Examinations That Support Collaborative Learning: The Students' Perspective

By Georg W. Rieger and Cynthia E. Heiner

We used surveys and classroom observations to examine student reactions to two-stage exams, where students first do the exam individually and then redo it collaboratively. Our results show why both students and instructors appreciate this examination format: Two-stage collaborative examinations are relatively easy to implement, have a high potential for learning, and support the collaborative learning approach used in many sciences classes. A look at survey data from an introductory physics class shows that a vast majority of students (76%) had a positive opinion of this exam format (expressed in 236 comments) whereas only 10% expressed an overall negative opinion in 30 negative statements. Most of the positive comments relate to how this benefits learning. In this article, we describe how to implement two-stage exams, discuss advantages and disadvantages, and present the students' view.

University instructors increasingly use interactive engagement and social/collaborative learning methods in their science classes to achieve better learning outcomes (National Research Council, 2012). Such methods result in deeper engagement by the students and incorporate more formative assessment to support learning. A number of research-based methods, such as peer instruction (Mazur, 1997), think-pair-share (Johnson, Johnson, & Smith, 2011), and cooperative group problem solving (Heller & Hollabaugh, 1992; Heller, Keith, & Anderson, 1992), share some basic features that are recognized to support learning across a wide range of contexts. These features include intense engagement by students, collaborative learning where students develop their thinking, and immediate feedback through the interactions with their peers (National Research Council, 2012). In this article we discuss an exam format—two-stage exams—that uses these same features.

Frequently, collaborative learning and formative assessment will be used in classroom instruction, but the course exams will remain in the traditional format in which students solve problems in isolation and only receive feedback several days later. Exams send very powerful messages, and such an exam format does not support the message that collaborative learning is important. Moreover, although individual exams produce a uniquely intense engagement with the material, that engagement provides little or no contribution to learning—defined as acquiring new ideas—because of the lack of timely and useful feedback (Black & Wiliam, 1998).

The two-stage exam is a relatively simple way to solve these problems. In a two-stage exam, students first complete and turn in the exam individually and then, working in small groups, answer the exam questions again. This makes the exam itself a valuable learning experience while also sending a consistent message to the students as to the worth of collaborative learning. We see indications that the use of this exam format goes beyond ensuring consistency across course components, in that it positively impacts how students approach the other collaborative components in the course. The two-stage exam accomplishes this while still providing summative assessment of individual performance.

Collaborative tests have been used for some time in a variety of formats (see summaries in Leight, Saunders, Calkins, & Withers, 2012; Zipp, 2007). The two-stage format discussed in this article (sometimes referred to as *group test* Cortright, Collins, Rodenbaugh, & DiCarlo, 2003; or *pyramid exam*, Cohen & Henle, 1995) has also been used in the past, in particular in team-based learning as part of the readiness assurance process (see

e.g., http//:www.teambasedlearning.org). This process, which uses scratch-and-win type testing cards during the group part to reveal the answers to all questions, follows up on the assigned reading, and provides a low-stakes way to ensure that students have the background knowledge necessary for the problem-solving activities that follow. However, administering high-stakes examinations such as midterm or final examinations in a two-stage format is still relatively rare. Stearns (1996), for example, mentions increased student performance on the (individual) final exam in a research method and statistics class after taking the midterm exams in a two-stage format, as well as decreased dropout rates, higher enjoyment of the course, and increased collaborative skills. Only a few studies have attempted to measure the benefits of two-stage exams on learning in science: In a recent study, Gilley and Clarkston (2014) reported knowledge gains (increases in student learning, i.e., the original acquisition of knowledge by students) due to the collaborative part of the exam in a science course on natural disasters, whereas other studies in biology (Leight et al., 2012) and physiology (Cortright et al., 2003) have focused on the retention of content. A positive impact on student motivation, reduced test anxiety, increased collaborative skills, and improved perception of the course were also mentioned in a number of other studies (see references in Gilley & Clarkston, 2014; Leight et al., 2012; Zipp, 2007). Potential limitations of two-stage exams are a reduced number of questions on the tests (to make time for the group portion) and a slightly higher administrative effort. In addition, differences in group composition may limit the effectiveness of this approach in groups with one dominant student or in groups with free-riders (see discussion in Zipp, 2007). Our survey results, however, indicate that this occurs only in a small number of groups.

This article was inspired by seeing both the success of two-stage exams and how popular they have been with both students and instructors across the Faculty of Science at the University of British Columbia (UBC). This exam format was first introduced in the UBC Faculty of Science 3 years ago and is now being used in at least 20 science courses. The faculty members value the widespread intense engagement by their students during the second stage of the exam, and as discussed below, students see them as valuable learning experiences. Next, we describe how to implement two-stage exams, discuss their benefits, and present the students' view.

Implementation of two-stage exams

The particular format of the two-stage exam we use is relatively easy to implement and has worked well in UBC science courses.

- Stage 1 (individual, between 3/4 and 2/3 of the examination time): This is a standard formal examination students complete working alone.
- Stage 2 (small groups, remainder of the examination time): The group portion begins after all individual exams are collected. Students work in groups of three or four students on (mostly) the same problems as in the individual portion (Figure 1). They must come to a consensus on the answers and hand in one copy with the names and student ID numbers of all group members. Because students have already seen each problem during Stage 1, solving the same problems again in Stage 2 usually takes much less time than in Stage 1, including the time for discussions and agreeing on a solution.

As an example, the two-stage exam given in our introductory physics course ($N = 178$) had a total duration of 90 minutes that was split as 55 minutes for individual effort (Stage 1) and 30 minutes for group effort (Stage 2), with 5 minutes in between for making the switch from Stage 1 to Stage 2. During the switch, instructors and teaching assistants first collected the individual exam copies, and then students were instructed to sit with their predetermined group members (3–4 students per group). In some courses, these groups are preformed (e.g., same as collaborative groups in class or groups put together by the instructor), whereas in other courses, students are free to choose their groups. Once the groups were assembled, the second part of the exam was distributed. Generally the switch can be done in less than 5 minutes—even in large classes, if there is at least one instructor or teaching assistant for 50 students.

A two-stage exam in a 50-minute lecture time slot is doable, but having a 90-minute time slot is easier. In some courses, instructors have replaced their 50-minute in-class midterm exams with 90-minute evening exams, so that similar content can be covered. Concerns about the length of an exam can be addressed by repeating only the conceptual questions of the individual part in the group portion and/or by turning short-answer questions of the individual part into multiple choice or ranking tasks in the group portion; see Figure 1 for examples.

Grades from the individual and the group portion are combined for the total examination mark, weighted between 75% to 90% for the individual portion and 25% to 10%, respectively, for the group portion. The group exam score has no effect on the differentiation between stu-

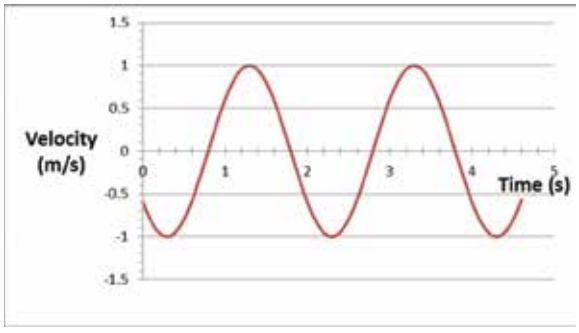
dents (i.e., a student's performance relative to the class), yet even the small weight of the group portion provides sufficient motivation for students to take this part seriously. For example, an 85/15 (individual/group) split used in our physics class resulted in an average increase of the midterm mark due to the group portion of 3.3% and an average increase in the final exam score due to the group portion of 1.6 %. The resulting impact on the average course grade of the group part of the exams was 0.5% from the midterm and 0.7% for the final exam, where the standard deviation of course grade distribution was 9.7%.

