CAN TWO-STAGE EXAMS IMPROVE RETENTION AND DECREASE ACHIEVEMENT GAPS?

Nathan L. Kirk, Lori J. Kayes, Jeff Chang and Indira Rajagopal

Abstract

Two-stage exams are summative assessments taken in two parts: 1) a traditional individual exam and 2) a group exam. These exams encourage collaborative discussion to promote deeper thinking and understanding of classroom material transforming exams into additional learning experiences. Two-stage exams can improve student performance, learning, long-term retention, and even reduce student test anxiety. We implemented a two-stage exam in a ~1,100 student Principles of Biology for majors course. To assess the efficacy of the group exam, we examined changes in student performance for questions differing in their level Bloom's level taxonomy. We also measured short-term and long-term knowledge in subsequent courses. In self-reported data, a majority of students felt they benefited from group discussions, better understood and remembered content, and retained more of the material when they had questions on the group exam and individual exam versus on the individual exam only. Scores increased on questions in the group exam compared to the individual exam demonstrating peer instruction and productive discussion of material within a majority of the groups. There were even increases among top performing students indicating the exam was helpful for a majority of students by facilitating useful classroom discussion and increasing their performance.

Introduction (Subject/Problem)

Active and collaborative learning has been shown to increase retention and comprehension of material (Chi and Wylie, 2014; Freeman et al. 2014; Snyder et al. 2016). Thus as a response to improve learning at institutions of higher education, many classrooms and faculty have moved to a collaborative and active learning model in the classroom (but see Stains et al., 2018). However, typically collaborative and active learning are limited to the in-class environment and used for formative assessment. To maximize the benefits of collaborative learning, there is a potential benefit of taking collaboration one step further to the summative assessment. While there are multiple ways a collaborative assessment could be implemented, we will discuss the "two-stage" exam model where individual student learning is still assessed while allowing for collaboration on part or all of the assessment.

Two-stage exams are summative assessments that are taken in two parts: 1) A traditional individual exam and 2) a group portion that encourages collaborative discussion (Stearns, 1996; Gilley and Clarkston, 2014). Two-stage exams have been shown to increase student learning in the short term (Zipp, 2007; Bruno et al. 2017), increase exam averages (Zimbardo et al. 2003; Bruno et al. 2017), increase performance on lower level exam questions (Bleedlove et al. 2012), improve classmates relationships with one another (Sandahl, et al. 2010), and provide immediate feedback (Stearns, 1996; Gilley and Clarkston, 2014). Additionally, the two stage exam has the potential to turn the exam into a learning experience (Bruno et al. 2017), to reduce test anxiety (Lusk and Conklin, 2003, Zimbardo et al., 2003; but see Bleedlove et. al. 2004 and 2012), help students build on their strengths, increase student motivation to learn and to prepare, and improve student thinking (Zimbardo et al. 2003). While studies have shown that a two

stage exam increases student learning in the short-term (see above), there were limitations on many of these studies due to lack of control groups, other changes to courses, or potential differences in course composition. Additionally, no studies to date have examined the long-term retention of having items on a group exam versus not.

Another potential benefit of collaborative and active learning is a decrease in achievement gaps often seen in underrepresented groups especially in STEM classrooms (Haak et al. 2011; Eddy and Hogan, 2014; Snyder, et al. 2016). In particular, two-stage exams have the potential to decrease the achievement gaps between students (Bruno et al. 2017). We implemented collaborative, active learning in our classroom to make it more inclusive and equitable for *all* students and to improve retention of content. The two-stage exam was implemented due to the potential for increasing student knowledge and decreasing achievement gaps. However, data were limited on the two-stage exams and reduction in achievement gaps in specific underrepresented groups. Therefore, in addition to examining long-term impacts of the two-stage exam on retention, we sought to delineate the impacts of two-stage exams on different demographic groups to determine if, not only could we make our classroom more inclusive, but even go so far as to make our exams more inclusive as well.

As we implemented a two-stage exam we asked the following research questions: **1)** Are group exams effective for increasing student **performance**? **2)** Is there a correlation between student **perception** of benefits of the group exam and their course or exam performance? **3)** Does participating in a group exam increase **short-term** (< 2 mos) retention on specific topics? **4)** Does participating in a group exam increase **longer-term**

(> 8 mos) retention on specific topics? **5)** Does the group exam decrease **achievement gaps** for underrepresented (UR) groups?

