LEARNING BY EXAMS: THE IMPACT OF TWO-STAGE COOPERATIVE TESTS*

Decades of research have documented the positive impacts of cooperative learning on student success: increased learning, retention through graduation, improved critical thinking, and intrinsic motivation. One cooperative teaching technique, however, has received relatively little attention. In the two-stage cooperative, group, or "pyramid" exam students first take an exam individually—as in traditional testing—and then take the same exam together with their learning group, with the exam grade being a weighted combination of their individual and group scores. This approach uses the exam itself not only for evaluation, but also as a learning tool. Although some researchers have studied group exams, they have not isolated the impact of group tests from individual achievement—an important omission. Using data from a mass lecture introductory sociology course, I found that holding individual achievement constant the group exam process significantly increased learning both for students who knew the material initially and for those who did not. This suggests that cooperative exams not only enhance learning but also allow for the process and form of testing to become more closely linked to the process and form of teaching and learning.

JOHN F. ZIPP
University of Akron

DECADES OF RESEARCH have documented the positive impacts of cooperative learning on student success: increased learning, retention through graduation, improved critical thinking, and intrinsic motivation (Astin 1993; Johnson, Johnson, and Smith 1991; McKeachie 2002; Millis and Cottell 1998; Quarstein and Peterson 2001; Slavin 1990). No less of an authority than McKeachie (2002) concludes: "The best answer to the question: What is the most effective method of teaching? is that it depends on the goal, the student, the content, and the teacher. But the next best answer may be: Students teaching other students" (p. 188).

Scholars offer various explanations for why cooperation improves learning. Some locate its philosophical basis in John Dewey's emphasis on experiential learning (Davidson 1990). Johnson and his colleagues (Johnson and Johnson 1989; Johnson et al. 1991) draw on the social psychological theories of Kurt Lewin and Morton Deutsch, especially with regard to the positive role of "social interdependence": cooperation tends to generate (and be generated by) mutual help, trust, and the sharing of resources. Astin (1993) contends that cooperative learning works because students become more motivated and involved when they know that they will be judged by their peers. McKeachie (2002) also points to the effects on student motivation and offers several additional reasons for the effectiveness of peer learning: cognitively it allows students to engage in "elaboration"—putting things in their own words; it makes students

*The author would like to thank the Department of Sociology at the University of Akron for financial support for this project, Ron Severtis and Heather Boughton for research assistance, and Jan Yoder and three anonymous reviewers for their helpful comments. All interpretations remain the author's. Please address all correspondence to the author at the Department of Sociology, University of Akron, Akron, OH 44325-1905; email: jzipp@uakron.edu.

Editor's note: The reviewers were, in alphabetical order, Kevin Dougherty, Michael Polgar, and Lee Streetman.

take greater charge of their own learning; and classmates can model helpful learning strategies.

Excellent resources are now available for sociologists interested in cooperative learning; the Center for the Advancement of Teaching at Illinois State University (http://www.cat.ilstu.edu/additional/active.php) is an especially good clearinghouse of information. Despite the quality and quantity of these resources, one cooperative teaching technique that has received relatively little attention is the two-stage cooperative, group, or "pyramid" exam (Cohen and Henle 1995; Cortright et al. 2003; Stearns 1996; Yuretich et al. 2001). Although there are variations in how they are used, the basic approach with these cooperative exams is that students first take an exam individually—as in traditional testing—and then take the same exam (sometimes with harder questions added) together with their learning group, with the exam grade being a weighted combination of their individual (typically 75-80%) and group scores. This approach uses the exam itself not only for evaluation, but also as a learning tool. When students discuss and debate exam answers in small groups, an exam can become an active learning exercise.

In my more than twenty-five years of teaching experience I have observed that many colleagues would heartily embrace an alternative to the traditional exam, especially if this modification increased student learning. Grzelkowski (1987) may have spoken for many instructors when she argued that our methods of examining students often are at odds with our lofty learning goals and pedagogical values.[1] Researchers and instructors are increasingly touting and using group work and active learning as effective learning strategies, but cooperative learning, it seems, typically ends on test day.

Like many of my colleagues, I have often used cooperative group exercises and assignments in my courses but stopped short of using any group component on exams. Sharing some of Grzelkowski's (1987) lack of comfort with this disjuncture, I recently re-organized my mass lecture (approximately 200 students) introductory sociology course to include two-step cooperative exams. On our four testing days, students first took a multiple choice test individually and then re-took the same test with their group, the latter time handing in one answer sheet per group. Because I repeated questions from each test on a final exam, I was able to see if how they did on the group exam, over and above their individual performance, affected how they scored on the final exam. In this paper I examine whether or not this modest attempt at two-stage cooperative testing improved relatively short-term student learning.

TESTING, TESTING

> One of the most odious tasks of teaching is testing. (Murray 1990:148)

In their comprehensive book on testing, Jacobs and Chase (1992) note that tests have four purposes: (1) as the basis of student grades, (2) to help instructors deliver content, (3) to increase student learning, and (4) to increase student motivation to learn (p. 2). The problem, they continue, is that too many instructors see tests solely or primarily in terms of the first of these purposes, ignoring the ways in which tests themselves are pedagogical techniques that can improve both teaching and learning. It is no surprise, then, that much of the literature on testing concentrates on such matters as how to write better questions, how to develop questions for different educational

[1]Believing that typical in-class, anxiety-producing exams were at odds with her humanistic perspective, Grzelkowski (1987) used take-home multiple-choice exams in her lower-level sociology courses, allowing students to work together on them. Although students reported greater satisfaction, Grzelkowski's design did not allow her to distinguish the effect of the take-home format from the effect of students working together.

objectives, or how to grade (e.g., Bloom 1984; Carneson, Delpierre, and Masters 1998; Magnan 1990; McKeachie 2002; Ory and Ryan 1993).