On the basis of the collective experience at UBC across the science disciplines of physics, chemistry,

FIGURE 1

Examples of questions taken from a two-stage exam for physics.

Most questions will be the same for the individual and the group part. If questions are modified, it is usually to reduce the number of detailed calculations, which do not promote discussions, and replace with prompts to "explain your reasoning." Additionally, one or two more challenging questions may be added.

INDIVIDUAL PART	GROUP PART
A train is approaching the train station at velocity v_0 = 15 m/s relative to the ground in still air. The train operator sounds the train whistle, which emits a note with frequency f_0 = 2500 Hz. The sound of the whistle is heard by different observers: The train operator hears a frequency f_A; a person standing on the station platform watching the train approach hears a frequency f_B; the operator of a second train approaching the station from the other direction with velocity v_2 = 10 m/s hears a frequency f_C. What are the frequencies f_A, f_B, and f_C?	(Changed to ranking) A train is approaching the train station at velocity v_0 relative to the ground in still air. The train operator sounds the train whistle, which emits a note with frequency f_0. The sound of the whistle is heard by different observers: The train operator hears a frequency f_A; a person standing on the station platform watching the train approach hears a frequency f_B; the operator of a second train approaching the station from the other direction with velocity v_2 hears a frequency f_C; a passenger traveling on a slower train that has just been overtaken by the first train (and sees the first train move farther away) hears frequency f_D. Rank the frequencies heard by the observers (f_A, f_B, f_C, f_D) in order from the highest to the lowest frequency.
The graph shows the velocity vs. time graph of a harmonic oscillator.  **Determine** **a)** the angular frequency **b)** the maximum displacement **c)** the phase constant and the equation describing the position as a function of time.	(Replace part c) **a)** same **b)** same **c)** Sketch the potential energy curve as a function of time. Assume that we have a **horizontal** harmonic oscillator.

biology, math, statistics, Earth and ocean sciences, computer science, and land and food systems, we would further recommend the following:

- Students are told on the first day of classes that examinations will be conducted in this format and, more important, why this is done in this way.
- A policy is implemented that the group score cannot be lower than the individual mark. This will address concerns about fairness. In practice, it affects only a few high-performing students as groups perform equal or better than individual students in almost all cases.
- Clear instructions are given during the individual-to-group transition. For example, students should remain seated while their individual exam copies are collected. Remind and check that all names and student numbers are listed on the group exam.
- Students are discouraged from working on their own during the group portion and all members are encouraged to be involved in discussing every problem. Teaching assistants and instructors can help with forming groups and encouraging collaborative work, but this is seldom needed.

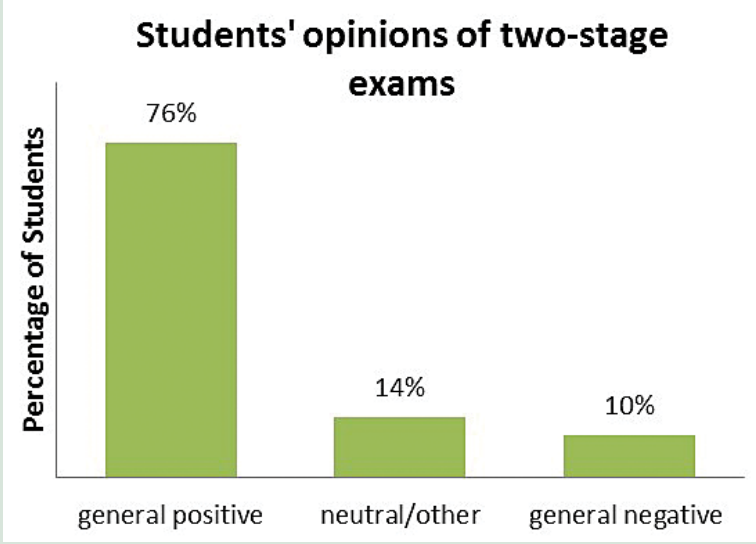
Overall, it does not take much effort to run a two-stage exam. From our experience, creating the group portion of the exam is easy because it is largely identical to the individual exam, and the additional marking time of the group copies is minor because most solutions are correct. To our knowledge, no instructor at our institution who has tried two-stage exams has abandoned this approach.

Benefits of two-stage exams

Here we offer some thoughts on why collaborative exams can increase learning and retention and add a few observations from several science classes at UBC.

FIGURE 2

Percentage of students with generally positive, negative, or neutral opinions about two-stage exams (*N* = 123 students). General positive means students found the exam format to be good or helpful for learning. *Neutral/other* means that students did not express a clear positive or negative opinion, or commented on other things. *General negative* means that student had overall negative comments about the exam format.



During the high-stakes environment of an examination, students are heavily invested in figuring out the correct answers. After the individual portion, all students are well prepared to discuss their approach in a group. In these discussions, students get immediate feedback on their solutions from their peers, which might help them clarify their thinking (Cortright et al., 2003, Gilley & Clarkston, 2014, Rao, Collins, & DiCarlo, 2002). Weaker students could benefit from the explanation that is targeted to their difficulties, higher achieving students might benefit from explaining concepts to others, and everyone may well benefit from critically evaluating others' ideas.

One may argue that these same benefits are also present in "normal" in-class collaborative learning activities; so why do this on an exam? Even a casual observation of the two situations reveals the difference: We routinely see nearly 100% engagement during the group part of the exams, presumably because of the high-stakes situation of an examination. As confirmed through both observations and student self-reports, most groups discuss the questions until all members agree on an answer; even during open-book, two-stage exams, it was very unusual to see students looking through the book to find the answers instead of discussing them and figuring them out themselves. Those students who are usually too shy to speak up during in-class activities will defend their answers vigorously during the second stage of the exam. By comparison, the discussions during normal in-class activities, such as clicker question discussions, do not have nearly the same intensity. This is probably because the stakes are

lower, and it is not necessary for students to reach an agreement because they usually submit their own (individual) answers. The students also know that they will receive expert feedback from the instructor following the discussion, so they don't have to evaluate as carefully what their colleagues are saying. Finally, students are better prepared to carry out peer discussions in a two-stage exam than they are during lecture because (a) they have studied for the exam, and (b) each student is forced to think deeply about the questions during the individual portion of the exam before the discussion starts in the group portion.

In our introductory physics class, we noted an additional beneficial effect of a two-stage midterm exam: It increased the engagement during in-class collaborative activities following the examination. Although students regularly participated in peer discussions of clicker questions and worksheet problems before the midterm and the instructors explained the benefits of collaborative learning, it appears that the two-stage exam convinced the students (more) of the value of peer discussions. It is also possible that, after the midterm exam, students think of the in-class activities as more directly related to exams.

Impact on student opinions

For illustrative purposes, we examine in detail how two-stage exams impacted student opinions in one course; however, these results are similar to what has been seen in other science courses.

We gave both the midterm and final exams in a two-stage format in the aforementioned calculus-based introductory physics course. The students filled out a 20-question online survey after the final exam; four questions probed their views on the exams. Of the 179 students, 123 completed the survey. Eighty-seven percent supported the use of the two-stage exam format for midterms,

TABLE 1

Coding scheme and results as applied to students' written comments regarding their experience with two-stage exams in Physics 101.

