starting with interactions and then moving to main fixed effects. Akaike's information criterion with correction for a finite sample size (AICc) was used to measure the fit of the model. AICc scores were recorded after each sequential model adjustment, and the model with the lowest AICc score was determined to be the best fit to explain the data (see Table 1 for parameter estimates of selected models). Models with a difference in AICc score of ±2 were considered to be equivalent, and in that case, to satisfy guidelines of parsimony, the model with the fewest parameters was selected (Burnham and Anderson, 2002).

Most analyses were carried out using JMP Pro (SAS Institute, 2016). Analysis of the number of practice exams completed was carried out in R (R Core Team, 2017). Model details for each research question are detailed in the following paragraphs.

Question 1: Do Students Complete All or Most of the Weekly Practice Exams, Even Though They Only Receive Points for Completion and Are Not Graded by the Instructor? For research question 1, we modeled the number of practice exams completed using a Poisson distribution, appropriate for count data.

Question 2: Does Peer or Self-Grading Have a Differential Impact on Student Performance on Practice Exam Questions or on Course Exams? For research question 2, we ran two linear models: one for average expert practice exam grade and one for total course exam grade (based on total exam points for all course exams, maximum of 550 points). For both student performance models, we also included average grading accuracy as a fixed effect. For the total course exam grade model, we also included average expert practice exam score in our most complex model.

Question 3: Does Peer or Self-Grading Have a Differential Impact on Student Grading Accuracy? For research question 3, we modeled the average percent difference between student grade and expert grade using a linear model.

This research has been approved by the Human Subjects Division of the University of Washington (application #51527).

RESULTS

Question 1: Do Students Complete All or Most of the Weekly Practice Exams, Even Though They Only Receive Points for Completion and Are Not Graded by the Instructor?

The majority of students completed all of the practice exams. Of the 550 students, 75% completed all nine, and 90% completed at least eight. The model that best explained the number of practice exams completed was the intercept-only model (Table 1). Therefore, practice exam completion was not influenced by incoming GPA; gender; EOP, URM, or first-generation status; or treatment.

Question 2: Does Peer or Self-Grading Have a Differential Impact on Student Performance on Practice Exam Questions or on Course Exams?

The average expert grade across all nine practice exams was 47%. The model that best explained student performance on practice exam questions included incoming GPA as a fixed effect (Table 1). Therefore, there was no differential impact of the peer- or self-grading treatment on practice exam score. Students with higher GPAs earned more points, on average, than students with lower GPAs. There was no influence of average grading accuracy, gender, or EOP, URM, or first-generation status on practice exam performance.

On average, students performed better on course exams (mean = 70.8%) than on practice exams. The best model explaining total course exam performance included incoming GPA, gender, and expert grade of practice exam as fixed effects. Therefore, there was no differential impact of the peer- or self-grading treatment on course exam performance (see Figure 1). There was no influence of average grading accuracy or EOP, URM, or first-generation status on course exam performance. GPA is positively correlated with exam performance, as students with higher incoming GPAs earn more exam points. Females earned an average of about 1% fewer points (3.9 exam points) than males on course exams. When controlling for GPA and gender, average expert grade on practice exams is still positively correlated with total exam points.

Question 3: Does Peer or Self-Grading Have a Differential Impact on Student Grading Accuracy?

Grading accuracy was assessed using average percent difference in practice exam grade between students and experts (see