Perhaps reflecting Murray's (1990) remark at the top of this section, a small amount of the testing literature chronicles meaningful options to traditional forms of testing. These are typically grouped under the rubric of "alternative" testing (e.g., Davis 1993; Jacobs and Chase 1992; Murray 1990) and include such possibilities as second-chance exams, paired-testing (see also, Hendrickson, Brady, and Algozzine 1987), answer justification, open book, take home or oral examinations, group tests (e.g., Breedlove, Burkett, and Winfield forthcoming, 2004; Helmericks 1993; Russo and Warren 1999; Zimbardo, Butler, and Wolfe 2003), and pyramid exams (Cohen and Henle 1995; Cortright et al. 2003; Stearns 1996; Yuretich et al. 2001). Since the latter two approaches are closest to the task at hand, I will probe them in more detail.

Most of those using group or pyramid exams appear to be searching for alternatives to traditional forms of testing, without necessarily articulating a clear theoretical explanation for *why* cooperative testing as a pedagogical approach improves learning. The exceptions to this are Breedlove, Zimbardo, and their colleagues (Breedlove et al. forthcoming; Zimbardo et al. 2003). Collectively they offer several reasons why cooperative tests work: they may reduce test anxiety and stress; by working together, students can build on each other's strengths; collaboration may increase the motivation to learn; students may prepare more so as not to let down their peers; and discussion can help students think at higher levels and recall information better.

With or without a theoretical rationale, instructors in a variety of disciplines have relied on group exams in the hopes of increasing learning. Russo and Warren (1999) used them in English composition classes and report being disappointed to find almost no higher scores from the group exams compared to previous semesters' individual tests. Without any controls for differences in the quality and preparation of students across semesters, it is hard to draw much from their study. In a similar vein, Helmericks (1993) used three group midterms along with individual final exams in his sociology statistics course, contrasting them with the individual exams he gave in the same course in the prior semester. Scores were higher on the collaborative midterms than on the individual midterms, but Helmericks was surprised to find that students who took collaborative exams did worse on the individual final than did students from the prior semester's traditional course. Helmericks explained the latter by suggesting that students did not need as many points to get their desired grades, though a skeptic might suggest that a plausible alternative explanation is that students did not learn the material as well.

Three other studies of collaborative exams are more methodologically rigorous, but still have important shortcomings. After students took a traditional exam in large introductory psychology courses, Zimbardo et al. (2003) offered students the option of taking the subsequent two exams with a partner. Although the share of students electing this option varied considerably—from 30 to 62 percent—Zimbardo and his colleagues found significantly higher scores on the group tests as compared to the individual ones. In an interesting twist, Breedlove and his colleagues (forthcoming, 2004) explicitly tried to isolate the impact on performance of collaborative testing in the absence of collaborative learning. Using seven sections of introductory sociology across two different academic years, they used a quasi-experimental design to compare scores from a section in which students took exams individually to those in sections in which students were randomly paired to take exams with a same-sex student. After controlling for a host of background factors, they found that collaborative testing (1) improved performance on lower-level but not higher-level test items and (2) did not re-

duce test anxiety. Despite the quality of these studies, neither assesses the degree to which group exams promote learning above and beyond individual preparation.

The small amount of research on two-stage cooperative exams also has important methodological shortcomings. Stearns (1996) reports a significant difference in final exam scores between her traditional research methods course and the one that used cooperative group exams; however, she did not consider any differences in the quality of students across the courses. Yuretich et al. (2001) compared two mass lecture oceanography courses, one taught in a traditional lecture format and one taught with two-stage cooperative exams and a whole range of other active learning innovations. Although they found statistically significant improvements on the final exam in the latter course, they too did not take into account any possible differences in class composition, nor were they able to separate the effects of the exam structure from their many other instructional changes. Cohen and Henle (1995) describe an elaborate pyramid exam protocol they use in calculus courses, involving a series of progressively harder individual and group exams that culminate in a final that the entire class takes as a group. Although theirs is an impressive effort, they provide no data on any changes in student learning. Finally, Cortright et al. (2003) split an exercise physiology class into two groups, randomly assigning students to answer some questions initially individually and others initially in groups. Although they found slightly higher retention for questions answered by groups (52.9% of these were answered correctly the second time, while only 46% of the individual "repeat" questions were answered correctly at time 2), they did not control for any possible differences in difficulty between these two sets of questions.

In addition to these methodological problems, researchers have failed to adequately define what they mean by a collaborative exam *working* or, in other words, improving learning. Almost without exception, researchers have examined this concept by looking at whether average scores on a cooperative exam are higher than on exams taken individually. It is not surprising to find that an average group grade is higher than an average individual grade, but this comparison does not speak to whether the testing process promotes learning. The central question should be: Do group exams help students learn—above and beyond their levels of individual ability and preparation? If so, do they do so when students know the material initially and when they do not? To assess this, we must test whether participation in a group improves individual performance beyond the day of testing. Only one of the studies (Cortright et al. 2003) attempted to ascertain this, but, as I noted above, their design did not take into account possible differences in question difficulty across groups, nor did they distinguish the effects of individual accomplishment from group performance. This study seeks to fill that gap, isolating the impact of group tests from individual achievement.