Methods (Design and procedure)

Student population and course demographics

In Winter Quarter of 2016 and 2017, we implemented a two-stage exam in a ~1,100 student introductory biology (life science majors) course series. We used the first year (2016) to assess student performance gains and satisfaction and the second year (2017) to examine short- and long-term retention. This 4-credit per quarter class with accompanying laboratory is comprised of mostly second year students (2016 = 47%; 2017 = 44%). The course was divided into two large sections (2016: 472 and 380 students; 2017:463 and 373 students) that were co-taught by two instructors and a third smaller (2016: 236 students; 2017: 234 students) section was taught by a third different instructor.

To assess short- and long-term retention we also implemented an optional extra credit quiz in the third term of the introductory biology course (Spring 2017) and a required pretest in a subsequent cellular and molecular biology course (Fall 2017). Using these courses allowed us to see if student knowledge of material was retained by this cohort of students for ~2 and ~ 8 months, respectively.

Setup and design

In the introductory biology course, there are two midterms and a final exam. We modified all three of these summative assessments into two-stage exams as described below. Approximately 80% of the exam grade was from the individual exam score and

20% from the group exam score. The two-stage exam was proctored across two days with separate individual and group portions. The first stage was an individual exam consisting of 50 multiple choice questions that we administered on a Monday night for 80 minutes. After the individual exam, we created four subsets of the exam (10 questions each) that would comprise of the group portion (second stage). The same week in lab (running Tuesday to Friday), lab groups were given one of the four versions of this subset exam to collaboratively answer in student groups of 3-4 individuals. Student groups were generally self-selected by students based on their seating arrangement in the lab rooms. Laboratory graduate teaching assistants proctored this group exam for 30 minutes. To control for potential intergroup variability, students were allowed to talk to all other students in their lab section. However, they had to reach consensus as their small group.

Performance gains

Part of the appeal of students taking exams collaboratively was turning the summative assessments into learning opportunities. In order to determine if students were benefiting from their discussions, we measured performance gains comparing individual exam questions to that of the group exams on both of the midterms. Given the two midterm exams, there were 80 questions that were asked individually of all students and by a subset of students on one version of the group exams. We used the discrimination index as a metric to validate questions with ≥ 0.3 as the cutoff. This allowed us to remove questions that high-performing students disproportionately got wrong. Unless otherwise noted, all statistical analyses in this study were conducted in the base package of R (v.3.3.1).

As students tend to do better on collaborative assignments than they do individually, we wanted to ensure that the two-stage exam did not lead to grade inflation. To avoid grade inflation, we attempted to move our exams up in complexity level to higher on the Bloom's score. There were few pedagogical changes between the courses (2014-6) besides adding the two-stage exam. To assess how our exams have changed over time, we categorized our questions according to Bloom's taxonomy levels of complexity (Bloom, 1956). We averaged the Bloom's taxonomy level for individual exam questions to get a exam level Bloom's taxonomy score. We then examined the effect of group exams on our grade distribution before and after implementation of the two-stage exam. Final grade distributions from 2016 were compared to 2014 and 2015 normalized to the same number of students using a chi square test.

Student satisfaction

To determine student satisfaction, we polled a subset of students using an in-class audience response system. At the beginning of a class, we asked students to respond to 5 opinion-based questions using a 5 point Likert scale. To assess student preference and opinion, we used binomial tests to compare positive responses (strongly agree and agree) to negative responses (strongly disagree and disagree) for each question. To determine if student perception was related to their overall performance in the class, we used a proportional-odds model comparing self-reported satisfaction to their final course grades and cumulative link model comparing self-reported satisfaction to whether or not their exam score increased after the group portion of the two-stage exam.

Short-term retention of material

To evaluate short-term retention (2 months) of course material tested on the group exam, we delivered the same assessment (a subset of 4 questions related to DNA structure and function) in the subsequent course (Spring 2017) as an online extra credit assignment. Two of these questions related to DNA backbone structure and two related to DNA basepairing. Students (n=238) saw all of these questions on the Winter 2017 individual exam and one of these questions again on a group portion. For each question, we used logistic regression to compares odds of getting correct answer as a function of whether they saw a similar question (i.e. backbone structure vs. base-pairing) on their previous term group exam.