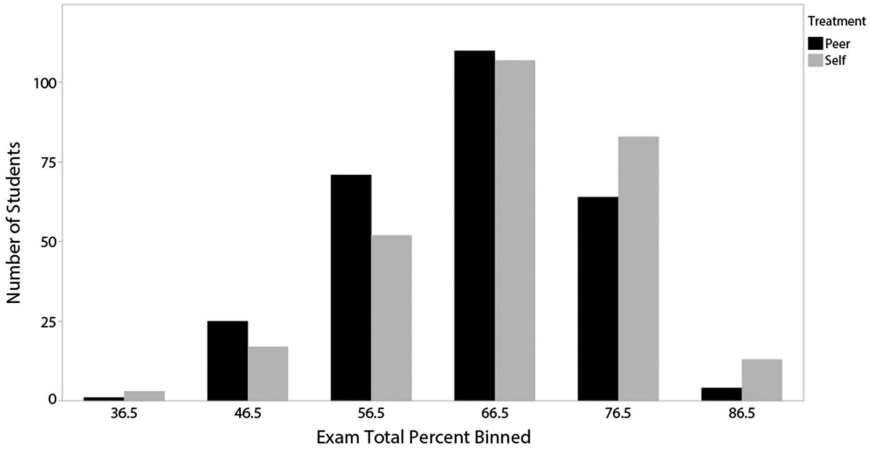

FIGURE 1. Histogram of course exam scores (in percent of total) for students in peer- and self-grading treatment groups. Raw scores are plotted, not model output. There was no overall effect of treatment on course exam performance when controlling for GPA.

Eq. 1 in *Methods*). A positive value indicates that the student awarded more points than an expert, while a negative value indicates the student awarded fewer points than the expert. Overall, most students awarded more points than experts (mean = 9.3% difference). The model that best explained grading accuracy included EOP status, incoming GPA, grading treatment, interaction of grading treatment and EOP, and interaction of grading treatment and GPA as fixed effects (Table 1). EOP students in the peer-grading treatment awarded more points on practice exams than EOP students in the self-grading treatment, while non-EOP students awarded similar numbers of points (Figure 2). Students in the peer-grading treatment with lower GPAs awarded more points than experts, and peer graders with higher GPAs awarded fewer points than experts (Figure 3). GPA had a smaller impact on the grading accuracy of students who self-graded.

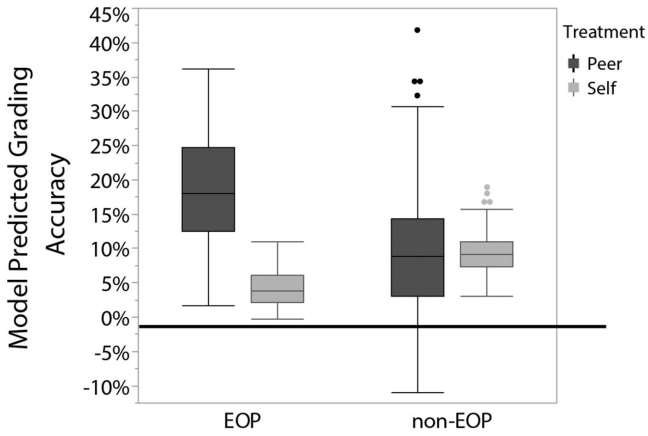

FIGURE 2. Interaction between EOP status and grading treatment on grading accuracy. Model output is graphed. See Eq. 1 in *Methods* for grading accuracy calculation. Line at zero designates complete accuracy compared with expert. Values greater than zero indicate more lenient grading; below zero is harsher grading than experts. EOP students who grade their own practice exams grade more similarly to experts than EOP students who grade peers' work.

DISCUSSION

We implemented weekly online practice exams to maximize student learning through deliberate practice. To decrease the instructor workload of providing feedback to each student in a large-enrollment course, we assigned students to either peer or self-grade the weekly practice exams using a detailed instructor-generated rubric.

Question 1: Do Students Complete All or Most of the Weekly Practice Exams, Even Though They Only Receive Points for Completion and Are Not Graded by the Instructor?

