An Experimental Study of the Effectiveness of Collaborative Testing in an Entry-Level Computer Programming Class

Mark G. Simkin
College of Business Administration
University of Nevada
Reno, Nevada 89557
simkin@unr.edu

ABSTRACT

In collaborative testing environments, students work together in small groups to answer examination questions. This study tested the hypothesis that group exams help student testing performance in IS classes. Quantitative and qualitative analyses of student scores on two examinations (a quiz and a formal, extensive midterm) found significantly higher group scores (compared to individual scores), and that superior group performance was particularly notable for the constructed-response portion of the midterm. Both direct observation of the group process and a survey of student perceptions *about* the group-exam process suggested that there were (1) few of the behavioral problems often attributed to group exams, (2) objective conflict resolution, and (3) favorable student perceptions of the process itself. This paper also provides several caveats that should be considered when interpreting these findings and suggests several avenues for future research.

Keywords: Collaborative testing, group testing, participative learning, computer programming, information systems education

1. INTRODUCTION

The term *collaborative testing* refers to a number of pedagogical tools that instructors can use to assess student understanding of course materials. Common to all of them is the requirement that students provide one common set of exam answers, presumably representing the collective wisdom of the group's members. Common formats for collaborative test questions include multiple-choice, true/false, or fill-in-the-blank, although such "constructed response questions" as short-answer or even essay questions are possible. Constructed-response questions are probably the most difficult to administer on a group basis and therefore not commonly used.

Proponents of group exams argue that such assessments better reflect the reality of the work place—i.e., the fact that such professionals as auditors, IT personnel, and advertising analysts typically work in teams, must create a single set of deliverables, and are judged on the basis of group, rather than individual, performance. If "teamwork" is so common to such diverse occupations and work settings, why should academic practice be any different?

At the university level, experiments in group exams have been conducted in a variety of academic settings, including classes in accounting, marketing, mathematics, and tax (see, respectively, Cottell and Millis, 1993; McIntyre, et. al., 1999; Berry and Nyman, 2002; and Hite, 1996). Less experimentation appears in the field of information systems—a surprising absence in light of the team orientation involved in so much of the industry's activities. Accordingly, the author sought to address this deficiency with some experimental trials in an entry-level computer programming class.

The next section of this paper discusses collaborative testing in greater detail and the rationale and concerns for the process. The third section of this paper describes a set of experiments and a survey that the author conducted to assess the effectiveness of group exams in an IS class setting. That section of the paper also provides the results of these experiments and describes the insights the author gained from this experiment. The fourth section of the paper provides some additional observations about the group-exam process and also describes some caveats that limit the findings described here or that might hamper the use of group exams in other classes. The final portion of this paper provides a summary of this work and conclusions about it.

2. COOPERATIVE LEARNING AND GROUP EXAMS

Much of the rationale for group exams rests in the more-general arguments for *cooperative learning*. Although teachers have used this term to encompass everything from group work on business cases to teamwork on intramural

sports activities, Johnson, et. al. (1991) suggest more rigorous criteria, which include: (1) *positive dependence*, in which positive rewards occur only when the group as a whole, rather than the individuals within it, succeed, (2) *face-to-face interaction*, (3) social norms of behavior requiring *individual accountability for group performance*, (4) *the need for collaborative skills* such as oral communication skills (including the ability to present cogent logic in persuasive arguments), leadership qualities, and organizational skills, and (5) *the ability to work with, and help, others* in order to accomplish group goals. The term "group" is not explicitly defined among these criteria, thereby allowing "cooperative learning" to apply to student sets as small as two individuals or as large as an entire class. In practice, groups of from three to six students appear to be the most common.

2.1 The Rationale for Cooperative Learning Pedagogy

One key argument favoring cooperative learning is *motivation* (Graham and Graham, 1997; Hite, 1996; Astin, 1993). This argument rests on the idea that students who are forced to perform tasks in group settings are also positively motivated to share what they know, listen to the ideas of others, critically examine evidence, and actively search for objective methods or knowledge sources for choosing among several alternative approaches or solutions. King (1992) also notes that working in groups motivates the participants to find impartial ways of resolving conflicts—an important social skill in almost any professional setting. Finally, Astin (1993) suggests that students may be motivated to study harder and therefore learn more—for example, because they will be answerable to their peers for wrong answers.