Long-term retention of material

To determine long-term retention (> 8 months) of course material, we delivered a pre-assessment in an upper-division course along with a self-reported demographic survey in Fall 2017. The pre-assessment included ten questions: the same 4 DNA structure and function questions from the short-term assessment, 2 additional novel DNA questions, and 4 additional questions from each of the group exam versions from Winter 2017 on different topics: information processing, protein folding, protein localization and the endomembrane system. Although 127 students took this pretest, 94 students provided consent and only 42 took this two-stage exam in Winter 2017. An additional 16 students took the two-stage exam in Winter 2016 and the remainder (n=36) took introductory biology elsewhere. Similar to short-term retention, we used logistic regression to examined each question separately and compared odds of getting correct answer to whether or not they had the topic on their group exam in the first term (Winter 2017). At the conclusion of the required pretest, students could provide self-reported demographic

data. We used these demographic data to determine if there was a performance gaps among different genders and among underrepresented groups by including them as fixed factors in our models. Finally, we compared between students that took introductory biology in Winter 2016 and Winter 2017 to see if there was an effect of when they took the course on this pre-assessment.

Results (Analysis and Findings)

Performance gains

In Winter of 2016, there was an increase in student performance on the group portion of the exams compared to the individual exams. Collectively, students earned significantly higher scores on the group portion for all but one of the questions (98.3%, 59/60, $t=8.9$, $df=118$, $p=2xe^{-14}$). On average, for all the questions, the percentage of correct answers increased from 66.8% to 91.7% on the group portion of the exam. This increase in student performance was largest for the questions that were ranked the highest on Bloom's Taxonomy scale of complexity (Fig 1).

In addition, we assessed increased student performance on the most difficult questions (n=23) from the individual exam where <60% of students selected the correct answer (average score=47%+/-11.8%). During the group exam, the average score on these questions was 82.2%+/-14.0, which was a significant improvement ($t=9.1$, $df=44$, $p=1.0xe^{-11}$) compared to the individual exam average. On these questions, 87.0% (20/23) improved to ≥60% selecting the correct answer, 73.9% improved to ≥75% correct and 34.8% (8/23) improved to ≥90% correct indicating productive conversations during the group portion of the exam.

The group component of the two stage exam is unlikely to have contributed to undue grade inflation in this class (Fig 2). When comparing course grade distribution over time, there was not a significant difference between 2014 and 2016 (χ^2=15.3, df=11, p>0.10). In addition, DWF rates (number of students that earned D's, F's or withdrew from the course) remained consistent with introduction of the two-stage exam (Fig 3). However, this was associated with a significant increase in Bloom's level (KW=1964, p=9.7xe^{-11}) for the questions asked (Fig 3).

Student satisfaction

In general, students indicated were positive reactions regarding the two-stage exam implementation. Of the ~170 students that responded to our in-class survey, a large majority of students felt that they benefited from the conversations (91% agree or strongly agree) during and better understood what the exam questions were asking (80 %) following the group portion of the exam (Table 1). Many students (74%) also thought that the group portion of the exam helped them retain information longer. Of the students in this class, 93.5% of them actually benefited from the group exam in terms of overall exam score. There was a slight negative correlation between belief that they benefited from the group exam and their final course grade indicating that some high-performing students did not perceive benefit from the two-stage exam (β=-0.003, p=0.02). However, there was no correlation between actual group exam score and student perceived benefit (β=0.062, p=0.53).

Short-term retention of material

There were significant increases in student performance on the reasked questions 2 months later compared to how they did on the individual exam during the first iteration

(binomial test: $p= 4.8xe^{-05}$). On average, students outperformed their original scores for these 4 questions by 0.3 (7.5%) ± 1.2 (30%) (Fig 4). When each question was analyzed separately, there was not a significant effect of previously having a DNA base-pairing or backbone structure question previously on the group portion of the exam during the first iteration on the student performance on these questions 2 months later. Although not significant, on 3 out of 4 questions students were more likely to get the question correct 2 months later if they had matching topics on their group exams during the first iteration.