While we were concerned that students might not complete the practice exams if the practice exams were not graded by the instructor and points were not given for correctness, the majority of students in fact did complete all the practice exams. However, while the majority of students completed all practice exams, performance on practice exams was significantly lower than on course exams. This relative performance is consistent with findings of Freeman and Parks (2010), who also compared peer versus expert grading of practice exams. In their study, students received course points for correctness; therefore, we think that the low practice exam performance we observed was not solely dependent on student effort based on how points were awarded. We purposefully made practice exams low-stakes participation-based assignments to encourage students to use them as formative assessments. Therefore, we think more likely explanations for the low performance on practice

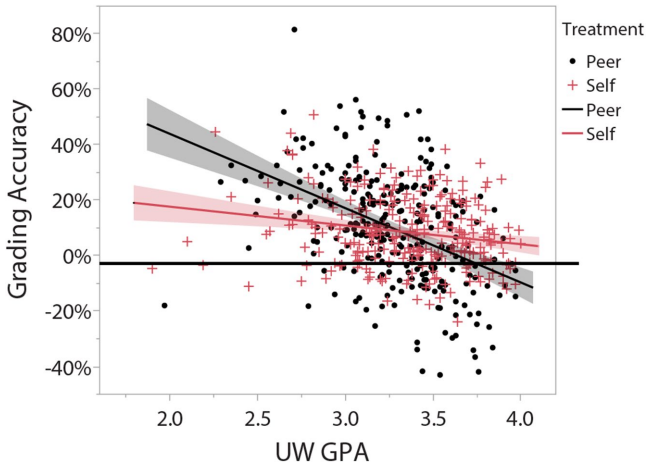

FIGURE 3. Interaction between incoming GPA and grading treatment group predicts accuracy of grading practice exams. See Eq. 1 in *Methods* for grading accuracy calculation. Line at zero designates complete accuracy compared with expert. Lines are best fit of the raw data. Most students fall above the zero line, indicating overestimation of practice exam grades. Students below the zero line are harsher graders than experts. There is less discrepancy in grading accuracy across GPA in the self-condition vs. the peer condition.

exams are that students just did not study for practice exams or that they used practice exams for a knowledge check to determine gaps in their knowledge before doing serious studying for the exam.

Question 2: Does Peer or Self-Grading Have a Differential Impact on Student Performance on Practice Exam Questions or on Course Exams?

Whether students peer or self-graded did not impact their practice exam grades as determined by experts. Interestingly, students who knew they would be grading their own work and could have taken the opportunity to put less effort into the practice exam did not, as the expert grades for their practice exams were no lower than those of peer graders. Given this, it is likely that both peer and self-graders put equal effort into answering the questions.

We found that students with higher incoming GPAs in either grading group scored higher on both practice exams and course exams. The ability to perform well on exams may reflect both students' understanding of the course material and their ability to plan for and complete college-level assignments. This ability to plan is a form of self-regulated learning that contributes positively to student learning (Pintrich and De Groot, 1990; Zimmerman and Schunk, 2001).

In this introductory biology course, there is a gender bias on course exam performance. This gender bias was surprising, given that females are the majority of students in the class and the instructor is female—all factors that argue against a gender bias. Nevertheless, females score about 1% lower, on average, than males on course exams, which is consistent with previous studies in biology courses (Wright *et al.*, 2016). While 1% seems small, it can impact a student's course grade by 0.1 GPA points (UW grades on a 4.0 scale in 0.1 increments). However, we did not observe a gender bias in practice exam performance. Many theories have been proposed for females' lower performance than males on course exams, including stereotype threat (Spencer *et al.*, 1999) and test anxiety (Cassady and Johnson, 2002; Ramirez and Beilock, 2011). It is possible that practice exams pose less of a threat or invoke less anxiety than the actual exam and therefore allow females to realize their academic potential.

Question 3: Does Peer or Self-Grading Have a Differential Impact on Student Grading Accuracy?

Our finding, that most students are easier graders than experts, is consistent with others who have used peer and self-grading in their courses (Freeman and Parks, 2010; De Grez *et al.*, 2012). Although most students were more lenient graders, we did find that there was heterogeneity within the student population. Students with lower incoming GPAs tended to assign more points to a response than an expert, while students with higher GPAs tended to assign a grade closer to that of the expert. This main effect of lower-GPA students overestimating performance in both treatments may be explained by the Dunning-Kruger effect, which addresses cognitive bias and illusory superiority (Kruger and Dunning, 1999).