A second characteristic of cooperative learning is the shift in learning focus from a "fact-driven" venue of lectures delivered by an instructor to a "process-driven" environment in which students interact with one another in peer groups. Within the context of mathematics subjects, for example, Crawford, et. al., (1998) found that students often perceive "learning" as a series of memorization tasks and rote rehearsal of calculations rather than as a process of discovery. In contrast, many university instructors now feel that, in order to produce graduates with the skills desired by future employers, group work and process-oriented learning skills may be even more important than any exacting grasp of standard curriculum materials (Berry and Nyman, 2002; Mitri, 2003). These considerations lead such instructors to believe that cooperative learning exercises may be better instructional formats than conventional lecture presentations.

A third dimension of group work is that it often results in heavy doses of *student teaching* as well as *student learning*. Both occur, for example, when one or more other members of a group are unfamiliar with particular theories, methodologies, or solution techniques, but are "brought up to speed" by others more knowledgeable in the subject area germane to the task at hand. A concomitant thought by Michaelson, et. al. (1985) is that because group work requires active, rather than passive participation, students who enjoy peer attention and positive feedback are motivated to work harder at group tasks in order to get these rewards.

Perhaps the most important question to ask about collaborate learning in general, and group testing in particular, is "does such pedagogy result in more student learning than traditional lecture formats?" Cortright, et al. (2003), argue that it does—for example, that student retention of course content increases with the use of collaborative testing. Similarly, when experimenting with group exams, Hite (1996) found that the final examination scores for over 250 students in her income tax class were "significantly higher" for those students who had participated earlier in group exams than those who had not. Stearns (1996) found similar results for her classes in a communications course, and McIntyre, et. al., (1999) observed the same results for the students in their marketing classes.

Finally, experimenters have noted a number of miscellaneous benefits from group exams. Among them are student perceptions that (1) group exams are challenging and good learning experiences, (2) courses using group exams more accurately reflect the professional evaluation norms used in the work place, and (3) the teacher is an outstanding instructor (Hite, 1996; Stearns, 1996). Graham and Graham (1997) also suggest that group exams reduce student fears about the testing process in general and encourage them to view an examination as a learning experience rather than as a chore or punishment. McIntyre, et al., (1999) also argue that group exams help students develop higher levels of cognitive, interpersonal, and communication skills.

2.2 Some Concerns

Despite the many positive qualities that are attributed to collaborative testing, some college-level instructors continue to harbor doubts. One concern is the potential lack of individual accountability in the group exam process. In discussing the use of small group assessment methods in mathematics classes, for example, Berry and Nyman (2002, p. 641) note that "collaboration by students is often seen as 'cheating' in the traditional mathematics classroom."

Graham and Graham (1996) also note that group examinations require faculty to relinquish their traditional role as "authoritative lecturer" and assume (at least temporarily) a more passive role as "facilitator." These researchers also note that the higher examination scores resulting from group examinations makes instructors wonder if "they had been too easy on the students."

Among the other concerns expressed by instructors contemplating the use of collaborative testing activities is that such cooperation encourages "hitchhikers" (who sit quietly and merely ride the coattails of their peers), "workhorses" (who do most of the thinking for the group), or "emperors" (who impose their wills on others regardless of the accuracy of their answers). Empirically, there has been little report of such developments. Cottell and Mills (1994) state that, to the contrary, the majority of the evidence collected so far about such cooperative learning activities has resulted in increased levels of achievement, and widespread load sharing of group responsibilities. Similarly, Hite (1996)

found that the superior performance of the students in her experimental group (who had participated in group examinations) was uniformly distributed among students with low, average, and high grade point averages.

2.3 Collaborative Testing at the College Level

University instructors can choose from a variety of test and administrative formats within the general area of collaborative testing. One set of variations lies in the test format itself (i.e., multiple-choice or short answer). Past applications of group testing have favored multiple-choice formats because these are easiest to grade, enable instructors to ask a wide, "shot-gun" selection of questions, and facilitate discussions because answers are limited to a few choices. Where large numbers of students are involved, machine-grading is particularly advantageous because repeated-testing of the type discussed here increases the grading burden by 20 to 50 percent (for group sizes of five or two, respectively).