THE PROJECT

I conducted this project at an urban, commuter university with 24,000 students; our department typically offers 15 to 20 sections of introductory sociology each semester, serving more than 2000 students each year. Several sections are mass lectures (100-250 students), while most of the rest have 40 to 50 students. Relying on a mix of full-time and part-time faculty and graduate students as instructors, more than a dozen years ago the department began to adopt a textbook (on a three-year cycle) and create a computerized test bank for it (the text in use during this project was Curry, Jiobu, and Schwirian 2005).

I had taught this course for five years, each time using the test bank to assess student learning of the material in the textbook. Because the test bank questions tend to be fairly difficult (exam scores typically average 65-70%), all introductory sociology instructors allow students a second attempt

for each test, typically with a day delay intended to be used for studying. Since questions on a particular topic (5 to 8 questions per topic) are randomly chosen each time a student takes a test, students are not likely to encounter the same question again. Scores tend to be 5 to 10 percent higher on the second attempt.

Despite the better performance on the second attempts, a fair number of students only take each test once, and I have never been sure how much students who take the tests again actually study or learn between attempts. In addition, even though for several years I have had students work in groups in the classroom on various projects and quizzes, I had never extended any group work to tests—a practice that appeared to be somewhat inconsistent with my learning goals. Recognizing this incompatibility and wanting to use the tests I give as learning opportunities, I decided to use questions from the computerized test bank, but administer them in class in the form of a standard paper and pencil test. Recognizing how difficult these questions tend to be, I decided to give students a second chance to answer them with their "learning" groups.

During the second week of class, I placed students into six-person, heterogeneous learning groups. I stratified groups by gender (most groups had an equal number of women and men, though some had more women due to the class composition) and students' self-reported scores on a "verbal-linguistic" learning styles test available in the textbook (Curry et al. 2005: xxii-xxiii). This allowed me to create groups that comprised both women and men with a combination of individuals who saw themselves as having differing strengths as verbal-linguistic learners. Students sat with their groups every day, and throughout the semester I used a variety of individual and group exercises and assignments.

I gave four exams during the semester (each covered four chapters), along with an optional, cumulative final, and I administered each of the exams as a two-step cooperative test during a regular 100-minute class period.[2] Each test contained 40 multiple-choice questions (with four answer choices for each question) that students first answered individually, recording their answers both on the question sheets and on separate answer sheets. After everyone handed in their individual answer sheets, students answered the same questions in their groups, turning in one answer sheet per group. Immediately after all the groups handed these in, I told the class the correct answers to each question. I gave students approximately 20 to 30 minutes for the individual exam (rarely did they need this long), with most of the rest of class time (typically 60-70 minutes) for the group test.

Students' grades were a combination of their performance on the individual and group exams: students received the grade on their individual exam, plus half the difference between the group average (based on the six individual scores) and the group score. For instance, suppose that the six individuals in Group X got the following grades on their exams: 90, 80, 70, 60, 50, and 40. The average of these individual scores is 65. Suppose also that Group X got an 85 on their group exam. This means that every member of Group X would get 10 extra points on this exam (85 [the group score]—65 [the average individual score within the group] divided by 2), raising their scores to 100, 90, 80, 70, 60, and 50.[3]

[2]Introductory sociology is a four-credit course; this class met twice a week for 100 minutes each time.

[3]These weights, either explicitly or implicitly, represent an instructor's differential rewarding of individual and group achievement. Others (e.g., Yuretich 2003) use a weighted average of the individual (75%) and group (25%) exams to comprise a student's grade, with the proviso that they do not allow a lower group grade to pull down a higher individual grade. I chose not to use this approach, as it (1) gives more points to students who do the poorest on the individual exam (e.g., in the illustration above, the student who got an 80 on the test would only receive 1.25 additional points, while the student with a 40 would be given an additional 11.25 points); (2) it could be a disincentive for the

As might be expected, average scores on the group exams (84.0%) were much higher than on the individual exams (65.5%). In fact, across all exams only 3.8 percent of students scored better than their group on a particular test.

With almost 200 students in the course and wanting to avoid having to field make-up exams, I also gave an optional, cumulative final exam. This exam, I told students, would consist entirely of questions asked on one of the four previous exams (10 questions from each exam). If students took all five tests, their four highest grades would count; if they missed an exam, their score on the final would replace it. Of the 194 students in the course, 122 (62.9%) took the final. With the exception of having in-class two-stage cooperative exams, this is the *same* testing practice that I have used in the past: a set of tests during the semester and an optional cumulative final consisting of questions already asked.[4]

My primary reason for using the two-stage cooperative group exams was academic: I wanted students to work together, to debate and discuss their answers, so that all the exams became learning opportunities for each student. The way that I did this, however, created the possibility for a quasi-experimental design for testing the impact that the two-stage cooperative exams had on (relatively short-term) student learning, a design that balanced my academic goals with an appropriate test for determining the impact of group exams.[5]