While lower-GPA students were more lenient graders in both treatments, the impact of GPA on grading accuracy was exacerbated in the peer-grading treatment. This indicates that lower-GPA students can grade more accurately when they grade themselves, but when they grade peers, they are not as accurate. Therefore, these results cannot be explained by a lower grading ability or just the Dunning-Kruger effect. Instead, lower-GPA students might award higher grades to peers because they have a low biology academic self-concept (a low perception of their ability in biology; Cooper *et al.*, 2018) and hence doubt their ability to be critical of other students. These students may also have more of a challenge in interpreting the answers of peers that deviate from or paraphrase the rubric and, therefore, may give their peers the benefit of the doubt. As lower-GPA students in the self-grading group were familiar with their own writing and thought processes, they therefore were able to more critically grade their own answers.

Grading accuracy improved as GPA increased in both treatments; however, as GPA increased, peer graders became more accurate than self-graders. Furthermore, at the highest GPAs, peer graders became harsher than experts. It may be that higher-GPA students who are self-grading are not as willing to acknowledge when their answer did not meet the criteria of the grading rubric, while peer graders with higher GPAs had both the ability, confidence, and willingness to identify the flaws in the answers of others.

Even when controlling for GPA, EOP students, compared with non-EOP students, were more lenient graders in the peer-grading treatment. Again, this indicates that EOP students can grade more accurately when they are grading themselves; hence, they do not have a lower grading ability. However, when they are grading a peer, they do not grade as accurately. A possible explanation for why EOP students are less accurate when grading a peer rests in the value that EOP students place on maintaining social networks and community (Stephens *et al.*, 2012; Eddy and Hogan, 2014). EOP students may feel that grading a colleague harshly could jeopardize those important ties to community. Additionally, EOP students might have a low biology academic self-concept (a low perception of their ability in biology; Cooper *et al.*, 2018) and hence doubt their ability to be critical of other students.

Limitations and Future Directions

In this study, we investigated whether peer or self-grading of practice exams had a differential impact on student course performance. Whether students peer or self-graded also did not impact course exam performance. Both grading methods appear to similarly impact student learning. However, we do not know the impact, if any, the practice exam assignment had on course performance due to the lack of the relevant comparison groups. We did not have a no–practice exam control, as Freeman *et al.* (2011) had shown that adding practice exams to a highly structured course improved student course performance. Therefore, we felt that having a no–practice exam control would be unethical.

As the practice exam assignment has multiple components, each of which has the potential to improve course performance (answering exam-like questions, grading an answer using an instructor-generated rubric, and receiving feedback on your performance), it would be necessary to design experiments to test each component. For example, it could be that receiving feedback on your answer is not as important as grading an answer using an instructor-generated rubric, which could help students learn what type of answers earn full credit.

To investigate the possible mechanisms underlying the impact of the practice exam assignment on course exam

performance, we suggest that future research explore the individual components of the practice exam assignment. We would propose a series of experiments testing the following two variables and levels that could be combined in a factorial design. Variable 1: grading an answer using an instructor-generated rubric (levels: grading your own answers, grading a peer's answers, grading an example answer, not grading). Variable 2: receiving feedback on your performance that you are required to read or reflect upon (levels: self-feedback, peer feedback, instructor/TA feedback, no feedback).

In this study, we did not train students to grade or give them feedback on their grading ability. Training could improve grading accuracy of both groups and could bring peer graders to be more inline with self-graders. This improvement in grading might possibly change the impact of both peer and self-grading on exam performance. Therefore, we propose that future experiments on practice exam grading incorporate a training component for all students. Instructors may use tools such as Calibrated Peer Review, a Web-based software that trains students how to evaluate written responses based on a series of exemplary responses generated by the instructor. Previous findings suggest that problem-based writing with peer review improves performance in a physiology course (Pelaez, 2002).

We were especially intrigued to find the GPA/EOP by grading treatment interactions that showed EOP students and lower-GPA students were more lenient graders in the peer treatment. To investigate the hypotheses that academic self-concept, sense of community, or self-confidence can explain this pattern, we suggest surveys of these constructs could be developed and administered during future exploration of the optimal use of practice exams.