A priori, nothing *requires* instructors to use MC formats, and this factor was also of particular interest to the author. Alternate formats such as short answer or essay questions are possible, but it is not known how useful or effective group testing is in such venues. One obvious concern is the lack of time required by in-class testing, and therefore the inability for groups to frame answers for less-structured questions in reasonable amounts of time. Another is the extra grading burden created by constructed response questions.

Instructors also have a range of options in how to administer the tests. One choice is to give only one test, requiring all students to work in small teams and awarding a single score to the group effort. An alternate scheme is to administer the same test twice—once as individuals, and again in groups—and to then take a weighted average of the two scores. Equal weightings are one choice, but alternate weightings are also possible—for example, the weights of 80 percent (individual) and 20 percent (group) used by Rao, et al., (2002).

3. TWO EXPERIMENTS

The benefits that appear to accrue from group exams—for example, that students perform better in groups, learn more, and perhaps enjoy many of the fringe benefits described earlier (e.g., that group work motivates the participants to find impartial ways of resolving conflicts)—encouraged the author to experiment with them in an introductory procedural language (VB.NET) programming class at his university. In particular, the author was interested in testing the following hypotheses:

H_1: Students perform equally well as individuals as they do in groups on multiple choice tests.

H_2: Students perform equally well as individuals as they do in groups on constructed response tests.

Of further interest to the author were student attitudes *about* such exams. How do students feel about group tests? Do they feel they learn anything? If so, what do they feel they learn? And finally, the author was interested in the mechanics of the group process itself. How does it work and what additional benefits and drawbacks accrue from such activities?

Without prior experience, the author favored the approach of giving the same test twice and then equally weighting the results to determine a final exam grade for each student. Reasons for this preference included the ideas that: (1) it incorporates individual accountability as well as group performance in the test results, (2) the scores on the first ("individual") tests provide a benchmark against which to measure group performance, (3) it forces students to form individual opinions about challenging test questions prior to meeting in groups, and (4) it encourage discussions when test takers do not agree upon a correct answer. Accordingly, the author performed two experiments to test the merits of group exams. These are described below.

3.1 Experiment 1

The author had no experience in administering group exams, and was therefore reluctant to devote too much time to such an uncharted activity. Of special concern was using valuable class time for a collaborative test that might prove disastrous. Accordingly, he began with a limited trial in which all the students in the class took a quiz twice—once as individuals and once in groups of size two or three.

Students first answered the test individually. Then, after the instructor collected their answers, students took the same test (using the same test booklet) in groups. For the group portion of the test, students were free to choose their own teams. The resulting partnerships resulted in ten groups—five groups of size 2 and five groups of size 3. Students could collaborate with their own team members, but not with the members of other teams. In answering test questions, they were permitted to use their class notes, past homework, and class handouts to help them, but not computers.

The quiz itself consisted of 15 multiple-choice questions that tested their knowledge of Visual Basic programming techniques. There was no penalty for guessing. Figure 1 provides a sample question. The students answered the quiz questions using machine-gradable ("scantron") forms. There were separate forms from each individual in the first round of the quiz, and a single scantron form for each *group* in the second round of the quiz. Both rounds of the test were administered back to back in the same class period, and the whole process took no more than 30 minutes.

1. Assume that *x*, *y*, and *temp* are Integer variables. Which of the following lines of code swaps the value of x and y?				
A. **x = y** **y = x**	B. **x = temp** **x = y** **y = temp**	C. **temp = x** **x = y** **y = temp**	D. **x = y** **temp = x** **y = temp**	E. None of these

Figure 1. Example question from a quiz on VB.Net

To keep the time devoted to this task manageable, students were only given fifteen minutes to complete their tests as individuals, and another fifteen minutes to take their test in

groups. Instructors with experience in group testing recommend that such tasks be done at the end of class, thereby enabling those who finish early to hand in their assignments and leave class. The author followed this advice, and found that everyone finished within the time frame allotted, leading him to conclude that "insufficient time" was not a performance factor here.

Figure 2 reports mean scores and test statistics for results of the quiz. These results confirm what might be intuitively expected—by any statistical measure, students performed better in groups. In this sample, for example, 92 percent of the students did as well as, or better than, they had as individuals when they took the test in groups, while only 8 percent (2 individuals) did worse in groups. Using a simple difference-of-means test, the resulting t-statistic was "4.9"—i.e., a value suggesting that the group performance was significantly better than individual performance.