Overall code	Detailed code	Description of code	No. of times mentioned (*N* = 123 students)
General positive (Total: 236)	G-E	Good, enjoy, benefit, great, liked, useful, OK, interesting	56
	H	Helpful	30
	C	Increased confidence	9
	LE	Good learning experience, good way to review exam	21
	LE-D	Learning from: discussions with others, hearing other approaches, comparing with others, explaining yourself, collaborating	48
	IF	Immediate feedback: good to know if right or wrong	34
	IF-LM	Immediate feedback: learning from mistakes	16
	GD-pos	Positive mention of group working together, group members, meeting friends, group preparation, cooperation, and references to grade boost	22
Neutral/ other	Misc	Random comments not fitting into the above categories as well as suggestions	15
General negative	NEG-gen	Negative mention of group not working so well together, not everyone pulling their own weight, hard to explain to others, and concerns about unfair grade boost to weaker student, not fair for the individual	15
	NEG-em	Dislike, frustrating, not helpful, feeling sad or depressed, less confident	15

whereas 74% supported the use for both midterm and final exams. A possible reason for the difference could be that students view the midterm as being part of learning and perhaps feedback on their studying, whereas they see the final exam as a kind of "certification," similar to many instructors. Many students see this course as their final exposure to physics, so although students may see the second-stage feedback on what they did wrong for the midterms as productive, they may not appreciate it as much for the final, where there is no hope of using the feedback for future improvement. To explore this further, one would need to conduct interviews with students.

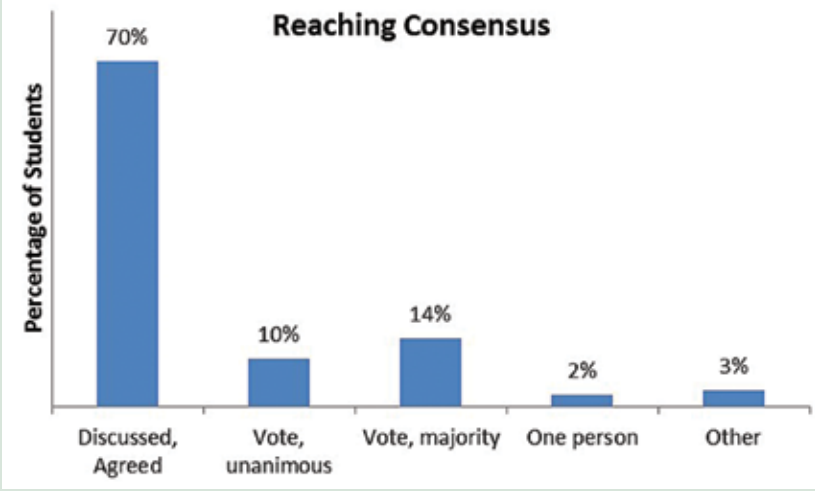
The survey included a question in which students were asked to describe their experience with the group exam in one or two sentences. All students who completed the survey answered this question. As shown in Figure 2, most students had a generally positive opinion.

The detailed analysis and coding scheme we developed for classifying the comments is shown in Table 1. Many students' responses fell into multiple categories; from the 123 students, we coded 283 comments. The comments were coded independently by each researcher and then compared. The interrater reliability for the comments was 95%, with differences in the coding being discussed until an agreement was reached. A few examples of student comments and coding are as follows:

- Student A: "It was a good experience since going over the exam with my peers reassured me about my answers. As well, I was able to learn from my mistakes through the group exam." (G-E; C; IF-LM)
- Student B: "It was surprisingly very helpful. I would say I contributed as much as I could. When I got a different answer I always commented why I chose the answer that I did and our group would discuss it. I think I was also very lucky to meet kind people during lecture." (H; MISC; LE-D; GD-pos)
- Student C: "It was sort of depressing to know what you got wrong right after writing the exam. I think it ends up being worth it, though, because you learn from your mistakes and the way classmates explain things could be easier to understand at times than the way it's explained in the textbook." (NEG-em G-E; IF-LM; LE-D)
- Student D: "The group exam was interesting and a good opportunity to go over the answers and talk about the questions. Just did not like when some members did not do anything." (G-E, LE, NEG-gen)

FIGURE 3

Student survey results on group decision making (*N* = 123). Students were asked: "During the group exam, my group usually____." Full answer choices (left to right): "discussed EACH question until ALL members agreed on an answer and explanation," "took a VOTE and if unanimous moved on, otherwise discussed the question until all members agreed on an answer," "took a VOTE and used the MAJORITY to determine the answer," "USED the answers from the ONE PERSON in the group who knew the most physics."



These comments give us insight into why students generally value two-stage exams: they felt it was a good learning experience, and a good way of reviewing the exam, they learned from discussing with other students and hearing other students' approaches, they enjoyed working together, and they valued the immediate feedback in the group part.

Students also expressed concerns about the exam format in 30 negative comments, half relating to group work and the other half to the emotional impact of getting immediate feedback. However, of the 15 students who criticized the group work, only six students rated their overall experience on the two-stage exam as negative. Three of the six students mentioned concerns about "weak students unfairly gaining marks." Nine students were critical of the group work but

still had an overall positive experience (as Student C). Fifteen students commented that it was "sad" or "depressing" to learn about their mistakes, but nine students still had an overall positive view.

In the survey students were specifically asked about how their group reached consensus. The results are presented in Figure 3. Clearly, most students worked on the group exam in the intended collaborative way. Only three students, two of whom commented about bad dynamics in their group, claimed to have "used the answers from the one person in the group who knew the most physics." These responses support our observations of classwide participation in the second stage of the exam and the intensity of the physics discussions that ensue.

Summary

Two-stage exams are valuable instructional tools that offer a combination of formative learning and assessment. They can easily be implemented in many courses and are popular with students and faculty members who use them. Surveys show that this exam format is popular with students for the right reasons—students recognize the value of immediate feedback that takes place and the learning that results. The exam format is similar to the collaborative in-class activities and therefore strengthens the link between exams and the peer instruction activities in class. We have noted an increase in engagement during in-class peer activities after a group midterm exam. Further studies are necessary to establish that this is mainly a result of the two-stage exam. It would also be interesting to find out if students acquire better group skills through participation in the group part of the exam. The experience in our science faculty has shown that the two-stage approach contributes to the overall coherence of any course that is using techniques of collaborative learning and formative assessment, as well as allowing students to learn while completing the exam. We therefore highly recommend this exam format to any instructor looking to add a formative element to their summative assessments. ■

Acknowledgments

The authors gratefully acknowledge the Carl Wieman Science Education Initiative (CWSEI) for funding and support. We thank Carl Wieman for assistance with the preparation of the manuscript. We also thank all the CWSEI Science Teaching and Learning Fellows for providing information on their experiences with two-stage exams, in particular Brett Gilley and Bridgette Clarkston.

References

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5,* 7–74.

Cohen, D., & Henle, J. (1995, July). The pyramid exam. *UME Trends, 10,* 2, 15.

Cortright, R. N., Collins, H. L., Rodenbaugh D. W., & DiCarlo, S. T. (2003). Student retention of course content is improved by collaborative-group testing. *Advances in Physiology Education, 27,* 102–108.

Gilley, B. H., & Clarkston, B. (2014). Collaborative testing: Evidence of learning in a controlled in-class study of undergraduate students. *Journal of College Science Teaching, 43*(3), 83–91.

Heller, P., & Hollabaugh, M. (1992). Teaching problem-solving through cooperative grouping. Part 2: Designing problems and structuring groups. *American Journal of Physics, 60*, 637–644.