Long-term retention of material and achievement gap

There were a few significant effects of the group exam on long-term retention (>8 months later). We detected a significant effect of group exam in 5 of the 10 questions asked including 3 of 6 on DNA structure and function. On these questions where we detected an effect of the group exam, students that took a two-stage exam (either 2016 or 2017) were 2.7-5 times more likely to get the DNA structure function questions right compared to students that did not have a two-stage exam (Table 2). There was significant effect of when they took our course for 3/10 questions (Table 2), so we chose to examine only students who recently took the class. When examining only the students that took the Winter 2017 class (n=41), there was no effect of having a DNA backbone or basepairing question on their performance on these questions 8 months later. However, students were 3.9 times more likely to get an endomembrane system question right if they had this topic on their group exam (p=0.04). Finally, there was no effect of underrepresented group or gender as all students did equally well regardless of self-reported status (Table 2).

Discussion

Two-stage exam can improve student performance and short and long-term retention on biology exams. Supporting findings by previous studies (Zimbardo et al. 2003; Bruno et al. 2017; Zipp, 2007), we also showed that two stage exams increased student exam averages and there were gains in short-term retention of material. Additionally, we see some positive trends on long-term retention that may be attributable to two-stage exams. However, it did not appear to matter if the students had seen a questions on a previous group exam or not in relation to long-term retention and the long-term retention data are confounded by the fact that *all* of the students that participate in our courses during the group exam showed increased performance in the following fall compared to students that either took the Introductory Biology course prior to 2014 or took at a different institution. Therefore, it is difficult to directly tie this increased performance to the group exam but rather it could be attributable to vertical alignment between our courses. However, we are cautiously optimistic in saying that the two-stage exam appears to increase student retention over time and we are collecting more data to help substantiate this premise. While we did not find that the two-stage exam contributed to a decrease in the achievement gap between male/female or underrepresented/non-underrepresented groups, was no difference in the long-term retention for underrepresented groups (underrepresented minority or women) compared to non-underrepresented groups. However, we did not collect data to substantiate whether there is an achievement gap in courses initially or historically at this point in our degrees. Thus, if there was an achievement gap it was eliminated by the beginning of the third course.

Student perceptions of the benefits of the group exam were indeed correlated with their course performance but not with their group exam score. It is not surprising that higher performing students' might perceive less benefit from the two-stage exam than lower performing students, since they already performing well on the exams. Since we have implemented the two-stage exam to increase the inclusivity and equity in our assessments, we feel that the lack of correlation of the group exam score and perception of benefit supports the idea that the two-stage exam is in fact contributing to the inclusivity of these assessments. The perceived benefit is likely due to the two-stage exam allowing lower performing students to review concepts, move towards more expert like knowledge and improve their exam scores by having discussions over content during the group portion of the exam. Additionally, no students are negatively impacted by the two-stage exam. Perhaps more compelling is the increase in student perceived retention, satisfaction and benefit from the group exam coupled with the fact that faculty were able to increase the cognitive complexity (based on Bloom's (1956) level) of their exams without grade inflation. The grade buffer provided by the two-stage exam allowed for faculty to change their summative assessment to ask questions that applied concepts that were never previously asked on exams without disadvantaging the students. Additionally, reworking portions of the exam with a group contributed to students gaining a fuller understanding of the questions that were being asked rather than potentially leaving an exam being lost, deflated or confused.

When considering implementation a two-stage exam, some of the benefits to students are: 1) earnest discussion of course material, 2) exams become learning opportunities, 3)

students more willing to question each other than instructors, 4) potentially increased retention of material in pretest in a follow-up class, and 5) increase satisfaction with exam experience. For faculty, many of the same benefits are possible as for students and, as mentioned previously the ability, to increase the cognitive complexity of the exams. One anecdotal benefit of the group exam is that students no longer referred to exam questions as "tricky" after they worked through them in groups and exam questions ceased to be contested due to unfairness. Some challenges or considerations that faculty need to consider prior to implementing a two-stage exam are 1) how to obtain student "buy-in", 2) potential group dynamic issues, 3) faculty or staff workload associated with regrading and group grading, 4) loss of class time for other activities (individual exam, class or lab time), and 5) potential disparities in group assignment if not randomly assigned. Some recommendations that we make for effectively implementing a group exam are to: 1) have the group exam count for points, 2) to alleviate unbalanced groups, we allowed students to work with anyone in the room on the group portion of the exam (except instructors), 3) contact your disability office prior to implementation to ensure fair access to all, and 4) have a plan for inputting group grades and the additional grading load (if not multiple choice exam especially.