Figure 1. Initial and Final Exams for Individuals and Groups

	Student Score	Group Score
Initial Exam	#1	#2
Final Exam	#3	#4

Since each of the 40 questions on the final exam had been asked on a prior test, there are four scores for each student for each question (see Figure 1): how the student (score #1) and her or his group (score #2) performed on that question initially on one of the tests held during the semester and how the student (score #3) and her or his group (score #4) performed on that question on the final. Because the measure of student learning (my dependent variable) is how students performed on their own in the final exam (score #3 above), there is no reason to look at the group performance on the final exam (score #4 above). This leaves us with the first two scores, and these are my explanatory variables: how students did on the question in the initial individual exam (score #1)—a measure of individual achievement—and how students did on the group exam (score #2)—a measure that reflects the impact of group performance. I used these two variables to form four quasi-groups (see Figure 2): questions that were answered

best students, perhaps leading them not to contribute as much to the group exam (e.g., the student who earned a 90 on the individual exam would not gain any points from the group exam because this weighting formula would produce a score of 88.75); and (3) it increases any errors involved in the construction of the groups. My weighting relied much more on individual performance; across the exams, 86.7 percent of each student's composite test grade came from the individually-completed test, with only 13.3 percent stemming from the group exam.

[4]There were 900 total points in the course, with 400 (44.4%) coming from these cooperative exams. An analytical paper, completed individually and in a series of steps, was worth 200 points, while the remaining 300 points came from 20 in-class quizzes/problems, each of which was worth 15 points and was done in a format similar to the exams: students first attempted it individually (10 points) and then worked in their groups (5 points).

[5]Perhaps the ideal way to test the impact of group exams would have been to randomly assign some students to take the tests individually, others to take group tests, and to alter who was in what condition across the semester (a randomized crossover design). Although this might have been methodologically more rigorous, not allowing every student the opportunity to take a group exam at each testing time would have sacrificed one of my academic goals at the methodological altar. This project was reviewed by and received the approval of the university's Institutional Review Board (IRB).

Figure 2. Individual and Group Performance on Initial Exams

		Group Exam	
Individual Exam		**Incorrect Answer**	**Correct Answer**
Incorrect Answer		Cell A	Cell b
Correct Answer		Cell C	Cell D

incorrectly on both the individual and the group exams (Cell A), questions answered incorrectly individually but correctly on the group exam (Cell B), questions answered correctly on the individual exam but incorrectly on the group exam (Cell C), and questions answered correctly on both (Cell D).

I refer to these as "quasi-groups," since they represent not a fixed subset of students but a student-question combination that differs by each question. For instance, a student could have answered Question 1 correctly both individually and in the group (Cell D), but Question 2 incorrectly on both (Cell A), and so on. Thus, each observation is a combination of performance on the individual and group exams for each student. With 40 questions across 122 students, this means that there are 4880 possible data points (40 x 122 = 4880) that I sorted into one of the four cells in Figure 2.[6] It is important to keep in mind that although the unit of measurement is the student, the unit of analysis is the test question.

My central question concerns whether the processes involved in taking a group exam (e.g., the debate, discussion), taking into account how students perform on their own individual test, can impact relatively short-term learning. In my terms, I am testing if, despite how a question was answered initially on the individual exams, being in a group that answered the question correctly means that this item is more likely to have been answered correctly on the final. This involves two sets of comparisons of the cells across rows in Figure 2: questions answered incorrectly initially (Cell A vs. Cell B) and correctly initially (Cell C vs. Cell D). In both cases, these comparisons hold individual achievement constant and allow me to test my central hypothesis: *Independent of whether a question was answered correctly or not initially on the individual exam, being answered correctly on a group test increased the likelihood that the question was answered correctly on the final exam*.

RESULTS

Before turning to my analyses of the effects of group exams, I will describe the distribution of student-question combinations into the four cells depicted in Figure 2. Table 1 shows that 59.2 percent of the questions were answered correctly on both the individual and the group tests, 25.2 percent were wrong initially but right on the group exams, 4.1 percent were correct on the initial exam but were incorrect on the group tests (my suspicion is that students either guessed right when initially taking the exam, were less confident in their answer than other group members were, or were not able to persuade their other group mem-

[6]Five students missed the first test, three missed the third test, and six missed the fourth test; thus, there only are 4740 data points.

Table 1. Comparing Performance on the Individual and Group Exams

	Group Exam		
Individual Exam	**Incorrect Answer**	**Correct Answer**	**Total**
Incorrect Answer	11.6%	25.2%	36.8%
(N)	(548)	(1195)	(1743)
Correct Answer	4.1%	59.2%	63.2%
(N)	(191)	(2806)	(2997)
TOTAL	15.6%	84.4%	
	(739)	(4001)	(4740)

Note: Entries are cell percentages; e.g., 11.6% of students answered incorrectly on their individual and group exams; 36.8% answered incorrectly on the individual exams.

bers to adopt their answer), while 11.6 percent of the questions were incorrectly answered both times.

Table 2 contains the results that bear on my central hypothesis. I hypothesized that the learning process was affected by both individual and group factors. We can gauge the impact of the first effect by examining the marginals in Table 2. The far right column of Table 2 shows that individual achievement played a sizeable role: 50.6 percent of the questions answered wrong initially were answered correctly on the final exam, while 83.2 percent of those answered correctly the first time also were answered correctly on the final ($t=23.65$; $p < .001$, two-tailed test). This difference, of course, is not surprising, as it suggests that those who knew the material initially were more likely to know it at the end of the semester. On the positive side, almost half of the questions answered wrong initially were answered right on the final, suggesting that some learning took place. On the negative side, almost 17 percent of the questions answered right the first time were not answered correctly on the final; this may represent those who guessed right the first time but wrong on the final.