Heller, P., Keith, R., & Anderson, S. (1992). Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving. *American Journal of Physics, 60,* 627–636.

Johnson, D. W., Johnson, R. T., & Smith, K. A. (2011). Lecturing with informal cooperative learning groups. In J. Cooper & P. Robinson (Eds.), *Small group learning in higher education: Research and practice* (p. 46). Oklahoma City, OK: New Forums Press.

Leight, H., Saunders, C., Calkins, R., & Withers, M. (2012). Collaborative testing improves performance but not content retention in a large-enrollment introductory biology class. *CBE—Life Sciences Education, 11,* 392–401.

Mazur, E. (1997). *Peer instruction: A user's manual*. Upper Saddle River, NJ: Prentice Hall.

National Research Council. (2012). *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering.* Washington, DC: National Academies Press. Available at http://www.nap.edu/catalog.php?record_id=13362

Rao, S. P., Collins, H. L., & DiCarlo, S. E. (2002). Collaborative testing enhances student learning. *Advances in Physiology Education, 26,* 37–41.

Stearns, S. A. (1996). Collaborative exams as learning tools. *College Teaching, 44*(3), 111–112.

Zipp, J. F. (2007). The impact of two-stage cooperative tests. *Teaching Sociology, 35,* 62–76.

Georg W. Rieger (rieger@phas.ubc.ca) *is an instructor 1 in the Department of Physics and Astronomy and a member of the Carl Wieman Science Education Initiative (CWSEI), University of British Columbia (UBC), British Columbia, Vancouver, Canada.* ***Cynthia E. Heiner*** *was a teaching and learning fellow in the CWSEI at UBC at the time the article was written and is now at the Free University Berlin in Germany.*

Physics Exams that Promote Collaborative Learning

*Carl E. Wieman, Georg W. Rieger, and Cynthia E. Heiner,** University of British Columbia, Vancouver, BC, Canada

The two-stage exam is a relatively simple way to introduce collaborative learning and formative assessment into an exam. Their use is rapidly growing in the physics department at the University of British Columbia, as both students and faculty find them rewarding. In a two-stage exam students first complete and turn in the exam individually, and then, working in small groups, answer the exam questions again. During the second stage, the room is filled with spirited and effective debate with nearly every student participating. This provides students with immediate targeted feedback supplied by discussions with their peers. Furthermore, we see indications that the use of this exam format not only ensures consistency across interactive course components, but it also positively impacts how students approach the other collaborative course components. This is accomplished without losing the summative assessment of individual performance that is the expectation of exams for most instructors. In this paper we describe how to implement two-stage exams and provide arguments why they should be part of physics courses that use interactive engagement and social/collaborative learning methods.

Why two-stage exams?

Two-stage exams are not new. They have been discussed and used in multiple contexts,[1] but they are still relatively rare in physics courses[2] despite some of the clear advantages they offer. Exams are typically individual problem solving in isolation, in stark contrast to problem solving in the real world and in courses that stress collaborative learning activities. As cognitive psychologist Dan Schwartz puts it, "If you ask someone else for help on a problem in an exam, you are cheating, but if you don't ask someone for help on a problem in the real world, you are a fool." Individual exams miss an excellent opportunity for formative assessment that has been shown to be strongly linked to learning.[3] Students are more intensely engaged with the material during an exam than at any other time during the course. However this opportunity for formative assessment is lost, because the feedback on exams is typically very limited—mostly "right/wrong" and coming a substantial time after completion of the exam. Both of these factors reduce the value of feedback to learning. Also, as many instructors have observed, and we have confirmed by monitoring website use, most students only review midterm exam solutions when they are studying for the final exam. During the second stage of the two-stage exam, students receive immediate, targeted feedback on their solutions from their fellow students. Gilley and Clarkson have shown that essentially all members of the group take away from the exam nearly the mastery achieved by the group as a whole during the second stage, a level that is well above that shown by most individuals during the first stage.[4]

How to implement two-stage exams

The particular format of a two-stage exam that we use is relatively easy to implement and has worked well in numerous UBC physics courses. The second-stage "group portion" begins after all individual exams are collected. Students work in groups of three or four students on (mostly) the same problems as in the individual portion. They must come to a consensus on the answers and hand in one copy with the names and student ID numbers of all group members. Since the students have already carefully thought about each problem individually during stage 1, the discussions and agreeing on a solution during stage 2 usually takes less time. In our large introductory courses we allot 55 minutes for the individual effort (stage 1) and 30 minutes for the group effort (stage 2), with five minutes for making the switch from stage 1 to stage 2. Some instructors use two-stage exams in a one-hour timeslot, but it is more challenging. Although there usually is sufficient time to redo the entire exam, to save time when there are many long problems, we often repeat only the conceptual questions of the individual part in the group portion and/or turn short answer questions of the individual part into multiple choice or ranking tasks in the group portion. Box 1 shows two examples of questions that were modified for the group portion.

In determining the exam grades, we have used weightings of the individual to group portions of the exams of both 75/25% and 85/15%, and did not see any difference in the student behavior for the two cases. With either weighting, the impact of the group exam is typically a few percent on a student's total exam score, and less than one percent on his or her overall course grade. Students are told on the first day of classes how two-stage exams work and why examinations will be conducted in this format. They are also told about the stated policy that if the group score is lower than the individual exam grade, the group exam will not reduce their exam grade. In practice, this is relevant to only a few students because the groups nearly always perform as well or better than the best individual students. Overall grading time increases only slightly due to the group exam since a large fraction of the solutions are entirely correct, which makes grading easy and quick.

Students' reactions to two-stage exams

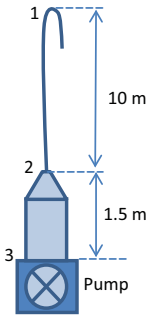
Witnessing the intense productive discussions in which nearly all students are engaged during the second stage has been the most convincing reason for most faculty for using

Box 1. Examples of questions taken from a two-stage exam for physics.

Most questions will be the same for the individual and the group part. If questions are modified, it is usually to reduce the number of detailed calculations, which do not promote discussions, and replace with prompts to "explain your reasoning." Additionally, one or two more challenging questions may be added.

Question

Assume you want to design a water fountain for your local park. The fountain is supposed to shoot water up to a height of 10.0 m above the exit nozzle, which is located 1.5 m above a pump that pumps water into a vertical tube of 5.0 cm diameter. The pump has a gauge pressure of 100 kPa.



Individual Part	Group Part
a) Rank the pressures at points 1 (at the top), 2 (at the exit of the nozzle), and 3 (at the exit of the pump). b) What is the diameter of the exit nozzle?	Part b changed to ranking: b) Rank the velocities at points 1, 2, and 3.

Question

You and your little sister are out in the snow on a sled that has a mass of 11 kg. Your sister, who weighs 29 kg, is sitting on the sled and you want to push her along. You start applying a horizontal force and initially the sled doesn't move but you slowly increase your force until, suddenly, the sled does move. You maintain the same force that you were applying when the sled started moving for the next 5.0 s after which you let go.

(Assume that the kinetic friction coefficient is $\mu_k = 0.02$ and the static friction coefficient is $\mu_s = 0.08$ in this case.)