Because the sample values represent *pairs* of scores, the author also performed a (superior) matched-scores test on the quiz data (Berenson, et. al., 2002). This resulted in a t-statistic of 4.5—again confirming the statistical significance of the individual versus group quiz scores. Finally, to overcome the requirement that the underlying distribution of paired differences be normally distributed, the author performed a Wilcoxon rank sum test for the difference in medians. This resulted in a t-statistic of 4.5, again confirming the significant statistical differences between student performances in the two testing venues. These findings are also consistent with earlier experiments—for example, Rao, et. al., (2003) who found that test performance was significantly higher when students worked in groups compared to when they worked individually.

Statistic (n = 25)	**Performance Measures**
Mean (individual)	10.9 (1.8)
Mean (group)	12.7 (0.9)
t-statistic for simple difference of means test	4.9*
t-statistic for matched pairs test	4.5*
t-statistic for Wilcoxon rank sum test	4.5*

* t-statistics significant at the .001 (standard deviations in parentheses)

Figure 2. Performance on Quiz 1

3.2 A Survey of Student Attitudes about Group Exams

On the first class day *following* the administration of the quiz, the author also asked the students in this class to complete a survey that asked questions about their group-test experiences. Appendix A contains the survey instrument. Students answered these questions as anonymous individuals—*not* in teams. Responding was voluntary, but all students chose to complete the survey. Indeed, most seemed happy to respond because the final question asked whether they wanted more group exams in the future, and *all of them enthusiastically did so.*

To allow students to answer in any manner they wished, the survey instrument asked open-ended questions. The analysis that follows is an interpretation of the responses, but caution should be used here—a few of the "yes-but" answers could be classified in a number of ways.

Question 2 of the survey asked students whether all members of the group contributed to the discussions, or did one or more individuals simply defer to others. This question speaks directly to the concern that some instructors have regarding "hitchhiking," "workhorse," or "emperor" behavior. In their answers, however, all but two students (92 percent) reported that all members participated in the discussions. For the two exceptions, one student simply answered "no," while another student wrote "every member contributed to the discussion, but some dominated in some questions." Because students answered this survey anonymously, the author does not know whether these two students were in the same group.

Question 3 of the survey asked students how they resolved conflicts when members did not agree. In almost every case, the respondents indicated that they referred to their class notes, homework, and handouts for this task, and *also discussed the applicability of these reference materials in their deliberations.* Where differences remained, most students indicated that they voted and went with the majority decision. Typical responses were: (1) "we discussed why we chose what we did and then picked the best answer" and (2) "which one had the best evidence to support their answer when in conflict." One interesting response for this question was "The person who was right proved the others wrong."

Question 4 of the survey asked "Did you learn anything from other members of your group? If so, what did you learn?" Most students (72 percent) answered "yes" to this question, and mentioned a particular Visual Basic programming technique or concept. An example is "I learned that there is no such thing as a Triple data type." But two answers surprised the author. One student wrote "I learned how to discuss the question among my group members and come to a logical decision." A second student wrote "Yes, good discussion skills and ways to work out problems differently." It had never occurred to the author that interactive discussion skills would be the most important thing to emerge from this experiment, but these answers lend credibility to the claim that collaborative testing helps students develop skills in cognition, communication, and conflict resolution.

Questions 5 and 6 of the survey asked whether the author should give more group tests in the future, and if so, what percentage weights should be assigned to individual versus group performance. In response to question 5, 88 percent of the respondents answered "yes" while three students said the equivalent of "maybe." One student wrote, in part, "it was interesting, but in some respects I like individual tests—the whole group think concept can mess things up." A second student wrote "I'm not sure, because sometimes [the group] test will be worse." The third student wrote "Only if you take the highest score...because students should never be penalized for an experiment." It is easy to guess that the first two responses came from the two students who did *worse* on

the group test compared to their individual scores, but the author does not know if this is true.