The main question, however, concerns the impact of the group test. Some surely may have gotten the correct answers on the final because I gave the answers to all these questions from 1 to 10 weeks earlier. The key issue, then, is whether being a part of a group that got the question right, net of individual performance, increased the likelihood of getting the correct answer on the final. The two relevant comparisons appear in the rows in Table 2. In both cases, being part of a group that got the question right resulted in a statistically significant increase in likelihood of answering the question correctly on the final. For questions answered incorrectly on the individual test (the top row), being in a group that answered correctly resulted in 53.3 percent answering correctly on the final, as opposed to only 44.7 percent correct for those who were in groups that got the wrong answer ($t=3.34$; $p < .001$, two-tailed test). The differences are even more dramatic for questions that were answered correctly initially (the bottom row): 84.1 percent were correct on the final when the group got the question right, while only 71.2 percent were when the group got the question wrong ($t= 3.85$; $p < .001$, two-tailed test).

Thus, these results indicate that being part of a group that answered correctly in the two-step cooperative group exam improved performance on the final for students who knew the material originally and for those who did not. For instance, the effect of being in a group that answered correctly on the group exam was 8.6 percentage points (53.3% vs. 44.7%) for questions answered incorrectly on the individual exam. Further-

Table 2. The Impact of Group Exams

	Percent (%) of questions answered correctly on final exam		
	Group Exam		
Individual Exam	**Incorrect Answer**	**Correct Answer**	**Total**
Incorrect Answer	44.7%[b]	53.3%[b]	50.6%[a]
(N)	(548)	(1195)	(1743)
Correct Answer	71.2%[c]	84.1%[c]	83.2%[a]
(N)	(191)	(2806)	(2997)
TOTAL	51.5%	74.9%	71.2%
	(739)	(4001)	(4740)

t-tests (comparisons between cells with the same superscript)
a: $t=23.65$; $p < .001$
b: $t=3.34$; $p < .001$
c: $t=3.85$; $p < .001$

more, group discussion appears to have had a bigger impact on questions answered correctly initially, depressing learning when the group got it wrong (71.2% on the final) and reinforcing it when the group got it right (84.1%). This difference is almost 13 percentage points; in my grading scheme, the difference is substantial, as it distinguishes a C- from a B. This suggests that, above and beyond an individual's preparation and knowing the correct answer, being part of a group that answered correctly helped students retain material, at least in the short-run.

Before concluding this analysis, I will note several key points about how my quasi-experimental design may affect the results. First, my underlying model is that performance on the final was due to two factors: initial individual achievement (score #1 above) and group learning (score #2 above). Thus, I contend that those who understood the material better initially and/or whose group answered correctly should have done better on the final exam. However, two alternative explanations could also account for students doing differently on the final exam than on the previous exams: (1) differences in ability and/or preparation, and (2) differential guessing. Since the average correlation between the first four tests and the final (.493) was reasonably strong and positive, those who did well initially were more likely to do well on the final. This suggests that differential ability or preparation is unlikely to account for much variation in results between the initial tests and the final.

It is also possible that differences in guessing could explain differential performance on the final as opposed to earlier tests. Measurement theory tells us that those who disproportionately guessed right the first time should guess incorrectly on the final, and vice versa. This would cause regression to the mean between the initial tests and the final: those who did very well initially should do worse on the final, while those who did very poorly should do better. However, since my comparisons are within rows of Figure 2, the only way for this to affect my results is if individual guessing is correlated with how a group answered a question or if there was regression to the mean for groups.[7] On the first point, we commonly

[7]Regression to the mean on the final would raise the average score on the final for those who got the questions wrong initially (the marginal in the top row of Figure 2), while depressing the final grade for those who answered correctly the first time (the marginal in the bottom

observed just the opposite: students admitted to their groups that they guessed on a particular question in order to not play a prominent role in the group discussion on that question. Furthermore, if groups who guessed wrong the first time should be more likely to guess right on the final, with the opposite occurring for groups who guessed correctly the first time, this would reduce the impact of group exams, as average scores on the final would be higher in Cells A and C and lower in Cells B and D. Since my comparisons are between Cells A and B and between Cells C and D, any effect of differential group guessing would make it harder for me to find an effect attributable to the group exam process.

A third threat to the validity of my results is selection bias: it is possible that the 72 students who did not take the final exam differ in important ways from those who took it. Although I cannot rule out all possible effects of this, selection is not likely to be a serious problem since there were no significant differences between these two groups of students on three of the four semester exams.

On the other hand, in one way this design constitutes a conservative estimate of the impact of two-stage cooperative group exams. As I noted earlier, at the end of the initial testing period, each student knew the correct answer. Thus, regardless of whether they answered correctly or not individually or in groups, all students left the classroom on the initial testing day with the correct answers to each question. Because I have held their individual performance constant in my comparisons, any impact of being told the correct answer is likely to reduce the variation attributable to group performance.