CONCLUSIONS

Based on our findings, there is no clear benefit of peer grading over self-grading or vice versa. Thus, either can solve the major challenge instructors face in giving the personal feedback that is required to make practice exams an effective form of deliberate practice for students in a large course. If instructors choose to use peer grading in their classes, we suggest they will have to provide more support and feedback on how to grade accurately to students with lower GPAs or educationally or economically disadvantaged students (EOP in our population). Owing to the more complex timing and software required to implement peer grading, the additional instruction needed to assist all peer graders to be more accurate graders, and the fact that self-grading is more accurate for a wider range of student academic abilities, we have decided to use only self-grading in our classes.

ACKNOWLEDGMENTS

We thank all the students who participated in the study and Kyle Loucks for logistical assistance in performing the experiment. We also thank Sarah Farrell and Osman Salahuddin and the University of Washington BERG lab for helpful discussion.

REFERENCES

Adachi, C., Tai, J. H.-M., & Dawson, P. (2017). Academics' perceptions of the benefits and challenges of self and peer assessment in higher education. *Assessment & Evaluation in Higher Education*, *43*(2), 294–306. https://doi.org/10.1080/02602938.2017.1339775

Balch, W. R. (1998). Practice versus review exams and final exam performance. *Teaching of Psychology*, *25*(3), 181–185. https://doi.org/10.1207/s15328023top2503_3

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer Science.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27*(2), 270–295. https://doi.org/10.1006/ceps.2001.1094

Cheng, K. K., Thacker, B. A., Cardenas, R. L., & Crouch, C. (2004). Using an online homework system enhances students' learning of physics concepts in an introductory physics course. *American Journal of Physics*, *72*(11), 1447–1453. https://doi.org/10.1119/1.1768555

Colvin, G. (2008). *Talent is overrated: What really separated world-class performers from everybody else*. New York: Penguin.

Cooper, K. M., Krieg, A., & Brownell, S. E. (2018). Who perceives they are smarter? Exploring the influence of student characteristics on student academic self-concept in physiology. *Advances in Physiology Education*, *42*(2), 200–208. https://doi.org/10.1152/advan.00085.2017

De Grez, L., Valcke, M., & Roozen, I. (2012). How effective are self- and peer assessment of oral presentation skills compared with teachers' assessments? *Active Learning in Higher Education*, *13*(2), 129–142. https://doi.org/10.1177/1469787412441284

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education*, *24*(3), 331–350. https://doi.org/10.1080/03075079912331379935

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Eddy, S. L., & Hogan, K. A. (2014). Getting under the hood: How and for whom does increasing course structure work? *CBE—Life Sciences Education*, *13*(3), 453–468. https://doi.org/10.1187/cbe.14-03-0050

Ericsson, K., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist*, *49*(8), 725.

Ericsson, K., Krampe, R., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363.

Fakcharoenphol, W., & Stelzer, T. (2014). Physics exam preparation: A comparison of three methods. *Physical Review Special Topics—Physics Education Research*, *10*(1), 010108. https://doi.org/10.1103/PhysRevSTPER.10.010108

Freeman, S., Haak, D., & Wenderoth, M. P. (2011). Increased course structure improves performance in introductory biology. *CBE—Life Sciences Education*, *10*(2), 175–186. https://doi.org/10.1187/cbe.10-08-0105

Freeman, S., O'Connor, E., Parks, J. W., Cunningham, M., Hurley, D., Haak, D., … Wenderoth, M. P. (2007). Prescribed active learning increases performance in introductory biology. *CBE—Life Sciences Education*, *6*(2), 132–139.

Freeman, S., & Parks, J. W. (2010). How accurate is peer grading? *CBE—Life Sciences Education*, *9*(4), 482–488. https://doi.org/10.1187/cbe.10-03-0017

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Haak, D. C., HilleRisLambers, J., Pitre, E., & Freeman, S. (2011). Increased structure and active learning reduce the achievement gap in introductory biology. *Science*, *332*(6034), 1213–1216. https://doi.org/10.1126/science.1204820

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Liu, N., & Carless, D. (2006). Peer feedback: The learning element. *Teaching in Higher Education*, *11*(3), 279–290.