Individual Part	Group Part
a) How far do you have to run if you apply the force for 5.0 s? b) What is your sister's speed at $t = 5.0$ s? c) After letting go, how far do your sister and her sled move until she is stationary again? (In case you could not solve part b, assume that her speed is $v = 2.5$ m/s at $t = 5.0$ s.)	(Converting calculation to reasoning and representation with graphs.) a) Draw a qualitative diagram that roughly shows the net force acting on the sled as a function of time. (Qualitative means that it explains the overall behavior without using exact numbers.) b) Draw a second qualitative graph of the acceleration of the sled as a function of time. c) Draw a third qualitative graph of the velocity of the sled as a function of time.

the two-stage format. Students also see the benefits of these discussions. We rarely have to discourage students from working individually during the group portion, and students that are usually too shy to speak up during in-class activities will defend their answers vigorously during the second stage of the exam. As confirmed through both observations and student self-reports,[5] a large fraction of the groups discuss the questions until all members agree on an answer, or they take a vote in cases where an agreement cannot be achieved. The high stakes context of an exam combined with the fact that all students are well prepared to participate in the discussion, because (a) they have studied for the exam and (b) they thought carefully about the questions and committed to an answer just moments ago during the individual portion, produce the perfect environment for rich discussion. Although we introduce collaborative learning activities into the course before the exams and explain the benefits, for many students the value becomes more readily apparent during the two-stage exam.

We see this on survey responses and in the behavior of the class after the first two-stage exam. Students' response to the use of two-stage exams is overwhelmingly positive, with 87% of the students recommending continued use of two-stage midterm exams and only a few percent recommending against their use. Examples of typical positive comments are:

> ***Student A:*** "I was able to instantly learn from my mistakes."
>
> ***Student B:*** "It was good to compare methods and answers with others, and it allowed us to be more confident."
>
> ***Student C:*** "Interesting. All had different ways [of] approaching the question. Very helpful to understand everyone's response and why they thought their answer was correct."

An interesting subset of the comments were those that indicated that the students found the experience emotionally unpleasant because they immediately recognized what they had done wrong, but for that same reason, clearly supported learning by the students.

> ***Student D:*** "The group exam was useful because I was able to see what I did wrong and what I did correct. The only negative part to it was [that] I realized all the mistakes I made."

Summary

Two-stage exams are an easy way to turn exams into learning experiences. This exam format is very popular with students because they recognize the value of the immediate feedback provided and the learning that results from it. The two-stage exams also provide a consistent message to students in any course that uses group work and collaborative learning.

References

* Current address: Free University Berlin, 14195, Berlin, Germany.

1. P. Heller, R. Keith, and S. Anderson, "Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving," *Am. J. Phys.* **60**, 627–636 (1992), and P. Heller and M. Hollabaugh, "Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups," *Am. J. Phys.* **60**, 637–644 (1992). See also introduction and references in Gilley and Clarkston (Ref. 4).
2. For example, two-stage exams are not mentioned among the 24 research-based instructional strategies in a large-scale survey that examines the knowledge and practices of physics faculty: C. Henderson and M. Dancy, "Impact of physics education research on the teaching of introductory quantitative physics in the United States," *Phys. Rev. ST Phys. Educ. Res.* **5**, 020107 (2009).
3. *How People Learn: Brain, Mind, Experience, and School: Expanded Edition* (National Academy Press, 2000).
4. B. Gilley and B. Clarkston, "Collaborative testing: Evidence of learning in a controlled in-class study of undergraduate students," *J. Coll. Sci. Teach.* (in press).
5. G. W. Rieger and C. E. Heiner, "Examinations that support collaborative learning: The students' perspective," *J. Coll. Sci. Teach.* (in press).

Department of Physics and Astronomy, University of British Columbia
Carl Wieman Science Education Initiative, University of British Columbia, Vancouver, BC, Canada; rieger@phas.ubc.ca

PHYSICS EDUCATION RESEARCH SECTION

The Physics Education Research Section (PERS) publishes articles describing important results from the field of physics education research. Manuscripts should be submitted using the web-based system that can be accessed via the American Journal of Physics home page, http://ajp.dickinson.edu, and will be forwarded to the PERS editor for consideration.

Collaborative exams: Cheating? Or learning?

Hyewon Jang[a)]
John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138

Nathaniel Lasry[b)]
Physics Department, John Abbott College, Montreal, Quebec H9X 3L9, Canada

Kelly Miller[c)] and Eric Mazur[d)]
John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138

(Received 11 February 2016; accepted 5 January 2017)

Virtually all human activity involves collaboration, and yet, collaboration during an examination is typically considered cheating. Collaborative assessments have not been widely adopted because of the perceived lack of individual accountability and the notion that collaboration during assessments simply causes propagation of correct answers. Hence, collaboration could help weaker students without providing much benefit to stronger students. In this paper, we examine student performance in open-ended, two-stage collaborative assessments comprised of an individually accountable round followed by an automatically scored, collaborative round. We show that collaboration entails more than just propagation of correct answers. We find greater rates of correct answers after collaboration for all students, including the strongest members of a team. We also find that half of teams that begin without a correct answer to propagate still obtain the correct answer in the collaborative round. Our findings, combined with the convenience of automatic feedback and grading of open-ended questions, provide a strong argument for adopting collaborative assessments as an integral part of education.
[http://dx.doi.org/10.1119/1.4974744]

I. INTRODUCTION

Exchanging information during conventional assessments, which focus on measuring individual students' knowledge, is typically viewed as cheating. Students are isolated from their peers and usually have no access to resources during examinations. However, virtually all human activities involve collaboration and experts use all available resources—material, digital, and human—when solving problems. If the role of education is to prepare students for expert practice, why shouldn't students be able to use those resources during an exam?

When students work collaboratively and exchange information, their performance on academic tasks improves.[1–3] Peer discussions can lead to higher-level reasoning and understanding.[3] Students working together construct new knowledge, develop skills, and obtain: greater understanding of concepts.[4] As collaborative student-centered activities gain traction, student-assessment has come into greater focus.[5,6] Arguably, one drawback of conventional assessments is that they do not provide an opportunity to learn.[7] Indeed, conventional assessments try to characterize a student's knowledge state, not make it change.[7,8]

A number of studies have examined collaborative testing approaches and documented positive effects such as improved performance,[1–3,9–13] increased motivation,[14,15] decreased test anxiety,[10] positive rapport with classmates,[16] increased retention,[17] and greater appreciation from both students and instructors.[5,12,17] Many forms of collaborative assessments have been designed to help students learn during assessments, including approaches such as two-stage exams,[12] cooperative testing,[10] and readiness assurance in team-based learning.[19] Collaborative assessments typically have two phases. The first phase is a conventional individual assessment; and the second is a collaborative assessment. To be effective, collaborative assessments should be complex enough to engage students in productive discussions.[2,17] To be efficient, the collaborative component can use the same questions as those posed in the individual round. Although many science instructors will acknowledge the shortcomings of conventional assessments, collaborative assessments are not widely used in higher education. Therefore, instead of highlighting its benefits, we identify four barriers to the adoption of collaborative assessments and address them systematically.

The first barrier to adoption is individual accountability. How does one assess an individual's performance with a team grade? Two-stage collaborative assessments address this issue by incorporating an individual round. This first round, typically accounting for half of the total credit, assesses individuals much like a conventional exam. The remainder of the credit comes from the collaborative round where groups of students work together iteratively toward a correct answer.

