Learning Benefits of Collaborative Exams

Kseniya Garaschuk

University of the Fraser Valley

In this paper, we present a case study to examine student performance and perceptions of collaborative exams in a first-year calculus course. In line with Vygotsky's theory, we track students' individual performance versus group performance to examine how the knowledge is constructed within the active learning community through discussion, collaboration, and communication. We then apply grounded theory to analyze qualitative student comments, allowing themes to emerge from the surveys. Overall, the data reveals widely positive effects of group exams on student learning and improved attitudes in students' perceptions of formal assessment as a learning tool. We highlight the importance of setting up an appropriate learning environment and opportunities for group work throughout the term.

Keywords: collaborative exams, group assessments, group work

As argued by Vygotsky (1978, p.100), ``Language is the main tool that promotes thinking, develops reasoning, and supports cultural activities like reading and writing". Indeed, numerous studies in a variety of settings and disciplines show that collaborative work and peer instruction enhance student learning and increase student success (e.g. Crouch & Mazur, 2001; Freeman et al., 2014; Kogan & Laursen, 2015; Rieger & Rieger, 2020; Springer et al., 1999). In practice, the majority of active or social learning happens during regular class time and within take-home group projects. However, constructive alignment of course design dictates that there must be clear connections between learning outcomes, teaching practices and student assessments. (Biggs & Tang, 2011). In this paper, we focus our attention on assessments and the effects of adopting active components from the learning stage to the evaluation stage in a standard first-year calculus course.

While 'assessment for learning' is a better pedagogical practice than 'assessment for evaluating', there is still a strong prevalence for using a system of tests and exams during the learning process (Buhagiar, 2007). Continuous feedback regarding student progress appears to be the most reliable and effective way to improve student learning. So, to encourage a more

thoughtful and holistic approach to evaluating students' knowledge, various educational organizations, such as the National Council of Teachers of Mathematics (NCTM) (1995) and Mathematical Association of America (MAA) (1999; 2006), have developed frameworks for assessment; these frameworks agree on many of the same principles. The principles outlined by MAA (2006) recommend that assessments have a cyclic nature, inform both teaching and learning, promote follow-up actions, use multiple measures of performance, measure what is worth assessing, and provide opportunities for engagement to a diverse set of students. To be fully realized, several of these principles must be reflected in general course design and instructional approach. However, many of the principles can be achieved by simply adjusting the format of an assessment. If good assessment practices can be built into the design of the test itself, then even novice and inexperienced instructors can benefit from it. Taking it one step further, with proper implementation of an advantageous design, even summative assessments can become formative-summative as explained below.

There are a variety of collaborative exam formats. In this paper, we work with *two-stage group exams* consisting of the following stages:

- Stage 1, Individual: standard formal assessment.
- Stage 2, Group: students revisit the same/similar problems in small groups.

Student papers are collected after the individual stage; for the group stage, only one paper per group is distributed, collected and marked per group. The total exam grade is calculated as a combination of the grades from the two stages with the individual stage grade usually composing 70-80% of the total grade. This format has several advantages. By design, this assessment format satisfies several of the principles mentioned in NCTM (1995) and MAA (2006). Specifically, this format emphasizes the cyclic nature of assessment practices, necessitates follow-up, allows for diverse approaches to be shared and investigated. A summative assessment in itself, the format directly speaks to 3 out of 5 effective strategies for formative assessment identified by MAA (2006), namely providing feedback to move students forward, encouraging ownership of learning, and using students as learning resources for each other.