References

Bleedlove, W., T. Burkett, I. Winfield. (2004). Collaborative testing and test performance. *Academic Exchange Quarterly*, 4,36-40.

Breedlove, W., Burkett, T., & Winfield, I. (2012). Collaborative testing and test anxiety. *Journal of the Scholarship of Teaching and Learnin*g, 4(2), 33-42.

Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook I: The cognitive domain*. New York: David McKay Co Inc.

Bruno, B.C., J. Engels, G. Ito, J. Gillis-Davis, H. Dulai, G. Carter, C. Fletcher, and D. Böttjer-Wilson. (2017). Two-stage exams: A powerful tool for reducing the achievement gap in undergraduate oceanography and geology classes. *Oceanography* 30(2),198–208.

Chi, M. T. H. and R. Wylie (2014). The ICAP framework: linking cognitive engagement to active learning outcomes. *Educational Psychologist* 49(4): 219-243.

Eddy, S.L. and K.A.Hogan (2014). Getting under the hood: how and for whom does increasing course structure work? *CBE—Life Sciences Education* (13), 453–468.

Freeman, S., S.L. Eddy, M. McDonough, M.K. Smith, N. Okoroafor, H. Jordt, and M.P. Wenderoth (2014). Active learning increases student performance in science, engineering, and mathematics. *PNAS* (111), 8410–8415.

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

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

Ives, J. (2015). Measuring the learning from two-stage collaborative group exams. *Proceedings of the Physics Education Research Conference*: 123-126.

Lusk, M., and L. Conklin. (2003). Collaborative testing to promote learning. *The Journal of Nursing Education* 42(3):121–124.

Sandahl, S.S. (2010). Collaborative testing as a learning strategy in nursing education. *Nursing Education Perspectives* 31(3):142–147.

Snyder, J.J., J.D. Sloane, R.D.P. Dunk, and J.R. Wiles. (2016). Peer-led team learning helps minority students succeed. *PLoS Biology* 14(3):e1002398.

Stains, M., J. Harshman, M.K. Barker, S.V. Chasteen, R. Cole, S.E. DeCheene-Peters, M.K. Esson, J.K. Knight, F. A. Laski, M. Levis-Fitzgerald, C. J. Lee, S. M. Lo, L. M. McDonnell, T. A. McKay, N. Michelotti, A. Musgrove, M. S. Palmer, K. M. Plank, T. M. Rodela, E. R. Sanders, N. G. Schimpf, P. M. Schulte, M. K. Smith, M. Stetzer, B. Van Valkenburgh, E. Vinson, L. K. Weir, P. J. Wendel, L. B. Wheeler, A. M. Young. (2018). Anatomy of STEM teaching in North American universities. *Science* 359: 1468-1470.

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

Zimbardo, P.G., L.D. Butler, V.A. Wolfe. (2003). Cooperative college examinations: more gain, less pain when students share information and grades. *Journal of Experimental Education* 71: 101-25.

Zipp, J.F. (2017). Learning by exams: The impact of two-stage cooperative tests. *Teaching Sociology* 35: 62-76.