Finally, question 6 of the survey asked students to provide relative weights to give to individual versus group test scores, in percentages, if they wanted more group exams in the future. All students answered this question. The average grade weights were "45%" for individual scores and "55%" for group scores. These averages are surprising because there were no restrictions on what students could indicate for these weights, students knew in advance that over 90 percent of the class had performed better on the "group" exam compared to the individual exam, and *the optimal grade weights are 0 percent for individuals and 100 percent for groups.*

3.3 Experiment 2

Encouraged by the positive results of the first experiment, the author also used a group-exam format for his midterm—a test that counted as 25% towards the final course grade. Because the exam consisted of two parts, this required him to devote two entire class periods to administering the test. In the first class period, he gave a 40-question, multiple-choice test, and in the second class period, he gave a 60-point in-class programming (constructed response) test. This latter part required students to create lines of code in VB.Net to perform stated tasks. Figure 3 provides a sample coding question.

> A customer number must be exactly 7 characters in length and in the form XXX-NNN, where X is any character, the fourth character is a dash, and NNN must be numeric. Write a Function named TestIt with one string argument to validate this customer number. The function should return a Boolean value set to True if the customer number passes these tests and False otherwise. (6 points)

Figure 3. A sample coding question for a midterm examination in VB.Net.

The administrative format for both parts of the test was the same. In each class period, the students spent the first portion of the time answering the questions as individuals. In the second portion of the class, students answered the same questions in (the same) groups of size two or three. For Part I of the exam, students provided individual scantron forms, but were permitted to keep their test booklets. They then worked together and turned in a single scantron form for the entire group. Again, there was no penalty for guessing on this part of the examination.

Perhaps the most interesting part of the experiment was Part II of the midterm exam, which consisted of the coding questions just described. To the author's knowledge, this latter experiment—the administration of a group exam for constructed-response questions—is particularly novel. For this part of the examination, students first wrote short-answer questions as individuals. Then, after collecting these booklets, the instructor distributed an identical test booklet to all members of the class, but asked each *group* to provide a single set of answers.

Figure 4 provides the results of, and statistical tests for, both parts of the examination. Again, by any statistical measure, most students did better in groups than they did as individuals. For example, group scores averaged almost 6 points higher (out of 40) for Part 1 (multiple choice questions), and over 14 points higher for Part 2 (coding questions). These differences were statistically significant (at the .001 level) using a simple difference-of-means test as well as a matched-pairs test.

For this exam, it is also instructive to examine individual performance. For example, it was *not* true that everyone did better in groups than as individuals. However, for Part 1, only two students had lower individual scores than groups—both by 2 points (see the row entitled "Maximum losses" in the figure). For part 2, this same (2-point) differential was found for three students—i.e., there were three students who did better individually than in groups. Thus, over 90 percent of the class did better on *either* part of the exam working in groups than working individually.

Statistic (n = 25 for both parts of the examination)	Performance on Part I—40 points (Multiple Choice)	Performance on Part II—60 points (Coding)
Means (individuals)	25.4 (5.0)	31.9 (10.9)
Means (groups)	31.5 (3.7)	46.5 (3.2)
Average gain in groups	6.1 points	14.6 points
Maximum losses	2 points	2 points
Maximum gains	18 points	35 points
t-statistic for simple difference of means test	6.1*	6.7*
t-statistic for matched pairs test	6.1*	6.3*

*significant at the .001 level or better (standard deviations in parentheses)

Figure 4. Test statistics for the individual and group scores on the multiple-choice and coding portion of a midterm examination

What about individual *gains* when working in groups? Figure 4 indicates that the maximum positive differences were 18 points for Part 1 and 35 points for Part 2. In other words, some students did spectacularly better in groups than they did as individuals, strongly suggesting that they learned something in this process and explaining why they might *really like* group exams!

4. SOME ADDITIONAL OBSERVATIONS AND CAVEATS

Figure 5 graphically illustrates aggregate student performances, scaled to percent correct scores, for the quiz and two parts of the midterm administered in the subject class. The positive evidence found here suggests that group exams do indeed improve test performance. Compared to individual norms, student scores increased dramatically, and (as illustrated in Figures 2 and 4) statistically significantly, in every portion of every test.

This finding is especially noteworthy for the constructed response questions in Part II of the midterm exam. The numerical improvements (of group over individual test performance) for all three examination parts were approximately 12% (quiz), 18% (midterm Part I), and 24% (midterm, Part II). This finding suggests that, while the group process appears to increase student understanding when students take either MC or constructed-response tests, the greatest amount of learning may happen in unstructured venues. To the best of the author's knowledge, this testing dimension has not been examined previously and is worthy of further study.