There is an existing body of literature that evaluates the effectiveness of group exams and explores student perspectives on the format in a variety of disciplines (Drouin, 2010; Leight et al., 2012; Levy et al., 2018; Jang et al., 2017; Rao et al., 2002; Rieger & Heiner, 2014; Sandahl, 2010; Zomorodian et al., 2012). Several recent studies examined the two-stage format articular and its impact on student short-term material retention (Garaschuk & Cytrynbaum, 2019; Gilley & Clarkston, 2014; Levy et al., 2018; Kinnear, 2021; Rieger & Rieger, 2020). Within this body of literature, researchers analyzed several factors associated with two-stage exams, including student demographics, question format, test length, grade distribution, and long-term material retention. As this is our focus, for the rest of the paper, we use the word *exams* to refer to any type of formal assessment, such as quizzes, midterms, and final exams, and use *group exams* to mean two-stage group exams.

Here, we present a case study to examine student performance and perceptions of group exams. We analyze quantitative and qualitative student data from a first-year differential calculus course with four group exams given during the term. We analyze students' individual and group performance on constructed response questions and track the evolution in student answers to examine how the knowledge is shaped within each group. We also track changes in group compositions throughout the term. We also survey students to reveal their perspectives on the format. We apply grounded theory to analyze qualitative student comments, allowing themes to emerge from the surveys. Finally, we compare our findings to previous studies, interpret the results, and discuss future directions of study.

Data Collection

This study took place at a medium-sized public teaching-intensive Canadian university in a mainstream differential calculus course with 68 students enrolled at the beginning of the term (7 students withdrew from the course before completing it). The course spanned 13 weeks with four hours of class time each week.

From the beginning of the term, the instructor identified effective group work as one of the desired learning outcomes. In designing the course with constructive alignment, the instructor aligned the teaching and assessment practices to achieve this outcome. During the course, the instructor regularly used a variety of active learning techniques in class, specifically those involving group work, such as group worksheets and small group discussions. Students worked in groups of 2-4 people for a portion of every class. The course had bi-weekly written homework assignments consisting of 2-3 substantial problems that involved analyzing and synthesizing several class concepts with the focus on communicating mathematical findings and interpreting the results. These homework assignments could be done individually or in pairs with the majority of the class opting for the latter. Finally, the course was supported by optional supplemental instruction, which consisted of weekly two-hour, peer-assisted study sessions and additional midterm review sessions. As a result, the students encountered group work in every course component.

A two-stage exam format was used on four out of five bi-weekly exams with 25 minutes allotted for the individual portion and 15 minutes for the group stage. The total exam score was calculated as the maximum of 100% individual or 80% individual+20% group; each exam was worth 3% of the final grade. The number of students taking each exam varied between 68 (Exam 1) and 57 (Exam 4). For analysis presented in the next section, we choose three questions from three different exams, categorize student errors on these questions, and track their performance from individual to group setting.

All of the groups for the exams consisted of two, three, or four students. Students were allowed to choose their own groups for each exam, so the groups were formed at the beginning of the group portion of the exam and seemed to be largely dictated by where students were

sitting that day. The classroom used was a standard lecture hall with moveable chairs and tables. The desks were not moved into pods for the exams, but the moveable chairs allowed students some flexibility as they could form groups with their neighbours on any side (including those in the row behind them). Since the groups were formed independently for each exam, their composition changed from one exam to the next. Before each exam, the instructor briefly reminded students to listen to each other, to be respectful, and to be inclusive when working together. During the group portion of the exam, the instructor circulated around the room and lightly intervened, when necessary, to ensure the groups were working effectively. Further in this discussion, we discuss changes that occurred in group compositions and the nature of instructor interventions that resulted in some group changes.