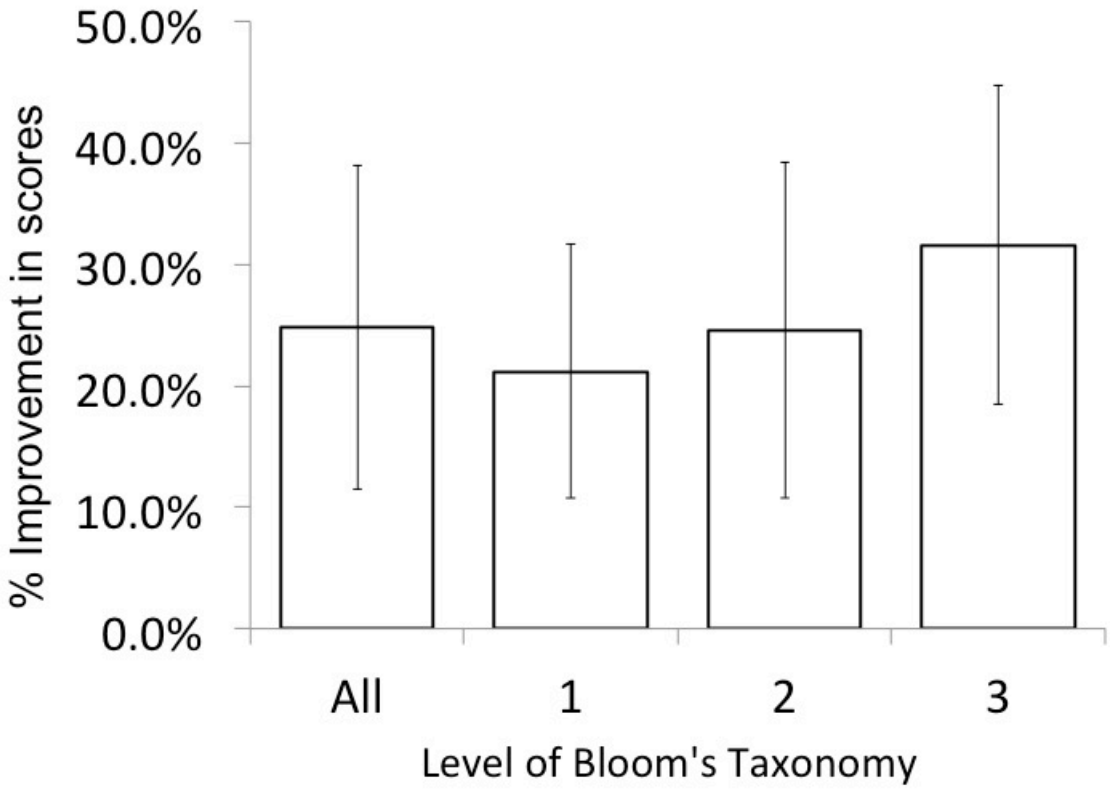

Figure 1. Improvement of scores on questions from the individual to the group exam (All). Improvement is further split by Bloom's taxonomy of the questions (1-3).

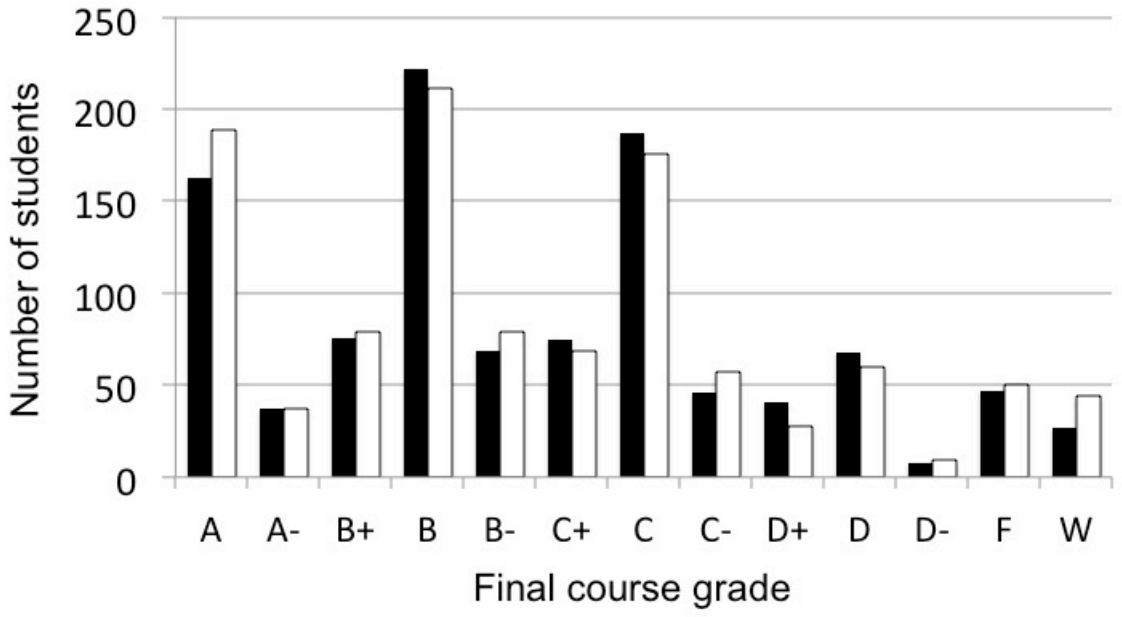

Figure 2. Histogram of course grade distributions for 2014 (Black) and 2016 (White). There was no significant change despite adding the group exam.