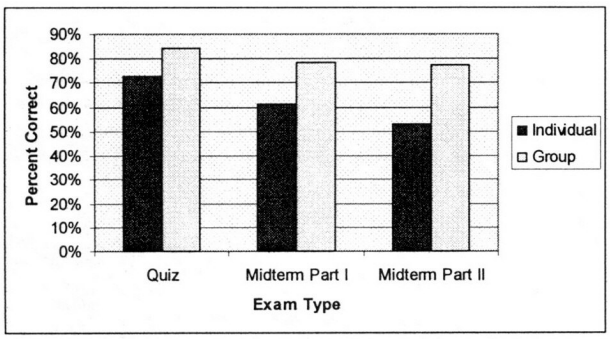

Figure 5. Mean student test scores, scaled to percent correct, for three measures of individual versus group test performance.

4.1 Some Additional Observations about the Collaborative Testing Process

What the statistical evidence provided so far does *not* convey is an idea of the process itself. In the author's experience, when the students in this class began to discuss their answers, he noticed that students became remarkably engaged in the task at hand and that the classroom filled with the "sounds of learning." The discussions were energetic and focused, and it was difficult for the author to imagine an alternate classroom format that could result in such concentrated attention on the material.

A second observation that is not evident from the statistics above was how happy students were when they found that they agreed on the answer to a given question. In effect, this agreement appeared to reinforce student understanding of a specific concept. A typical conversation began with one student's saying "Oh, good… I picked "A" but choice "C" also tempted me" and ended with a second student's explaining why that would have been a bad choice.

A third observation is that, for the MC portion to the test, most students "managed by exception" in their deliberations over test questions. This is to say that most students spent little time on questions that they felt they understood and had answered correctly, and focused most of their time and energies on questions that challenged their understanding or questions on which they disagreed. Although this only seems logical, this observation seems particularly important to the author. One reason for this feeling is a sense of efficiency. Unlike lecture formats that force all attendees to listen to material which might be self evident, the group-exam process allows students to select their own study materials, and therefore focus on matters of individual interest and challenge. In the author's opinion, it is difficult to imagine a more efficient use of class time.

A final observation is the idea that group exams allow—even encourage—instructors to ask more challenging questions on their examinations. Indeed, the author found that he could ask more challenging questions with some confidence that the resulting tests would not only better distinguish student understanding, but would also provide a higher level of learning in the group portions of the tests. What this means, is that, in effect group exams not only may result in more learning for students, it may also enable instructors to teach more, albeit indirectly.

4.2 Caveats

Finally, some caveats are also in order. One reservation is the fact that the test experiments described above were conducted in a class composed mostly of IS majors or minors in their junior or senior year. Thus, it is easy to guess that these students had a lot in common and probably knew each other from prior classes. It is therefore also possible that the potential consistencies in their backgrounds, temperaments, prior friendships, or common major helped them work together, and that these factors helped produce the positive outcomes observed here. The extent to which these factors overcame inherent deficiencies in the process itself is unknown.

A second caveat is the fact that the statistics derived for these experiments are necessarily small sample values. In the information systems arena, where industry ramp-up problems are notorious, it is easy to argue that group exams can *only* work in small classes. A counterargument is the fact that the findings of the present study reflect the findings of Hite (1996) and others who performed similar experiments with as many as 250 students. In the author's opinion, therefore, "class size" does not seem a particularly relevant barrier to collaborative testing procedures.

A third consideration is the range of resources available to the students taking group tests. Some instructors do not permit their students to use any test aides, while others, such as this author, permitted his students to use their textbooks, class notes, class handouts, and homework assignments as references (but not their computers). For the examinations described here, it is important to note that there were some questions in the multiple choice portion of the exams, and most of the constructed-response portion of the midterm examination, that were probably best answered by testing them on a computer. Thus, it is possible to argue that the range of resources available to the test takers have a bearing on how well students are able to generate viable test answers or solutions, and therefore how much they do or do not learn in group interactions.