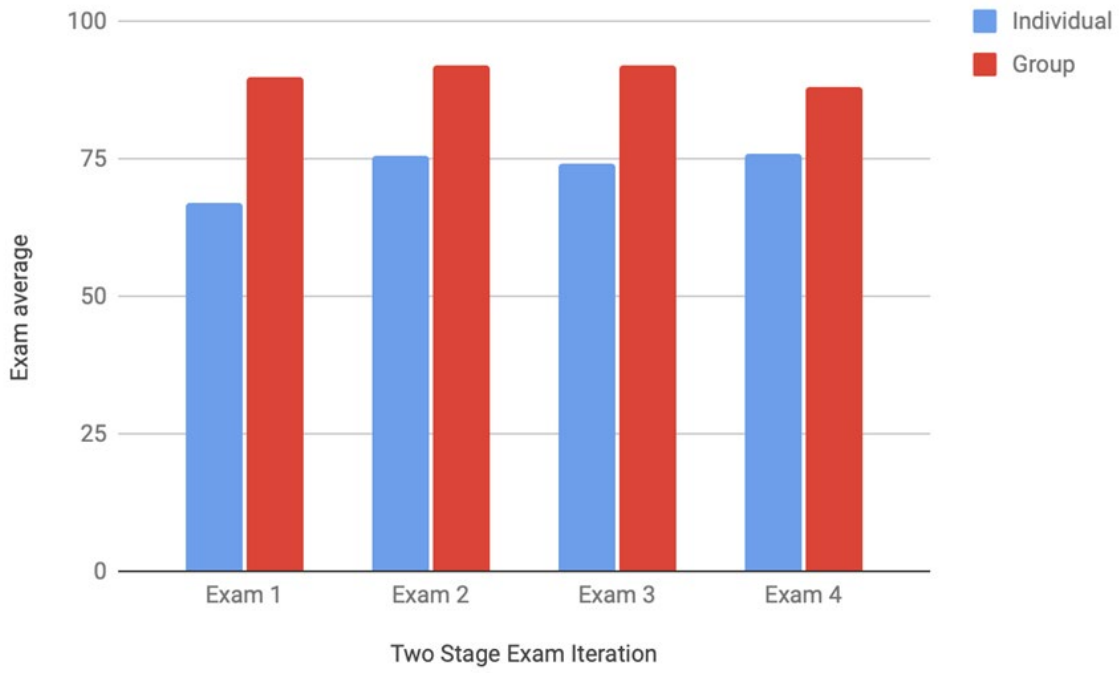
At the end of the term, we conducted student surveys to collect information on student perceptions of the two-stage exam format. Surveys were conducted anonymously in class and contained 3-point Likert scale questions as well as an open-ended question regarding group exams.

Student Success and Group Composition Results

Groups performed quite well: the lowest score on any of the group exams was 88%; Figure 1 shows exam averages for all two-stage exams. Figure 1, however, does not compare two identical sets of data since the individual and group exams were not identical. In fact, the group portion of each exam contained a new question that was either an extension of a question from the individual portion or a new conceptual question based on a more procedural one from the individual portion. As a result, the higher group exam scores actually represent higher scores on a more difficult exam.

Figure 1

Exam Averages for All Two-Stage Exams



Each individual exam contained at least one constructed response question that was repeated on the group exam. Below are the three examples we will analyze in detail:

A. Let f(x)=1/x. Find and simplify the difference quotient (f(x+h)-f(x))/h to the point when you can plug in h=0.
B. It has been determined that the number of insect species N is inversely proportional to the square of their length L (in millimetres), that is N=a/L^2 for some nonzero constant a. If a certain insect evolves to be a third of the size of its ancestor, by what factor does the number of its species change?
C. Amanda has 200 meters of fencing to build two pens: a square one for her goats and a circular one for her pigs. Set up the equation for the total combined area of the pens in terms of only one variable.

We analyzed how knowledge was constructed by the groups. Specifically, we characterized each group by whether or not it contained a member who correctly answered the question on the individual potion of the test. We further split the groups by whether or not they correctly answered the question on the group portion. Results are presented in Figure 2, where the labels are as following:

- "100% individual" means the group contained a member who correctly answered the question on the individual portion of the exam.
- "<100% individual" means none of the groups' members correctly answered the question on the individual portion of the exam.
- "100% group" means the group correctly answered the question.
- "<100% group" means the group did not correctly answer the question.

We examined whether groups scored higher, same, or lower than their top individual, which resulted in the following three cases.