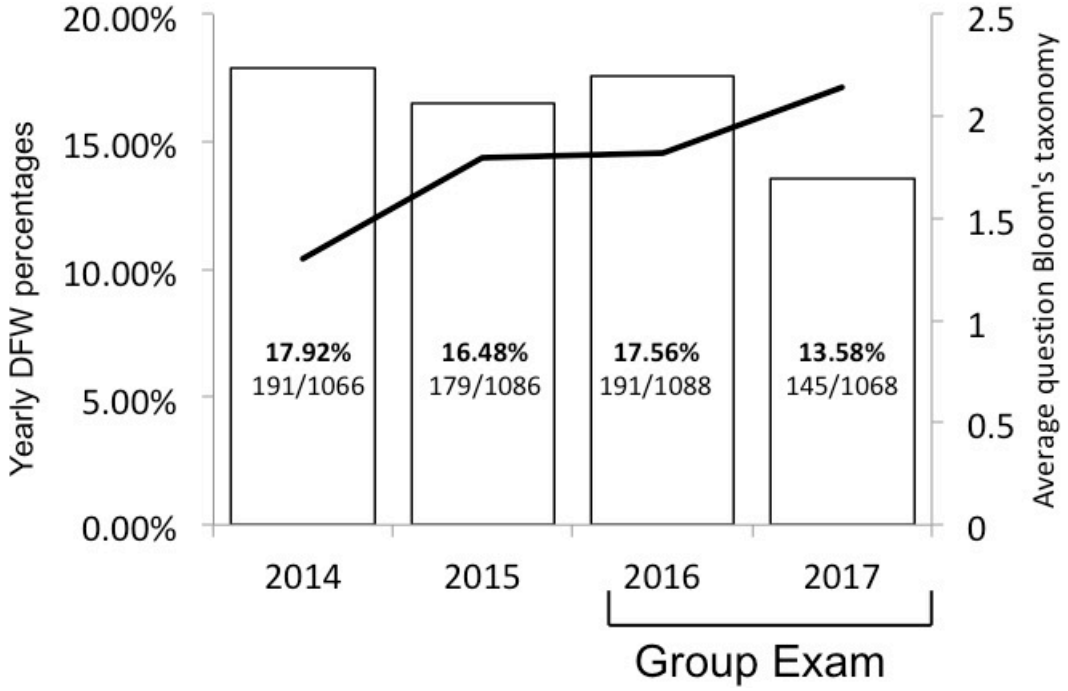

Figure 3. Percentage of students earning a D, F or withdrawing from the class (DFW) by year on the left axis. The right axis (black line) indicates the average exam question level for Bloom's taxonomy. The group exam was introduced in 2016 and 2017.