A fourth reservation concerns what we mean by "student learning." For the most part, past experimenters have argued that statistically-significant improvements in test scores are a necessary and sufficient condition to demonstrate the

effectiveness of collaborative test procedures. But are they? It is reasonable to ask "did students really learn during the collaborative portions of their tests, or did they merely acknowledge their mistakes?" The author's sense is that "learning" took place, but repeated assessments of student understanding were not conducted to prove it. This is a fertile avenue for future research.

A fifth observation is that a number of "testing mechanics" *do* appear to be important when administering group exams. One concern here is the permissible size of the test groups themselves. In the experiments described here, the author purposely limited this to "three" in order to minimize the hitchhiking behavior feared in such collaborative venues. But it is easy to speculate that larger group sizes increases the potential for this. A second concern here is the fact that those collaborative exams requiring both individual and group efforts take almost twice as much time to administer as exams that do not. Finally, examinations requiring constructed responses that cannot be graded by machine necessarily impose an extra grading burden on the instructor. Both of these latter considerations force instructors to weigh the benefits of group exams against these costs.

A final, mechanical concern is that group exams require enough classroom space for students to meet without bothering others. In the conventional, often filled, university classrooms in the U.S., this may require groups to work in close proximity to one another. The extent to which individuals from one group overheard the conversations of others, or used this information in their own examinations, is not known. It did not seem to be a problem in the present experiments—students *seemed* too immersed in their own group conversations to listen to others. But the author did not control for "overhearing" and acknowledges this concern—especially in cramped classrooms.

5. SUMMARY AND CONCLUSIONS

The term *collaborative testing* refers to a number of pedagogical tools that instructors can use to assess student understanding of course materials. Common to all of them is the requirement that students provide one common set of exam answers, and the presumption that such testing fairly and accurately represents the collective understanding of course materials by the group's members. Arguments favoring group exams include the belief that such testing motivates students to study, encourages students to critically examine evidence, facilitates student discussions and teaching, and forces students to adopt socially-acceptable ways of resolving conflicts. Concerns about group testing include a perceived loss of accountability for individual achievement, the fear that group exams encourage hitchhiking or "emperor behavior," and the potential to lose control over classroom activities.

To test the hypothesis that group exams foster student understanding of course materials and do enjoy the advantages identified above, the author conducted two experiments. The first—a quiz consisting of 15 multiple-choice questions—resulted in significantly higher group scores (compared to individual scores). A post-quiz survey also revealed: (1) very few of the behavioral problems often attributed to group exams, (2) most students used professional, objective means of resolving conflicts (and a few students indicated that this problem solving dimension was the most important thing they learned from the experiment), and (3) all students viewed group exams as a positive experience, and one that they would welcome again for additional examinations.

Because the students taking the quiz all indicated their preference for group examinations, the author also used a group test for the first and second parts of his midterm examination. Of special note is the fact that the second part of the exam required constructed response questions. However, the author observed the same positive statistical results for both parts of this experimental test as he did for the quiz. Group scores were significantly higher than individual scores, some remarkably so.

The author also found several additional positive factors that might encourage university faculty to experiment with collaborative exam formats. Among them were: (1) a remarkable amount of student engagement in the group exam process, (2) reinforcement when students agreed on the same answers, (3) a tendency to manage by exception in the discussion process, and (4) a finding that group exams encourage faculty to ask more challenging questions than they might otherwise, and potentially increasing the amount of learning in the classroom.

6. BIBLIOGRAPHY AND REFERENCES

Astin, A. W. (1993), "What Matters in College? Implications for Cooperative Learning of a New National Study" *Cooperative Learning and College Teaching* vol. 3, no. 3, pp. 2-8.

Berry, John and Melvin A. Nyman, (2002), "Small-Group Assessment Methods in Mathematics" *International Journal of Mathematical Education in Science and Technology* vol. 33, no. 5, pp. 641-9.

Berenson, Mark L., David M. Levine, and Timothy C. Krehbiel, (2002), *Basic Business Statistics* (Upper Saddle River, New Jersey: Prentice Hall,).

Cortright, R. N., Collins, H. L. Rodenbaugh, D. W, and DiCarlo, S. E. (2003), "Student Retention of Course Content is Improved by Collaborative-Group Testing" *Advances in Physiology Education* vol. 27, no.3, September, pp. 102-108.

Cottell, P. and B. Millis (1993), "Cooperative Learning Structures in the Instruction of Accounting" *Issues in Accounting Education* vol. 8, Spring, pp. 40-59.