At a more general level, all cooperative learning exercises present two potential problems: unanticipated troubles stemming from group construction and "free riders"—students who rely on the efforts of others without themselves making adequate contributions. Taking group composition first, most of the literature recommends that instructors form groups and create them to be diverse with respect to various sociodemographic and academic attributes (Millis and Cottell 1998). Heterogeneity will produce better balance than, for example, when students self-select their groups. As I noted earlier, I formed teams based on two factors: gender and self-reported scores on verbal-linguistic learning styles, the latter an attempt to measure speaking, reading, and written skills (Curry et al. 2005). As it turned out, this score was virtually uncorrelated with performance on the individual tests and in the course as a whole. Although it would have been better if I had been able to form groups on factors more related to academic achievement, not doing so is not a threat to the validity of my results since I am trying to ascertain the impact of group learning on short-term individual learning.[8]

The literature on cooperative learning suggests that fostering individual responsibility is a key in minimizing the threat of free-riders (e.g. Johnson et al. 1991; Millis and Cottell 1998). One way that I created individual responsibility in my course was by having almost 87 percent of the test grade stem from performance on the individual exam (see footnote 3); thus, any student expecting to do well mainly due to the

row of Figure 2). Since my comparisons are within these rows, differential guessing by individuals has no bearing on them.

[8]The real threat is that I may have created groups that were unbalanced with respect to academic abilities, thus potentially helping some students (e.g., a low achiever in a group of high achievers) while hindering others (e.g., a low achiever in a group of low achievers). Since many factors other than test scores shape performance (e.g., effort, attendance, etc.) it would be virtually impossible to eliminate this problem completely. However, it is comforting to note that, since there was virtually no correlation (.07) between the group's average verbal-linguistic score and its average performance on the four group exams, it appears that my method of forming groups did not have the unintended consequence of significantly privileging some while harming others.

group exam would pay this freight. Although I cannot rule out the possibility that some students were willing to take this chance, one bit of indirect evidence suggests that free riding was not a serious problem. In seven groups truly outstanding "A" students did not take the final exam; these students were by far the best students in their groups and it would be surprising if each group had not recognized that fact. If other group members came to exams expecting to ride the coattails of these students, we might expect these free-riding individuals to fare much more poorly on the final exam, as the top students were not there for help. In fact, if I eliminate these seven students from all exams—including the final—there was no significant difference on any exam between the performance of their groups and the average of all other groups. In other words, the students in groups with high performers—those who were in positions most likely to allow them to free-ride—scored about the same on their individual tests when that high performer was there and when that high performer was not there.

Two final caveats bear mentioning. I measured short-term learning by performance on a multiple-choice test, and although multiple-choice questions are the most widely used exam items (Jacobs and Chase 1992), they constitute only one measure of learning. Others (e.g., Cohen and Henle 1995; Cortright et al. 2003) have used two-step cooperative group exams that relied at least in part on short answers, essays, and papers. Since taken together these will provide a more complete assessment, it is perhaps better to see my results as analyzing one step in a broad process of learning.

In addition, it is important to note that my findings are based on the 122 students who took the final exam in one introductory sociology course at an urban, midwestern, state university. Our student population is predominantly first generation, commuting, white students (though my course was roughly one-third non-white), with African-Americans as the primary minority group. It is possible that these results might vary in other settings.

DISCUSSION

I began this paper by noting that previous studies of cooperative group exams did not distinguish between the impact of individual performance and group performance. Before discussing the impact of group learning, I must note that group learning is not as important as individual preparation and cannot take its place. One way to see this is to compare performances on the final for questions answered incorrectly individually but correctly in the group (upper right hand cell in Table 2) with those answered correctly individually but incorrectly in groups (lower left hand cell). As indicated in Table 2, the average score on the final was 53.3 percent for the former but 71.2 percent for the latter. This certainly indicates that the learning that took place in this two-step cooperative exam was not as important as ability and the work individuals did in preparing alone for these exams. A second point also bears noting: recall that 44.7 percent of questions answered incorrectly on both the individual and group exams were answered correctly on the final. Since groups got this question wrong, the most likely explanation is that this is the effect of being told the correct answer after the group test.

Despite these qualifications, my results support the notion that the processes associated with a group exam increased relatively short-term student learning, at least as measured by answers to multiple-choice questions on a final exam. This impact was net of individual achievement and manifested itself both for those who knew the material and for those who did not know the material, with it resulting in an improvement of more than a full letter grade for the former. And, since this exercise was a relatively conservative assessment of group learning, it is possible that this represents the lower boundary of the impact of these sorts of exams.

Having shown that being part of a correctly answering group compared to being part of an incorrectly answering group in a cooperative group exam increases short-term retention of material, it is reasonable to ask how the exam does this. Although it is not possible to answer this question completely without collecting detailed information on the actual group deliberations, my results allow me to shed some light on this important question. To begin with, it stands to reason that the group exam operates differently for questions answered correctly initially and those answered incorrectly initially. For questions answered correctly, it is likely that the two-step cooperative group exam *reinforces* individual learning. Recall that there was a 12.9 percentage point difference (84.1% vs. 71.2%) on the final exam questions between groups that answered correctly and those that answered incorrectly on the group exam—a difference of more than a full letter grade. This suggests that reinforcement is reasonably powerful and valuable in helping the retention of knowledge. In addition, the presence of this reinforcing effect may be an effective counter to those who might argue that high performing students benefit little from group work.

By definition, however, reinforcement cannot account for correct answers on the final for questions answered initially incorrectly. Instead, what may be occurring here is *group-to-individual transfer of learning* (see Johnson and Johnson 1989:50-2 for a summary): individuals learn the material in a cooperative group and then transfer this learning to a subsequent individual test. In this case, participating in a group that answered a question correctly helped students who initially got the question wrong to learn the material well enough to transfer it to the final exam.