Case 1: groups scoring the same as their top individual. Not surprisingly, on a question-by-question basis, most groups performed as well as their top individual(s): as seen in Figure 2, the majority of groups fell into categories of "100% individual, 100% group" or "<100% individual, <100% group" (42 out of 62 cases). In all cases, knowledge was transferred (at best) between group members to obtain the score of their top individual, but no new knowledge was gained collectively.

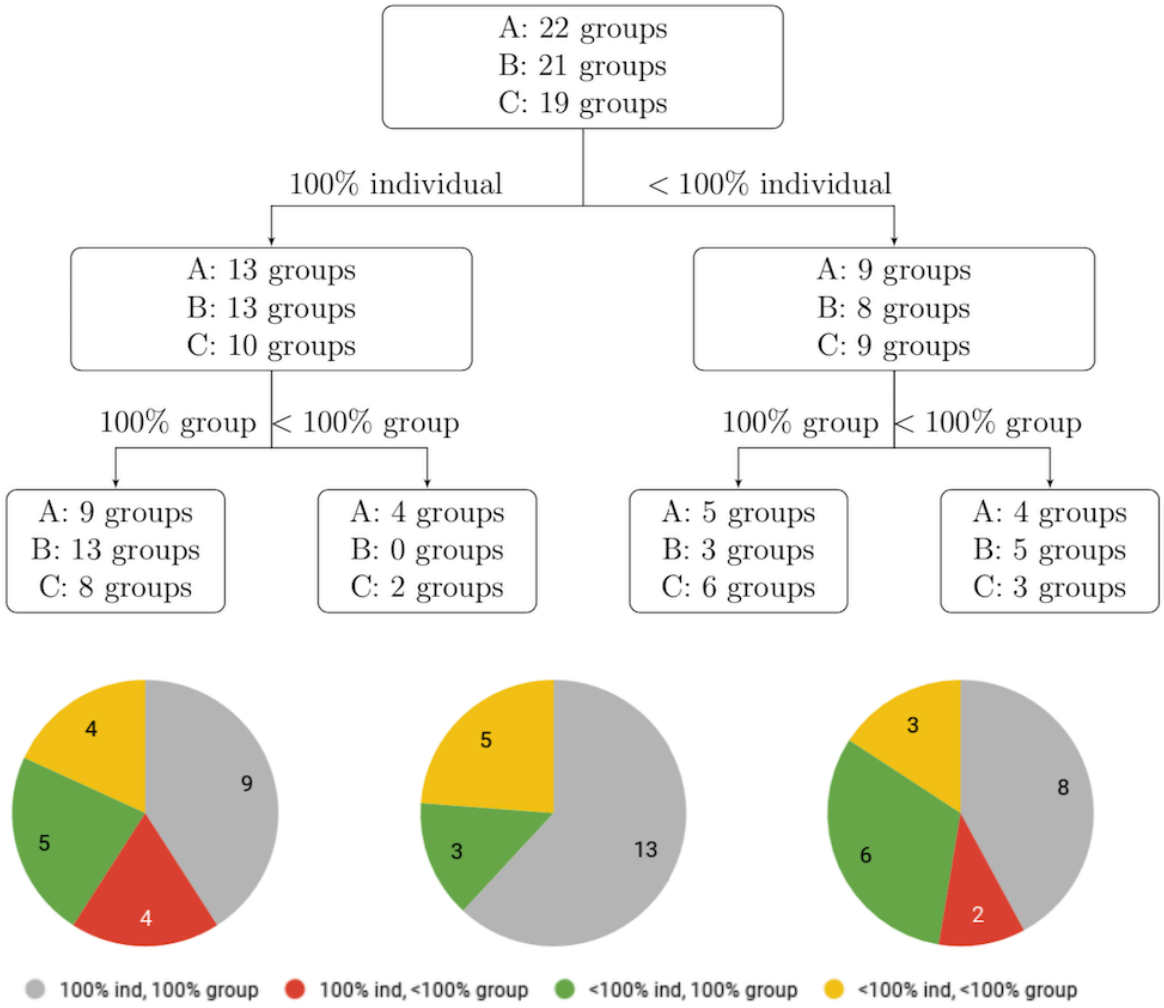
Case 2: groups scoring lower than their top individual. Some groups collectively did not perform as well as their top individual. Note that none of these types of groups appeared in Example B. Between Examples A and C, we have a total of 6 such cases out of 62 group scores. In 5 of those cases, a group had exactly one member who scored 100% on the individual test. In contrast, in 4 of these 6 cases, a group had at least one member who scored 0 individually; in the other 2 cases, one member made a major mistake on the individual portion.