A second barrier to adoption is the notion that collaboration is but a means to propagate correct answers within teams. Stated differently, collaboration simply allows higher-ability students to share correct answers with their teammates. We test this hypothesis by asking the following three questions: (i) Does collaboration only help weaker students who do not typically get the correct answer and provide no advantage to higher-ability students? (ii) Does collaboration work in groups where none of the members had a correct answer in the individual round? (iii) What impact does the rate of individual correct answers have on the problem solving ability of a team? Our results put to rest the notion that collaboration can be reduced to correct-answer propagation. We show that, regardless of their ability level, students respond correctly more frequently after collaboration than when they respond individually, and that correct answers are even generated in groups where there is no correct answer to propagate.

The third barrier to adoption is the inherent difficulty of administering and managing two-stage exams, particularly with open-ended questions. After the initial individual round, students work in teams to solve each question they first answered individually. Groups typically submit a consensus answer. In our implementation, immediate feedback is given on each group answer so that students may iteratively work toward a correct solution. To minimize the complexity of administering collaborative exams, our group designed an online response system.[18] This system automatically grades open-ended and multiple-choice questions answered by individuals, manages team collaboration, and iteratively provides automatic feedback to groups. All instructors need to do is provide the questions and solutions.

The last barrier to adoption, and possibly the most difficult to address, is the resistance to changing established practices. We believe changes are warranted because: individuals can be held accountable for their learning; all students gain from collaboration, since they are more likely to answer correctly after discussion; and the complexity of managing collaborative assessments can be reduced to below that of managing conventional assessments using online systems.

II. METHODS

We studied an introductory calculus-based mechanics class taught at Harvard University by one of us (E.M.). The course uses a combination of instructional methods: team-based learning;[19] project-based learning comprised of three, month-long team projects;[20] and Peer Instruction.[21] The class meets twice a week for a total of 6 h of instruction per week.

Our study comprises 67 students (33 males and 34 females), mostly engineering and premedical students. We collected the following data for each student: gender, school year (freshman, sophomore, junior, and senior), Force Concept Inventory (FCI) scores,[22] class test results, and previous team makeup. We balanced teams of four or five students by considering gender, grades, academic experience (school year), and background knowledge (as assessed by FCI pre-test score), using the following two rules: (i) team members must have complementary strengths (as assessed by grades and FCI pre-test score), and (ii) avoid isolation of women.[19]

A. Assessment process

Students are assessed using a technique adapted from the "Readiness Assurance Process" in team-based learning.[19] All assessments take place in two rounds and are administered using an online team-based assessment system.[18] Students use their own laptops or tablets to enter responses to a set of multiple-choice or open-ended questions, which are either conceptual, computational questions, or estimation problems. During the first round, students individually answer the questions. After completing the individual round, without receiving feedback on their performance, students are asked to provide the answers to the same set of problems as a team. Students obtain scores for each round (individual and team), and their final score is computed as the mean of both scores. In the individual round, each correct answer is worth four points and no feedback is given about the correctness of their answer. In the team round, students solve the same problems collaboratively with a single team answer submitted by one of the team members. The system then provides immediate automatic feedback about the correctness of this answer. A correct answer receives four points. If the initial answer is incorrect, the team has two more chances to submit a correct answer. Correct answers on the second attempt receive 2 points, while correct answers on the third attempt receive 1 point. If the answer on the third attempt is still incorrect, the system reveals the correct solution. The process is shown schematically in Fig. 1.

During both rounds of the exams, teams are seated at round tables separated from each other by moveable white boards. During the individual round, the room is completely silent as students input individual answers into the system. In

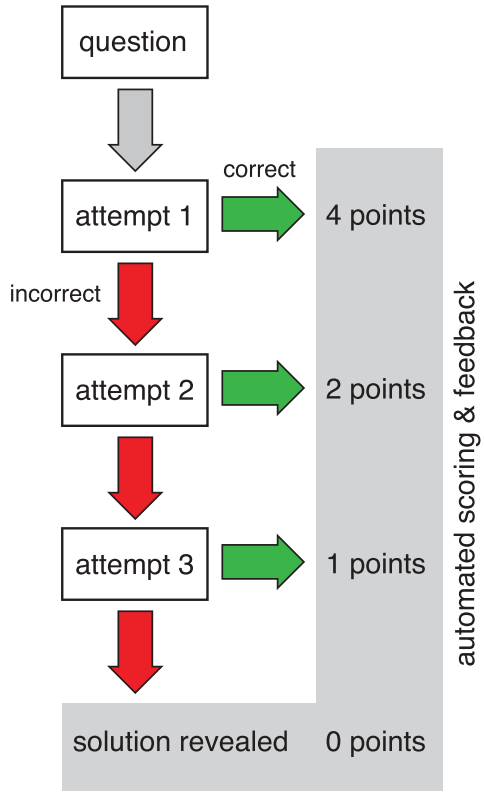


Fig. 1. Schematic representation of the team round in two-stage collaborative exams. Each time a team submits an answer, it receives instant feedback. If the answer is correct on the first attempt, the team receives a score of four points. If the answer is incorrect, the team has two more opportunities to answer. Correct answers in on the second attempt receive 2 points, while correct answers on the third attempt receive 1 point. After a third failed attempt, the solution is revealed.

the second round, students use the white board to discuss their answers to each question with their teammates. While there is no audio insulation between teams, once the collaborative round begins, teams appear sufficiently focused on their own discussions so that what is discussed by other teams does not appear to affect them. Also, the whiteboards provide an additional physical barrier to mitigate the audio interference between teams.

We administer five two-stage collaborative exams over the course of a semester. Each assessment takes approximately 90 min and has between seven and 11 questions. A total of 46 questions (31 open-ended and 15 multiple-choice questions) were posed across the five exams included in this study (sample questions are provided in the Appendix). To encourage collaboration, we designed relatively difficult questions for each assessment so that even the highest ability students do not score much above 50% in the individuate stage. Designing exams of high difficulty is important because it ensures that weaker students do not authoritatively follow higher-ability students (higher-ability students being wrong roughly half of the time) and encourages higher-ability students to be open to the thinking of their peers.

B. Assessment analysis

For the 46 assigned questions, we gathered 2954 individual responses and 644 team responses. We investigate the effectiveness of collaborative assessments by comparing the percentage of correct answers before and after the team collaboration. To account for the impact of students' ability level, we divided the students into three groups according to their average individual scores on the five assessments. Weak students are defined as those scoring in the lowest quartile, strong students are those scoring in the upper quartile, and average students as those between the lowest and highest quartiles.

III. RESULTS

A. The effectiveness of a single round of collaboration

Table I shows the mean score on individual assessments $\langle S_i \rangle$ and the mean score after the first team answer for each question $\langle S_{t1} \rangle$ on the five assessments given during the semester. Having purposely designed relatively difficult questions to encourage the collaboration of high-ability students, the mean scores for individual assessments range between 29% and 50% with a mean score across all students of 37%. Given that team answers are the result of a consensus, we use the non-parametric Wilcoxon signed-rank test (score distributions being non-normal) to compare the difference in scores for each individual before and after collaboration. We find that mean individual scores increase significantly ($p < 0.0001$) with increases ranging between 24% and 40% after a single round of collaboration (see Table I). The mean scores after the first team answer are roughly twice as large as the mean score on the individual round, a finding that quantitatively supports the notion that on average, groups perform better than individuals.[5,14,17,23]

Table I. Mean score for individual assessment, $\langle S_i \rangle$, and after a single round of team collaboration, $\langle S_{t1} \rangle$, for each of the five assessments administered. The numbers in parentheses represent the standard error of the mean for the reported values. (The differences in the rightmost column all have $p < 0.0001$ for the Wilcoxon Z-statistic.)