Crawford, Kathryn, Sue Gordon, Jackie Nicholas, and Michael Prosser (1998), "Qualitatively Different Experiences in Learning Mathematics at University" *Learning and Instruction* vol. 8, no. 5, pp. 455-68.

Graham, Reginald A. and Beverly L Graham (1997), "Cooperative Learning: The Benefits of Participatory Examinations in Principles of Marketing Classes" *Journal of Education for Business* vol. 72, no. 3, January/February, pp. 149-152.

Hite, Peggy A. (1996), "An Experimental Study of the Effectiveness of Group Exams in an Individual Income Tax Class" *Issues in Accounting Education* vol. 11, no. 1, Spring, pp. 61-76.

Johnson, D.W., R.T. Johnson, and K. A. Smith, (1991), *Active Learning: Cooperating in the College Classroom* (Edina, MN: Interaction Book Company,).

King, A. (1992), "Promoting Active Learning and Collaborative Learning in Business Administration Classes" in *Interactive Learning and Technology: Reaching for Excellence in Business Education* (Arthur Anderson Foundation), pp. 158-73.

Mallenby D.W and M. L. Mallenby (2003), "Use of Brief Collaborative Quizzes on New Quantitative Lecture Material" *Decision Sciences Journal of Innovative Education* vol. 1, no. 1, February, pp. 141-144

McIntyre, Faye S., James L. Thomas, Jr., and Russel W. Jones (1999), "Cooperative Testing in the Marketing Classroom" *Marketing Education Review,* Summer, pp. 45-51.

Michaelson, L. K, W. E. Watson, and C. B. Shrader (1985), "Informative Testing—A Practical Approach for Tutoring with Groups" *Organizational Behavior Teaching Review* vol. 9, no. 4, pp. 18-33.

Mitri, Michel (2003), "Applying Tacit Knowledge Management Techniques for Performance Assessment" *Computers and Education* vol. 41, no. 2, September, pp. 173-90.

O'Neil, Harold (2003), "Issues in the Computer-Based Assessment of Collaborative Problem Solving" Assessment *in Education* vol. 10, no. 3, November, pp. 361-74.

Quarstein, Vernon and Polly A. Peterson (2001), "Assessment of Cooperative Learning: A Goal-Criterion Approach" *Innovative Higher Education* vol. 26, no. 1, Fall, pp. 59-77.

Rao, S. P., Collins, H. L., and DiCarlo, S. E., (2002), "Collaborative Testing Enhances Student Learning" *Advances in Physiology Education* vol. 26, no. 1, March, pp. 37-41.

Stearns, Susan A. (1996), "Collaborative Exams as Learning Tools" *College Teaching* vol. 44, no. 3, Summer, pp. 111-112.

Williams, N. (2004), "Daily Shared Quizzes," *The National Teaching & Learning Forum*, vol. 13, no. 3, Mar., pp. 6-8. (Also at http://www.ntlf.com/restricted/v13n3/quizzes.htm)

AUTHOR BIOGRAPHY

Mark G. Simkin is a professor of Computer Information Systems at the University of Nevada. He earned his MBA and Ph.D. degrees from the University of California, Berkeley. His research in internet law, end-user computing, computer education, and computer crime appears in over 100 academic journal articles and 13 books, including *Decision Sciences, The Journal of Accountancy, Communications of the ACM, and Communications of the Association for Information Systems.* His most recent book is *Core Concepts of Accounting Information Systems* (New York: John Wiley and Sons, 2005).

APPENDIX A: SURVEY INSTRUMENT

The questions below refer to the "group test" you took in class last week. Please answer all questions directly on this form.

1. In the group portion of your test did you change any right answers to wrong ones? If so, approximately how many?

2. Did all members of your group contribute to the discussions, or did one or more individuals simply defer to the others?

3. How did you resolve conflicts when members of your group did not agree on the answer to a specific question?

4. Did you learn anything from other members of your group? If so, what did you learn?

5. Do you think we should have more group tests in the future? If so, why? If not, why not?

6. If you think we should have more group exams in the future, what relative weights (in percentages) should be given to individual versus group test scores? Please indicate your preferences here. (Note: they must add to 100%)

   Individual Score _____ % Group Score _____ %