Case 3: groups scoring higher than their top individual. This is the most encouraging case since these groups gained collective knowledge and managed to progress together further than any of their members did individually. This happened in 14 cases. Here we saw a variety of group compositions in terms of students' individual performance containing minor mistakes or major mistakes. In Example A, the majority of group members in each of these groups scored 0 individually: they failed to correctly use function notation to compute f(x+h). In Examples B and C, each group contained at least one member who made a minor mistake on the individual portion or did not complete the question individually. In these cases, group members managed to correct each others' mistakes and to construct a complete solution together.

Note that we did not encounter a single instance of a question where the number of groups with scores that decreased exceeded the number of groups with scores that increased. This means that while there is still a chance that a group will score lower than its top individual, there is a much higher chance that a group will match or increase the best individual score.

Figure 2

Group Composition Analysis of Three Constructed Response Questions



Changes in groups' compositions

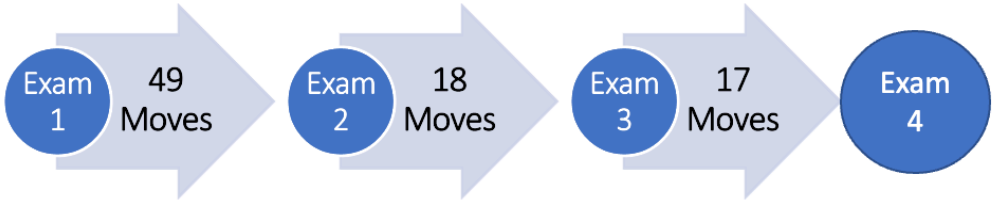
We recorded student groups throughout the semester and tracked student movement between exams. The goal was to see whether the groups stabilized or retained some fluidity in terms of their membership. Table 1 summarizes the number of students present and the number of groups formed for each exam. As the data illustrates, the average group size was between 3.1 and 3.2. In fact, 47% of all groups formed in exams throughout the term consisted of 3 people.

Table 1

Group Statistics

Exam Iteration	Number of Groups	Number of Students
Exam 1	22	68
Exam 2	21	67
Exam 3	20	62
Exam 4	19	57

To quantify the changes in the groups' compositions, we calculated the total number of student moves between each pair of consecutive exams. Specifically, we counted the number of students leaving their groups. For example, suppose for one exam Alice, Bob and Charlie are in one group. If they all appear in 3 different groups for the next exam, we count this as 3 moves. If Alice and Bob appear in the same group for the next exam, but Charlie joins another group, we count this as 1 move (Charlie's exit). We count the moves similarly for 4-person groups with one small exception. If the group of Alice, Bob, Charlie and Donna split into two couples of Alice/Bob and Charlie/Donna, we count this as 1 move since some of the group structure is preserved. This also distinguishes the case if Alice and Bob stay together as a group while Charlie and Donna join two separate groups: we count this as 2 moves. The total moves are then as follows:



Generally, students choose seats in specific areas of the classroom (see Zomorodian et al., 2012 for literature review of the seating arrangement studies), and our data supports the fact that the changes in seating location, and hence group formation, mainly occurs at the beginning of the term. Clearly, most moves happened between Exam 1 and Exam 2 with Exam 1 taking place on the 5th lecture and Exam 2 taking place two weeks later. The more surprising fact is that students did not form more permanent groups after Exam 2.