Assessment	N	$\langle S_i \rangle$	$\langle S_{t1} \rangle$	$\langle S_{t1} \rangle - \langle S_i \rangle$
1	67	0.50 (0.02)	0.90 (0.01)	0.40
2	67	0.31 (0.02)	0.60 (0.01)	0.29
3	64	0.33 (0.02)	0.57 (0.02)	0.24
4	61	0.29 (0.02)	0.68 (0.02)	0.39
5	63	0.37 (0.02)	0.75 (0.01)	0.38

Examining the 31 open-ended questions assigned, 38% of students responded correctly in the individual round. However, after a single round of collaboration, 74% of students answered correctly (see Fig. 2) showing a sizable and significant increase due to collaboration ($p < 0.0001$). This twofold increase in correct responses to open-ended questions cannot be ascribed to random guessing because open-ended questions provide no choices that students can pick from randomly. As this does not rule out teams converging on correct answers provided by individual team members, we investigate the propagation of correct answers within teams in Sec. III B.

B. Collaboration and propagation of correct answers

To disprove the idea that collaboration simply causes correct answers to propagate within teams, we first examine the effectiveness of collaboration in teams where no students answered correctly to open-ended questions in the individual round. That is, can collaboration work in teams where there is no correct answer to propagate? In our study, 14 teams responded to 31 open-ended questions, yielding 434 team-responses to open-ended questions. Of these, 114 responses (26%) were given by teams where no one answered correctly in the individual round. After the first attempt, one quarter of these teams (25%) submitted a correct answer. An extra 20% (roughly one quarter of the remaining 74%) answered correctly in the second attempt and another 7% obtained a correct answer on the third attempt. Thus, in teams where all members entered the team round with incorrect answers, a majority of teams (52%) figure out the correct answer after three attempts. These findings show that collaboration does not simply serve to propagate correct answers—teams frequently generate correct answers even when there is no correct answer to propagate.

Figure 3 shows how the number of correct answers obtained by team members in the individual round affects the rate at which groups converge towards a correct answer. In groups with two or more members having obtained correct answers in the individual round, 96% of them answer correctly on the first attempt and 100% of teams do so within the three attempts. For groups with just one member having answered correctly in the individual round, 75% of them answer correctly after the first attempt, three times more than

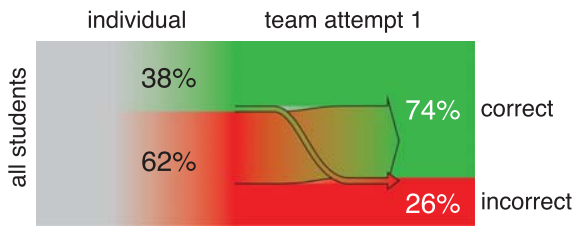


Fig. 2. Percentage of individuals migrating between incorrect (bottom, red online) and correct (top, green online) answers after the first team attempt. The data represent the averages over 31 open-ended questions.

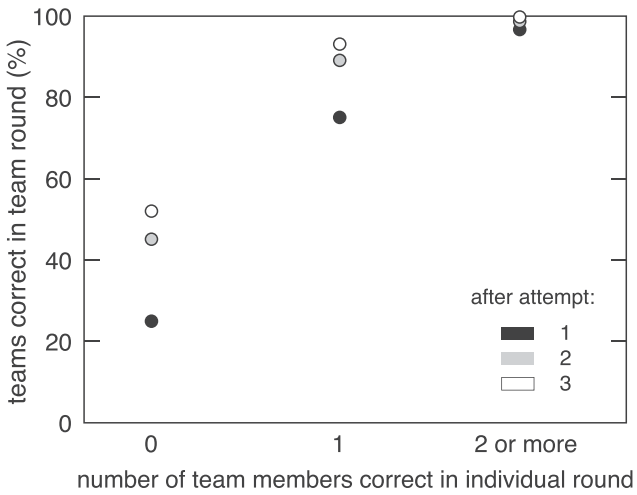


Fig. 3. Dependence of the cumulative percentage of teams obtaining the correct answer for 31 open-ended questions on the number of team members having a correct answer in the individual round. (Number of instances: 0 correct, $N = 114$; 1 correct, $N = 124$; and 2 or more correct, $N = 196$.)

when no one has the correct answer in the individual round. Note, however, that in this case, the teams never reach 100%. So, while the presence of a team member with the correct answer certainly improves the team's performance, the interactions that take place in the team round are more than just a mere propagation of the correct answer.

C. Benefits of collaboration

Next we examine the effect of collaboration as a function of the individual student's ability level. We test the hypothesis that collaboration mainly benefits weaker students by having higher-ability students share their knowledge. Using the average individual assessment scores for five exams, we divided the students into three categories: low ability (average scores in the lowest quartile), high ability (average scores in the top quartile), and average ability (the remainder). Figure 4 shows the mean individual correctness and collaborative correctness by the end of each trial for the low-, average-, and high-ability students. We find that the

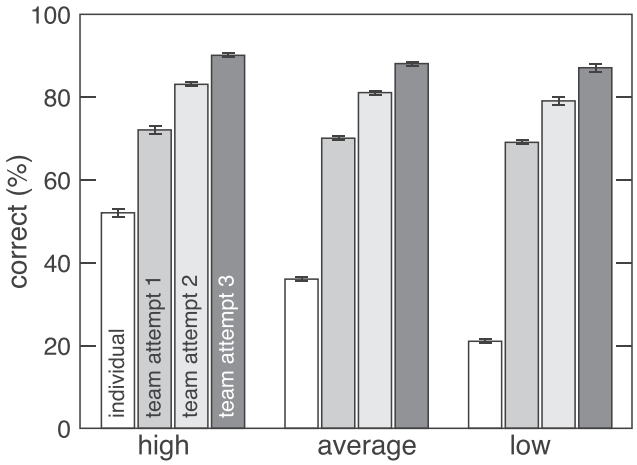


Fig. 4. Average individual (white) and team (gray) scores after each of three attempts. Students' ability levels are determined by the average individual score on five exams (low: bottom quartile; high: top quartile, average: the remainder). Error bars represent the standard error of the mean.

Table II. Median scores of all 14 best-in-team students before and after collaboration, for each of the five assessments. The Wilcoxon Z-statistic is computed to test the differences between best-in-team scores before and after collaboration. Best-in-team students systematically score significantly higher after collaboration. (The differences in the rightmost column all have $p < 0.0001$.)

Assessment	Median score for best-in-team student		
$N = 14$	Before collaboration	After collaboration	Wilcoxon Z-statistic
1	0.74	0.93	3.29
2	0.52	0.72	3.22
3	0.49	0.68	3.31
4	0.50	0.74	3.30
5	0.55	0.84	3.30

mean score of groups surpasses the mean score of all categories of individuals that comprise them. Specifically, we find mean increases ($\pm$standard error) of: 37% ($\langle S_i \rangle = 52 \pm 2\%$ vs $\langle S_t \rangle = 89 \pm 1\%$) for students with high ability, 53% ($\langle S_i \rangle = 35 \pm 1\%$ vs $\langle S_t \rangle = 88 \pm 1\%$) for those with moderate ability, and 66% ($\langle S_i \rangle = 21 \pm 1\%$ vs $\langle S_t \rangle = 87 \pm 2\%$) for those with lower ability. This finding suggests that, on average, students of all ability levels are more likely to answer correctly after collaborating with others than before. However, this analysis focuses on categories of students and cannot exclude the possibility that the strongest student in a team does not benefit from collaborating. We therefore examine the performance of best-in-team students, defined as the student in each team who obtains the highest score in the individual round. Table II compares the performance of these best-in-team students before and after collaboration. These data unequivocally show that these best-in-team students score significantly higher in the team round, indicating that even the strongest students gain from collaborating.