In addition to these documented benefits, I gathered impressionistic evidence regarding the process itself. As I noted, students sat with their learning groups throughout the semester (I gave almost daily quizzes or exercises, most of which were in the same two-step format; see footnote 4) and thus, were quite familiar with each other's strengths and weaknesses. During the group exam, the teaching assistants (two graduate assistants and one undergraduate peer tutor; at times, two other graduate students, both taking our College Teaching of Sociology course, were also there) and I circulated throughout the classroom, observing and sitting with groups. With four to six of us doing this throughout the semester, we were able to gain a pretty fair sense of how the groups worked on the cooperative exams.

With a few exceptions, almost every student was actively engaged in the group exam. The norm was for group members to begin by sharing their answers on each question; the least thoughtful groups took the modal answer as the group response without any real debate and discussion, while most of the groups debated disagreements—occasionally trying to get the teaching assistants or me to shed some light on the answer. Students were quite honest and open in their discussions; we frequently heard some admit that they just guessed, while others cited specific parts of a chapter, a lecture, or an example to argue for their answers. The discussions were loud and animated, with "I told you so's" cascading through the room when individuals learned that their answer was correct.

Although most of the groups were quite cohesive and generally looked forward to working with each other, there were a few problems with the testing design that could be improved in the future. The first of these concerns my requirement that each group agree on the answer for each question. Some individuals who had correct answers clearly got outvoted by their groups. In the future, it may be better to have group discussion but then let individuals turn in their own answer sheets. This could have the benefits of the group process but not force everyone to agree. A second improvement might be to extend this modification to require students to provide reasons for any change in answers between their individual and group exams, counting only the correct

answers that include correct reasons. This would, of course, mean fewer exam questions, but has the possibility of getting at deeper learning.

These results also bear on several broader questions regarding cooperative testing. At one level, the whole idea of cooperative testing appears to conflict with the purpose of grading: measuring what the individual student knows (Davis 1993). Since the best way to assess individual achievement is by having students work alone, the use of cooperative testing introduces error into the grading process. Because group averages are typically higher than individual ones, grades formed by cooperative testing are likely to overestimate individual achievement (Webb 1993). The problems associated with this may not be trivial, as access to many resources—jobs, graduate and professional schools, scholarships, awards, and so on—depends at least in part on grades that are presumed to be indicators of individual achievement. A student who happened to take a considerable number of courses relying on cooperative testing might obtain some rewards denied to a student of similar competence who took few, if any, such courses. Such a specter goes against the very meritocratic assumptions of higher education.

Although at face value these are valid concerns, a bit of probing suggests a more complex process. To begin with, it is important to explicitly acknowledge that grades are *"socially constructed and context-dependent"* (Walvoord and Anderson 1998:10; emphasis in original). As much as we might want to believe otherwise, there simply is no absolutely right grading standard. Milton, Pollio, and Eison (1986) show not only how the meaning and assignment of grades have changed historically (e.g., changes in grading scales over time; the introduction of pass-fail courses; average GPAs rising between 1965 and 1980, just as average SAT scores were falling), but also how grading practices are idiosyncratic by institution, discipline, and instructor. Many academics either implicitly or explicitly acknowledge this. For instance, in evaluating potential graduate students, a 3.0 at an Ivy League school may be interpreted as indicating greater individual achievement than a 3.5 at an open-admissions regional state university. I suspect that a fair number of faculty know a colleague who has a reputation of giving all A's, or someone who might (so to speak) "flunk their mother." Given this sort of variation and how much interpretation already is required, it is hardly likely that any grade inflation associated with cooperative testing will do much damage to what is already a somewhat problematic relationship between grades and individual competence.

Second, cooperative testing is one form of active learning, and as such, is open to the criticism that these sorts of activities reduce the amount of time available for covering content (Millis and Cottell 1998). For instance, in a recent *Teaching Sociology* article, I reported (Zipp 2002) that after describing my active learning exercise to a colleague he dismissively informed me that he had too much to cover to waste his time on those types of "games." One surely need not agree with this glib response to recognize that using cooperative learning groups, two-step exams, or active learning in general can consume considerable class time.

This being said, there are always tradeoffs in any course design. Having students present papers or guide discussion, both potentially valuable learning opportunities, also reduces the time that the instructor can lead class. The key point, however, is not the amount of class time that is devoted to teaching but how the course is organized to promote learning. Millis and Cottell (1998), for instance, describe ways in which cooperative classrooms, by increasing student motivation and preparation, can actually cover more content than traditional lecture courses. And, given the effect that cooperative group work has on learning, we might wonder why instructors do not use these approaches more frequently.

Finally, and coming full circle, the results contained here provide a method for those

who rely on cooperative learning to extend the group process to exam day. This not only enhances learning but it also allows for the process and form of testing to become more closely linked to the process and form of teaching and learning. The symbolic importance of this latter point should not be underestimated, as it provides an important way for faculty members to practice what they preach on exam day. Although we as faculty may see every class period as equally important, surely the average student focuses more on test day than on most lectures. Having a consistent format between lecture and testing days, then, sends a powerful signal to students about our commitment to group learning.

REFERENCES

Astin, Alexander W. 1993. *What Matters in College? Four Critical Years Revisited*. San Francisco, CA: Jossey-Bass.

Bloom, Benjamin S., ed. 1984. *Taxonomy of Educational Objectives*. New York: Longman.