As mentioned earlier, during the group portion of the exam, the instructor circulated around the room and ensured that the groups were working effectively. Instructor interventions were minor and concerned only group dynamics, not exam content. In one case, the instructor suggested a group of 4 split into 2 groups of 2 because of surfacing conflicting personalities; in another case, the advice was exactly the opposite to bring together a weak pair of students with a strong pair for mutual benefit. In one particularly delicate case, the instructor suggested an alternative group composition for 7 people (split into two groups of 3 and 4

students) pointing out a more convenient physical arrangement. While the real reason behind the suggestion was clashing personalities in one of the existing groups, students seemed to happily form new groups and those particular groups persisted until the end of the term.

Student Feedback

We ran an in-class anonymous survey during the last week of class, which consisted of 3 multiple choice questions and an open-ended question about group exams. Out of 61 students registered in the course at that point, 56 students were present in class that day and filled in the survey.

Student responses to Likert-scale questions

For each of the three multiple choice questions we used a 3-point Likert-scale for the questions. Below are the questions and the summary of student responses.

Question 1. Did you enjoy working on the exams with your classmates?

- Liked it: 54 students
- Neutral: 2 students
- Disliked it: 0 students

Question 2. How useful did you find group exams for your learning?

- Useful: 56 students
- Neutral: 0 students
- Not useful: 0 students

Question 3. If you were allowed to pick the quiz format for the whole term, which would you pick?

- Group only: 31 students
- Individual followed by group: 24 students
- Individual only: 1 student

This overwhelmingly positive student feedback stands in contrast to findings in other literature. For example, in a recent study (Garaschuk & Cytrynbaum, 2019), in response to the same questions, only 72% of students liked working with their classmates, 72% found group exams useful for their learning (with 14% each of Neutral and Not Useful responses) and 21% preferred individual-only exams; the differences in the mean responses of the two groups to each question are statistically significant. Iannone and Simpson (2014) found that students prefer 'traditional' assessment methods in mathematical courses, but it must be noted that all the considered exam formats were individual assessments (such as closed book, open book, presentation, project) and none had a collaborative group component.

Student responses to the open-ended question

We gave students an open-ended question with a prompt “Group exams are ...". We used grounded theory to recursively code student comments and allow themes to emerge from the collected data.

The single most used word throughout was the word “helpful" as it was mentioned by 46% of the participants (26 students). In fact, 36% of all responses began with that word with or without an intensifier, such as “very", “extremely" and so on. There was one single negative comment (discussed below). To interpret the otherwise positive student feedback, we developed codes and counted the number of their occurrences. Table 2 summarizes this analysis.

Table 2

Coding Scheme and Results of Students’ Written Comments Regarding Their Opinion of Two-Stage Exams

Code Description	Number of Occurrences
C- collaboration, cooperation	26
I – improve/increase understanding, reinforce concepts	13
E- explanations, communications	9
F -immediate feedback	8
L – learning experience	8
V – variety of approaches, perspectives	6
D – productive debate	4
Other positive	17
Total	91

Many of the students' comments fell into multiple categories and were recorded with multiple codes. The category “Other positive" contains one-word comments or positive comments that occurred at most twice. Below are several examples of student comments and the corresponding codes assigned to them:

- “Beneficial to me because there are people to bounce ideas off of and someone to help you catch silly mistakes. Also talking through problems with others helps me learn better." (C, E)

- “A good learning experience, they help me understand the content much better." (L, I)
- “The group quizzes were interesting to see how we could work together. Hopeful to hear how others work through questions." (C, V)
- “Helpful to correct mistakes that may have occurred on individual right away and discuss different approaches." (F, V)
- “Help concrete understanding because you have to work together and justify your thinking to the group verbally." (I, C, E)

As can be seen from the data, students value collaboration with their peers, recognize the importance of explanations, and appreciate immediate feedback. Four people pointed out the benefits of having an opinion and be able to support it through a healthy debate: “[...] if I feel like I am doing something right but my group members think they are right we are able to debate and point out each other's mistakes, which then helps me remember in the long term what I did wrong." This highlights the creation of a learning environment during an exam in which students are comfortable arguing for their points of view, being wrong, and ultimately learning from the experience.