IV. CONCLUSION

Education should be a means for students to construct knowledge and gain expertise. Although much progress has been made in higher education with respect to active-learning and collaborative pedagogies, collaborative assessment techniques are not widely used. Although many instructors view conventional assessments as inadequate, collaborative assessments are not typically considered a viable solution. In this paper, we identify and address four barriers to the adoption of collaborative assessments. First, we argued that two-stage exams incorporate individual accountability. Second, we address the notion that strong students simply provide correct answers to weaker students, and therefore, collaboration during assessments is tantamount to cheating. Our data clearly show that all students (on average) gain from collaborating, with sizable and significant increases in correct answers for students of all ability levels. To further dispel the notion that collaboration merely causes correct answers to propagate within teams, we show that more than half of the teams where no individual has the correct answer still converge to a correct solution within three iterations. Furthermore, while the presence of individuals with the correct answer certainly improves team performance, such teams do not necessarily converge on the correct answer. Third, to minimize the complexity of implementing collaborative assessments, we used an online

platform that automatically grades open-ended and multiple-choice questions and manages team assessments by assigning groups and providing iterative feedback. Given that instructors need only input assessment questions and that the management and grading are fully automated, this system makes the effort involved in implementing collaborative assessments comparable to that of conventional exams. While our implementation of collaborative assessments might involve more exam time over the course of the semester than in traditional physics courses (90 minutes per assessment, 5 times during the semester), the pedagogical purpose of these exams is not simply to assess students but also to provide students with an opportunity to discuss problems with one another and learn through this experience.

Given these findings, it is up to the readers of this paper to help overcome the fourth barrier: resistance to changing established practices. Although we do not address the nature or the complexity of the knowledge and skills acquired during collaborative exams, we show that even in institutions where established practices have a very long history, collaborative exams can be effectively implemented with significant benefit to all students. We hope the realization that collaboration can turn assessment into a learning opportunity will encourage instructors to adopt collaborative assessment practices more broadly.

ACKNOWLEDGMENTS

Several people contributed to this work. H.J. conceived of the basic idea for this work. E.M., K.M., and H.J. designed and carried out the experiments, and analyzed the results. E.M. supervised the research and the development of the manuscript. H.J. and N.L. co-wrote the manuscript; all authors subsequently took part in the revision process and approved the final copy of the manuscript. E.M. and K.M. provided feedback on the manuscript throughout its development. This research was supported by the National Science Foundation under Grant No. NSF DUE 1504664.

APPENDIX: SAMPLE ASSESSMENT QUESTIONS

1. Multiple-choice question

A classmate leaves a message on your voice mail betting that you cannot throw a stone hard enough so it lands on the roof of a 20 -m high building. As you stare out of your window pondering whether to accept the challenge, the well in the courtyard suddenly gives you an idea. You drop a stone into the well and note that you hear a splash 4.0 s later. You repeat the experiment with another stone, but this time, you throw the stone down as fast as you can. This time the splash comes 3.0 s after the stone leaves your hand. Armed with this information you carry out a quick calculation and then you call back your friend. Do you accept the bet?

(A) Yes, but it is close
(B) Yes, easily
(C) No, but it is close
(D) No, not by a wide margin
(E) Insufficient information in this problem

2. Open-ended question

Three books, each of inertia m, rest on the floor of an elevator. The elevator starts at the first floor and rises to the sixth floor. It travels at a constant speed between the second and fifth floors, as it rises by a total distance h. Enter an expression for the work done by the bottom book on the middle book during the passage from the second to the fifth floors in terms of m, h, and the acceleration due to gravity g.

[a]Electronic mail: wanypie@gmail.com
[b]Electronic mail: nlasry@me.com
[c]Electronic mail: kellymillervt@gmail.com
[d]Electronic mail: eric.mazur@gmail.com

[1]P. Heller, R. Keith, and S. Anderson, "Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving," Am. J. Phys. **60**, 627–636 (1992).
[2]P. Heller and M. Hollabaugh, "Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups," Am. J. Phys. **60**, 637–644 (1992).
[3]L. S. Vygotsky and Michael Cole, *Mind in Society: The Development of Higher Psychological Processes* (Harvard U.P., Cambridge, MA, 1978).
[4]W. Damon and E. Phelps, "Critical distinctions among three approaches to peer education," Int. J. Educ. Res. **13**, 9–19 (1989).
[5]G. W. Rieger and C. E. Heiner, "Examinations that support collaborative learning: The students' perspective," J. Coll. Sci. Teach. **43**, 41–47 (2014).
[6]H. Shirouzu, N. Miyake, and H. Masukawa, "Cognitively active externalization for situated reflection," Cognit. Sci. **26**, 469–501 (2002).
[7]P. Black and D. Wiliam, "Assessment and classroom learning," Assess. Educ. **5**, 7–74 (1990).
[8]J. P. Murry, "Better testing for better learning," Coll. Teach. **38**, 148–152 (1998).
[9]B. Gilley and B. Clarkston, "Collaborative testing: Evidence of learning in a controlled in-class study of undergraduate students," J. Coll. Sci. Teach. **43**, 83–91 (2014).
[10]M. Lusk and L. Conklin, "Collaborative testing to promote learning," J. Nurs. Educ. **42**, 121–124 (2003).
[11]M. K. Smith *et al.*, "Why peer discussion improves student performance on in-class concept questions," Science **323**, 122–124 (2009).
[12]C. E. Wieman, G. W. Rieger, and C. E. Heiner, "Physics exams that promote collaborative learning," Phys. Teach. **52**, 51–53 (2014).
[13]J. G. Lambiotte *et al.*, "Cooperative learning and test taking: Transfer of skills," Contemp. Educ. Psychol. **12**, 52–61 (1987).
[14]J. V. Shindler, "Greater than the sum of the parts? Examining the soundness of collaborative exams in Teacher Education Courses," Innovative Higher Educ. **28**, 273–283 (2004).
[15]P. G. Zimbardo, L. D. Butler, and V. A. Wolfe, "Cooperative college examinations: More gain, less pain when students share information and grades," J. Exp. Educ. **71**, 101–125 (2003).
[16]S. Sandahl, "Collaborative testing as a learning strategy in nursing education," Nurs. Edu. Perspect. **31**, 142–147 (2010).
[17]S. A. Stearns, "Collaborative exams as learning tools," Coll. Teach. **44**, 142–147 (2010).
[18]"Learning Catalytics Home Page," <https://learningcatalytics.com>.
[19]L. K. Michaelsen, M. Sweet, and D. X. Parmelee, *Team-Based learning: Small Group Learning's Next Big Step: New Directions for Teaching and Learning* (Wiley, Lexington, KY, 2011), Vol. 116.
[20]P. C. Blumenfeld *et al.*, "Motivating project-based learning: Sustaining the doing, supporting the learning," Educ. Psychol. **26**, 369–398 (1991).
[21]C. H. Crouch and E. Mazur, "Peer instruction: Ten years of experience and results," Am. J. Phys. **69**, 970–977 (2001).
[22]D. Hestenes, M. Wells, and G. Swackhamer, "Force concept inventory," Phys. Teach. **30**, 141–158 (1992).
[23]J. F. Zipp, "Learning by exams: The impact of two-stage cooperative tests," Teach. Sociol. **35**, 62–76 (2007).