Mahlberg, J. (2015). Formative self-assessment college classes improves self-regulation and retention in first/second year community college students. *Community College Journal of Research and Practice*, *39*(8), 772–783. https://doi.org/10.1080/10668926.2014.922134

Panadero, E., Romero, M., & Strijbos, J.-W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in*

Educational Evaluation, *39*(4), 195–203. https://doi.org/10.1016/j.stueduc.2013.10.005

Pelaez, N. J. (2002). Problem-based writing with peer review improves academic performance in physiology. *Advances in Physiology Education*, *26*(3), 174–184. https://doi.org/10.1152/advan.00041.2001

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1). 33.

Ramirez, G., & Beilock, S. L. (2011). Writing about testing worries boosts exam performance in the classroom. *Science*, *331*(6014), 211–213. https://doi.org/10.1126/science.1199427

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved October 12, 2017, from www.R-project.org/

Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment*, *11*(1), 1–31. https://doi.org/10.1207/s15326977ea1101_1

Sambell, K., McDowell, L., & Brown, S. (1997). "But is it fair?": An exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation*, *23*(4), 349–371. https://doi.org/10.1016/S0191-491X(97)86215-3

SAS Institute. (2016). *JMP Pro* (Version 13.0.0). Cary, NC.

Schinske, J., & Tanner, K. (2014). Teaching more by grading less (or differently). *CBE—Life Sciences Education*, *13*(2), 159–166. https://doi.org/10.1187/cbe.CBE-14-03-0054

Simkin, M. (2015). Should you allow your students to grade their own homework? *Journal of Information Systems Education*, *26*(2), 147–153.

Simkin, M., & Stiver, D. (2016). Self-graded homework: Some empirical tests of efficacy. *Journal of Education for Business*, *91*(1), 52–58. https://doi.org/10.1080/08832323.2015.1110554

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28. https://doi.org/10.1006/jesp.1998.1373

Stefani, L. A. J. (1992). Comparison of collaborative self, peer and tutor assessment in a biochemistry practical. *Biochemical Education*, *20*(3), 148–151. https://doi.org/10.1016/0307-4412(92)90057-S

Stephens, N. M., Fryberg, S. A., Markus, H. R., Johnson, C. S., & Covarrubias, R. (2012). Unseen disadvantage: How American universities' focus on independence undermines the academic performance of first-generation college students. *Journal of Personality and Social Psychology*, *102*(6), 1178–1197. https://doi.org/10.1037/a0027143

Sun, D. L., Harris, N., Walther, G., & Baiocchi, M. (2015). Peer assessment enhances student learning: The results of a matched randomized crossover experiment in a college statistics class. *PLoS ONE*, *10*(12), e0143177. https://doi.org/10.1371/journal.pone.0143177

Tanner, K. D. (2012). Promoting student metacognition. *CBE—Life Sciences Education*, *11*(2), 113–120.

Topping, K. J. (2009). Peer assessment. *Theory Into Practice*, *48*(1), 20–27. https://doi.org/10.1080/00405840802577569

Trussell, H. J., & Dietz, E. J. (2013). A study of the effect of graded homework in a preparatory math course for electrical engineers. *Journal of Engineering Education*, *92*(2), 141–146. https://doi.org/10.1002/j.2168-9830.2003.tb00752.x

Wright, C. D., Eddy, S. L., Wenderoth, M. P., Abshire, E., Blankenbiller, M., & Brownell, S. E. (2016). Cognitive difficulty and format of exams predicts gender and socioeconomic gaps in exam performance of students in introductory biology courses. *CBE—Life Sciences Education*, *15*(2), ar23. https://doi.org/10.1187/cbe.15-12-0246

Zimmerman, B. J., & Schunk, D. H. (2001). *Self-regulated learning and academic achievement: Theoretical perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.