It is important to note themes absent from the student comments. Recall that the total exam scores were calculated as the maximum of 100% individual or 80% individual+20% group. In 82% of all four two-stage exams (255 total exams), the better overall score was the group score. In these cases, the group portion provided students with a small grade boost. This fact in itself was not mentioned by a single student; in comparison, in Garaschuk and Cytrynbaum (2019), a total of 10% of coded responses referred to the small grade boost as a positive aspect of group exams. On the flip side, in 18% of cases in this particular study, the better overall score was the individual score. In those cases, the group portion did not provide a grade boost, but would have lowered the overall score if the “maximum of" option was not given. Yet, this non-benefit of the group portion regarding student grades or time spent on the activity was not commented on by the participants.

Another notable absence is the general lack of negative comments. In the previous study with large classes (Garaschuk & Cytrynbaum, 2019), 15% of the total coded responses were negative, with most of them regarding dysfunctional group dynamics or inadequately prepared group members. In this study, there was a single negative comment: “Great to get ideas from other people on how to solve the problem; good to still have individual part because some classmates don't study and rely on you to do the questions." Due to the virtual absence of negative comments regarding group structure and comments regarding grades, we conclude that current implementation of group exams was highly successful in terms of student perceptions: students were able to work effectively in groups and valued the process of collaboration and learning more so that the grade-associated outcome of the group portion of the exam.

Discussion

We set out to analyze how group knowledge is constructed from individual knowledge during the two-stage exams in a course designed with constructive alignment focusing on group work. The results in this paper are mutually consistent and support one another. Quantitatively, in agreement with Vygotsky's view of knowledge as social process ideas, we see that learning happens in the group stage. Qualitatively, students express high satisfaction with the format. Both results are much more positive than in Garaschuk and Cytrynbaum (2019). While there are numerous factors involved, such as class size and student demographics, we attribute this gain in effectiveness to one main idea: using constructive alignment in course design with effective group work as a desired learning outcome. That is, for the biggest impact and learning benefits, group work should be ingrained in the teaching approach, formative assessments, and summative assessments. By injecting group work into all course components, students began to see it as a part of the process, and then effective group work naturally became a learning outcome in itself. As a result, we see three main consequences.

- **Students are able to work productively during collaborative exams.** As supported by exam results, in over 90% of examined cases groups did as well as or better than their top individual, while in 23% of the cases, groups outperformed their top individual resulting in meaningful learning taking place during the group stage of the exam.
- **Students form semi-permanent groups.** Groups retained some transience throughout the semester. While initial group formations underwent changes at the beginning of the term, groups did not fully stabilize over the semester. This fluidity in group structure can be attributed to students being comfortable working with slightly modified groups and not seeking full permanence in group composition.
- **Students begin to perceive exams as a learning opportunity.** Student feedback in open-ended responses revealed overwhelmingly positive attitudes towards group exams. The students' comments did not mention grades, but focused on collaboration, deeper understanding, and the importance of communication.

As the next step, it would be interesting to examine how collaborative knowledge transfers back to individuals; that is, how do individual students use what they learn as a group. To that end, one might have an individual assessment follow the group one or, alternatively, set up another individual exam following the two-stage exam. A study in the shape of the latter was performed by Gilley and Clarkston (2014) within a condensed 3-week course with multiple choice midterms. As effective group work is essential for effective performance on constructed answer questions, a full term course and open answer questions would be more appropriate for this study.