Breedlove, William, Tracy Burkett, and Idee Winfield. 2004. "Collaborative Testing and Test Performance." *Academic Exchange Quarterly* 4:36-40.

_____. Forthcoming. "Collaborative Testing and Test Anxiety." *Journal of Scholarship of Teaching and Learning* 4.

Carneson, John, Georges Delpierre, and Ken Masters. 1998. "Designing and Managing Multiple Choice Questions." *University of Cape Town*. Retrieved October 2006 (http://www.le.ac.uk/castle/resources/).

Cohen, David and James Henle. 1995. "The Pyramid Exam." *UME Trends* 10:2,15.

Cortright, Ronald N., Heidi L. Collins, David W. Rodenbaugh, and Stephen E. DiCarlo. 2003. "Student Retention of Course Content is Improved by Collaborative-Group Testing." *Advances in Physiology Education 27*:102-8.

Curry, Tim, Robert Jiobu, and Kent Schwirian. 2005. *Sociology for the Twenty-First Century*. 4th ed. Upper Saddle River, NJ: Prentice-Hall.

Davidson, Neil. 1990. *Cooperative Learning in Mathematics*. Reading, MA: Addison-Wesley.

Davis, Barbara Gross. 1993. *Tools for Teaching*. San Francisco, CA: Jossey-Bass.

Grzelkowski, Kathryn P. 1987. "A Journey toward Humanistic Testing." *Teaching Sociology* 15:27-32.

Helmericks, Steven G. 1993. "Collaborative Testing in Social Statistics." *Teaching Sociology* 21:287-97.

Hendrickson, Jo M., Michael P. Brady, and Bob Algozzine. 1987. "Peer-Mediated Testing: The Effects of an Alternative Testing Procedure in Higher Education." *Educational and Psychological Research* 7:91-101.

Instructional Technology and Development Center, Illinois State University. 2006. "Active Learning Strategies." *Center for Teaching, Learning, and Technology*. Retrieved October 2006 (http://www.cat.ilstu.edu/additional/active.php).

Jacobs, Lucy Cheser and Clinton I. Chase. 1992. *Developing and Using Tests Effectively*. San Francisco, CA: Jossey-Bass.

Johnson, David W. and Roger T. Johnson. 1989. *Cooperation and Competition: Theory and Research*. Edina, MN: Interaction.

Johnson, David W., Roger T. Johnson, and Karl A. Smith. 1991. *Active Learning: Cooperation in the College Classroom*. Edina, MN: Interaction.

Magnan, Robert, ed. 1990. *147 Practical Tips for Teaching Professors*. Madison, WI: Atwood.

McKeachie, Wilbert J. 2002. *Teaching Tips: Strategies, Research, and Theory for College and University Teachers*. Boston, MA: Houghton Mifflin.

Millis, Barbara J. and Phillip G. Cottell. 1998. *Cooperative Learning for Higher Education Faculty*. Phoenix, AR: American Council on Education and the Oryx Press.

Milton, Ohmer, Howard R. Pollio, and James A. Eison. 1986. *Making Sense of College Grades*. San Francisco, CA: Jossey-Bass.

Murray, John P. 1990. "Better Testing for Better Learning." *College Teaching* 38:148-52.

Ory, John C. and Katherine E. Ryan. 1993. *Tips for Improving Testing and Grading*. Newbury Park, CA: Sage.

Quarstein, Vernon A. and Polly A. Peterson. 2001. "An Assessment of Cooperative Learning: A Goal-Criterion Approach." *Innovative Higher Education* 26:59-77.

Russo, Antonio and Susan H. Warren. 1999. "Collaborative Test Taking." *College Teaching* 47:18-20.

Slavin, Robert E. 1990. *Cooperative Learning: Theory, Research, and Practice*. Englewood Cliffs, NJ: Prentice-Hall.

Stearns, Susan A. 1996. "Collaborative Exams as Learning Tools." *College Teaching* 44:111-

2.
Walvoord, Barbara E. and Virginia Johnson Anderson. 1998. *Effective Grading: A Tool for Learning and Assessment*. San Francisco, CA: Jossey-Bass.
Webb, Noreen M. 1993. "Collaborative Group versus Individual Assessment in Mathematics: Processes and Outcomes." *Educational Assessment* 1:131-52.
Yuretich, Richard F. 2003. "Encouraging Critical Thinking: Measuring Skills in Large Introductory Sciences Classes." *Journal of College Science Teaching* 33:40-6.
Yuretich, Richard F., Samia A. Khan, R. Mark Leckie, and John J. Clement. 2001. "Active-Learning Methods to Improve Student Performance and Scientific Interest in a Large Introductory Oceanography Course." *Journal of Geoscience Education* 49:111-9.
Zimbardo, Philip G., Lisa D. Butler, and Valerie A. Wolfe. 2003. "Cooperative College Examinations: More Gain, Less Pain when Students Share Information and Grades." *Journal of Experimental Education* 71:101-25.
Zipp, John F. 2002. "The Impact of Social Structure on Mate Selection: An Empirical Evaluation of an Active Learning Exercise." *Teaching Sociology* 30:174-84.

John Zipp is professor and Chair in the Department of Sociology at the University of Akron. His primary teaching responsibilities are in introductory sociology and statistics, with current research interests in inequality and the scholarship of teaching and learning. He currently is the Chair of the ASA Section on Teaching and Learning in Sociology.