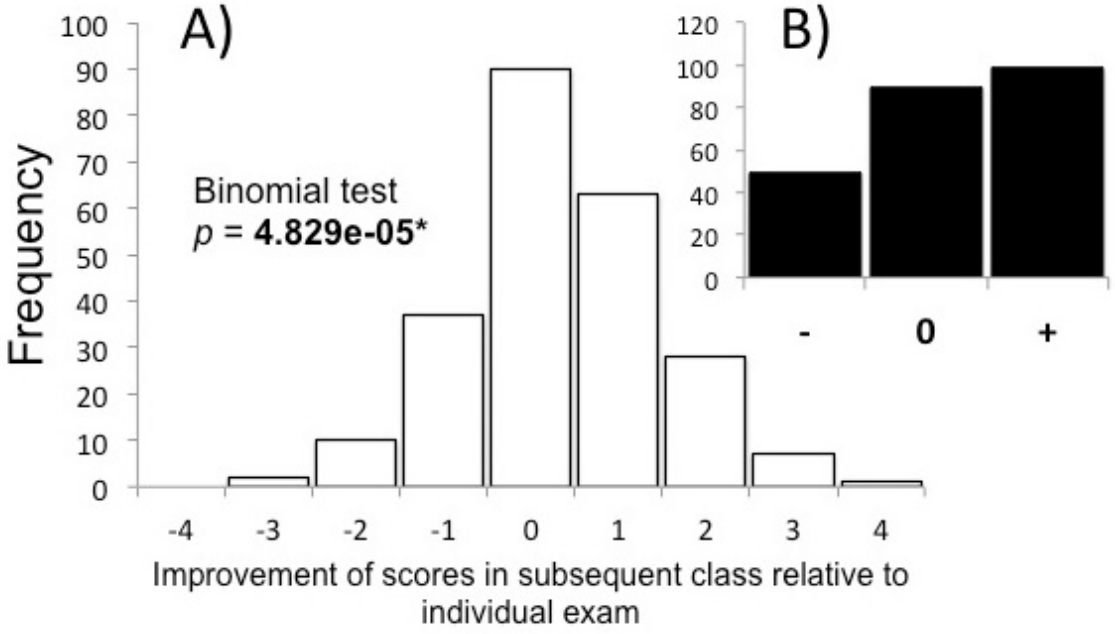

Figure 4. Comparison of scores for the 4 questions taken on the individual exam and again as a pre-test of a subsequent class (short term retention; ~2 mo.). **A)** The change in exam score is given for these 4 questions. B) Change in exam score collapsed into whether the student increased (+), decreased (-) or did not change their performance on these 4 questions in the subsequent class.

Table 1. Student satisfaction self-reported data survey and results. Most popular responses for each question are denoted in bold.

1) I understand what the questions are asking better during the group exam, than during the individual exam.

a) Strongly agree	39% (64)
b) Agree	**41% (67)**
c) Neither agree nor disagree	15% (7)
d) disagree	4% (7)
e) Strongly disagree	2% (3)
	(n=165)

2) I benefit from the discussions that occur during the group exam.

a) Strongly agree	**46% (77)**
b) Agree	45% (74)
c) Neither agree nor disagree	5% (9)
d) disagree	4% (6)
e) Strongly disagree	1% (2)
	(n=168)

3) I remember questions that are reasked on the group exam better than questions that are only on the individual exam.

a) Strongly agree	30% (52)
b) Agree	**38% (67)**
c) Neither agree nor disagree	24% (42)
d) disagree	6% (10)
e) Strongly disagree	2% (4)
	(n=175)

4) I remember material reasked on the group exam better than material only covered on the individual exam.

a) Strongly agree	25% (43)
b) Agree	**40% (70)**
c) Neither agree nor disagree	27% (47)
d) disagree	6% (11)
e) Strongly disagree	2% (3)
	(n=174)

5) I feel that the group exams help me retain information longer.

a) Strongly agree	31% (54)
b) Agree	**43% (75)**
c) Neither agree nor disagree	21% (36)
d) disagree	2% (4)
e) Strongly disagree	3% (6)
	(n=175)

Table 2. Long term effects of the group exam split by the 10 questions (Q) asked on the pre-assessment from a subsequent course (~8 mo. after). Coefficients from binomial regressions (response = question answered correctly) are presented using the following explanatory variables: whether they took a group exam (Group), were part of an underrepresented minority group (URM), gender and whether they identified as latin@. Significant effects are highlighted in bold. Year represents if there was a significant difference if they took the group exam in 2016 or 2017.

Q	Group	std. error	odds ratio	p value	Year	URM	gender	latin@
1	**1.4263**	**0.517**	**4.163266573**	**0.0058**	no	0.61	0.18	0.63
2	**1.5997**	**0.4832**	**4.951546738**	**0.00093**	**yes**	0.7213	-0.2935	-0.07361
3	0.4055	0.4559	1.500052339	0.374	no	0.507	-0.6966	-1.3275
4	**1.5404**	**0.5559**	**4.66645648**	**0.00559**	**yes**	-0.7783	-0.2181	0.6614
5	0.7397	0.4827	2.095306828	0.125	no	0.4032	-0.6246	0.4412
6	**1.1896**	**0.5586**	**3.285766638**	**0.03321**	**yes**	1.707	0.3232	0.3864
7	**1.0116**	**0.4893**	**2.749997493**	**0.0387**	no	0.7695	-1.091	1.768
8	1.38E-01	4.50E-01	1.148205168	0.759	no	0.3405	-0.78	0.6885
9	-0.1278	0.4574	0.880029367	0.78	no	0.1904	-0.2148	3.84E-01
10	0.009368	0.453927	0.990675743	0.984	no	-0.0823	0.1407	0.9209