In summary, the two-stage exam format is an active learning technique that transforms a formative assessment into a formative-summative one, provides opportunities to create new knowledge, and improves students' attitudes toward assessments. However, we emphasize

that for highest effectiveness, course assessments should be aligned not only with course learning outcomes, but also with the general approach to teaching and learning facilitation.

References

Buhagiar, M. (2007). Classroom assessment within the alternative assessment paradigm: revisiting the territory, *The Curriculum Journal*, *16* (1), 39-56.

Biggs, J. B., & Tang, C. (2011). *Teaching for quality learning at university: what the student does*. Maidenhead: McGtaw-Hill.

Crouch, C. & Mazur, E. (2001). Peer Instruction: Ten years of experience and results. *American Journal of Physics*, *69* (9), 970-977.

Drouin, M. A. (2010). Group-based formative summative assessment relates to improved student performance and satisfaction. *Teaching of Psychology*, *37*(2), 114-118.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N. , Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23): 8410-8415.

Garaschuk, K. & Cytrynbaum, E. (2019). Feasibility and effectiveness of group exams in mathematics courses, *Problems, Resources, and Issues in Mathematics Undergraduate Studies*, *29* (10), 1061 -1079.

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.

Iannone, P. & Simpson, A. (2014). Students' preferences in undergraduate mathematics assessment. *Studies in Higher Education*. *40*. 1-22. 10.1080/03075079.2013.858683.

Jang, H., Lasry, N., Miller, K., & Mazur, E. (2017). Collaborative exams: Cheating? Or learning?. *American Journal of Physics*. *85*. 223-227. 10.1119/1.4974744.

Kinnear, G. (2021), Two-Stage Collaborative Exams have Little Impact on Subsequent Exam Performance in Undergraduate Mathematics, *International Journal of Research in Undergraduate Mathematics Education 7*(2), DOI:10.1007/s40753-020-00121-w

Kogan, M. & S. L. Laursen. (2014). Assessing long-term effects of inquiry-based learning: A case study from college mathematics. *Innovative Higher Education*, *39*(3): 183-199.

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 Science Education*, *11,* 392-401.

Levy, D., Svoronos, T. and Klinger, M. (2018). Two-stage examinations: Can examinations be more formative experiences?. *Active Learning in Higher Education*. 146978741880166. 10.1177/1469787418801668.

Mathematical Association of America (1999), *Assessment Practices in Undergraduate Mathematics.* Editors: Gold, B., Keith, S. Z., Marion W. A. ISBN 0-88385-161-X.

Mathematical Association of America (2006), *Supporting Assessment in Undergraduate Mathematics.* Editor: Steen, L. A. ISBN 0-88385-820-7.

National Council of Teachers of Mathematics (1995), *Assessment Standards for School Mathematics*, 102 pages, ISBN 0-87353419-0.

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

Rieger, G. W. & Heiner, C. E. (2014). Examinations That Support Collaborative Learning: The Students' Perspective. *Journal of College Science Teaching*, *43*(4), 41–47.

Rieger, G. W. & Rieger C. L. (2020), Collaborative Assessment That Supports Learning, Chapter in *Active Learning in College Science*, 821-837

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

Smith, C. (2008). Design-focused evaluation. *Assessment and Evaluation in Higher Education 33*(6), 631-645.

Springer, L., Stanne, M. E., & Donovan, S. S (1999). Effects of Small-Group Learning on Undergraduates in Science, Mathematics, Engineering, and Technology: A Meta-Analysis, *Review of Educational Research*, *69*(1), 21-51.

Vygotsky, L. S. (1978). *Mind in Society: The development of higher psychological processes*. Cambridge, MA, Harvard University Press.

Zomorodian, K., Parva, M., Ahrari, I., Tavana, S., Hemyari, C., Pakshir, K., Jafari, P., & Sahraian, A. (2021). The effect of seating preferences of the medical students on educational achievement, *Medical Education Online*, retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3355379/