COLLABORATIVE TESTING IN STATISTICS: GROUP INTERACTION, ANXIETY, AND CLASS PERFORMANCE

SUSAN KAPITANOFF
American Jewish University
skapitanoff@aju.edu

CAROL PANDEY
Los Angeles Pierce College
pandeycj@piercecollege.edu

ABSTRACT

Seventy-one students in two community college Statistics for the Social Sciences classes took six exams either individually or collaboratively. Assignment to test condition was randomly determined for each exam.. Scores on collaboratively-taken exams were significantly higher than those for individually-taken exams, particularly for students with low GPAs and high test anxiety. Women's, but not men's, performances on the mid-term and final exams was related to the quality and quantity of their collaborative interactions. Thus, examining both quantity and quality of collaboration adds to our understanding of the underlying mechanisms of collaborative testing.

Keywords: *Statistics education research; Collaborative learning; Statistics anxiety; Test anxiety, Women and statistics*

1. INTRODUCTION

1.1. COLLABORATIVE TESTING

Collaborative testing, where students work together to complete an examination, is a recognized and important aspect of collaborative learning. It should be examined in detail to determine its specific contribution to the collaborative learning process. In many studies, beginning as early as 1944 (Klugman), collaborative testing has been found to reduce anxiety, increase critical thinking and enthusiasm, provide experience in cooperation and communication, and improve academic performance. The difference in scores between collaborative and traditional individual testing has ranged considerably (e.g., 3.82%, Pandey & Kapitanoff, 2011; 27%, Cortright, Collins, & Di Carlo, 2005) probably because the implementation of collaborative testing has varied greatly from application to application.

Collaborative testing is hypothesized to improve scores not just because of the sharing of answers but through enhanced cognitive processing. To investigate how collaborative testing might serve as a learning experience, Dahlström (2012) had female students in their third year of a speech and language pathology program enrolled in a course in basic methods and statistics take a pretest individually before taking their final exam collaboratively. They then took a posttest individually. This procedure is a variant of the popular *two-stage exam* collaborative method. Results indicated that the collaborative testing experience led to learning at several levels of knowledge and that posttest scores increased more for low performers on the pretest than for high pretest performers. In a review, LoGiudice, Pachai, and Kim (2015) concluded that there are cognitive mechanisms involved in collaborative testing that facilitate learning through such processes as retrieval effort and error pruning.

Collaborative testing has been employed in many different disciplines and academic levels. As mathematics anxiety tends to be high among American students (Campbell, 2004), it seems probable that collaborative testing would be especially valuable in mathematics and mathematics-related classes. Statistics, in particular, is considered by many students to be both mathematics-related and difficult.

The high levels of anxiety surrounding statistics experienced by these students may impair their achievement in the subject (Fitzgerald, 1997; Onwuegbuzie & Wilson, 2003), lead them to avoid taking statistics, and impact their choice of careers (Watson, 1988).

The focus of this paper is on how collaborative testing affects performance in college students taking statistics classes. This is particularly relevant for today's students as many of them must take statistics regardless of their majors. Statistics is fundamental to STEM majors, disciplines in which women as well as minority students are underrepresented (National Science Foundation, 2017). Regardless of career goals and major, in today's increasingly complex world, some understanding of statistics is vital.

Results of collaborative testing in statistics classes have been mixed. Townsend, Moore, Tuck, and Wilton (1998) had students enrolled in an educational psychology course spend 60% of their time in the associated computer statistics lab working in small groups. Their lab grades, however, were based on individual reports. Although the students' "mathematics self-concepts" and confidence in dealing with statistics problems increased over the course of the term, no significant reduction in mathematics anxiety occurred, nor did performance improve. Helmericks (1993) had college students in an introductory Statistics for the Social Sciences class take their exams in mixed-sex groups assigned anew for each test. On these tests combined, they performed 13.46% better than had students from the previous semester who had taken the same tests individually and they dropped the class less frequently. They also reported increasing enjoyment of the collaborative testing experience as the course progressed. However, on the final exam, which was taken individually, the collaborative-semester students scored 5.75% lower than those in the previous semester.

In a Research Methods and Statistics class (Stearns, 1996), college students took the same four midterm exams individually and then again in groups. On the final exam, which was taken individually, they scored 10.84% higher than had students from the previous semester who took no tests collaboratively. They also dropped the class less frequently. Berry and Nyman (2002) found that college students taking a mathematics modeling class liked team testing more than individual testing. A study by Giraud and Enders (1998) of undergraduates enrolled in two introductory Statistics for the Social Sciences classes found that whereas performance on the midterms and final exam as well as self-reported study times did not differ significantly for individual and collaborative conditions, students' liking of cooperative testing increased significantly over the semester. Björnsdóttir, Garfield, and Everson (2015) and Dallmer (2004) also found that collaborative testing improved students' attitudes towards statistics over the course of the semester.

Mixed results have also been found for younger students taking various mathematics classes. Singer (1990) found that collaboratively-taken test grades in junior high prealgebra students improved by 9% over the course of the semester whereas during the previous term, when students had taken tests individually, scores dropped by 5%. These students also reported that they preferred to take tests in pairs rather than individually. Klugman (1944) had elementary school children complete one form of a standardized arithmetic reasoning test individually and the other form in assigned pairs. For the sample as a whole, as well as every subgroup of the sample (categorized by age, sex, etc.), the paired condition yielded significantly higher scores than the individual condition. But collaborative testing was not beneficial for all participants. Fifteen percent of the sample scored equally well under the two conditions, and 22% did better when tested individually. Working with elementary mathematics students of differing performance levels, Fuchs et al. (1998) concluded that collaborative testing may yield better results for above-average students than for those at or below grade level. Thus, it remains unclear what aspects of the collaborative process—cognitive or affective—lead to reduced anxiety and improved class performance.

1.2. QUALITY OF COLLABORATIVE GROUP INTERACTION

Johnson, Johnson, and Smith (1991), pioneers in the field of collaborative testing, maintain that in order to achieve productive levels of interaction among group members, preliminary training in interpersonal and communication skills is necessary. Attempts to relate the quality and types of group interaction to performance, drop-out rate, and enjoyment may help to explain why collaborative testing has not always resulted in improved final exams scores. Is there true collaboration going on or is there simply chatting or the passing of answers from one student to another in the collaborative group?

Of the many studies reviewed for this research, only a few mentioned trainings of any kind or examined the quality of interaction in group discussions. Most frequently, students were simply encouraged to discuss the questions quietly. Occasionally, more specific advice was given such as Stearns' (1996) suggestion to participants to discuss rationales for each answer rather than merely taking a vote, and Webb, Nemer, Chizhik, and Sugrue's (1998) instructions to ask questions of each other. In only a few instances (e.g., Nowak, Miller, & Washburn, 1996; Webb, 1993; Fuchs et al., 1998) did participants receive more extended instruction and practice. Based on the results of their study, Fuchs et al. (1998) concluded that group training is essential for effective collaboration.

But even comprehensive pretraining in group interaction does not guarantee that high quality interaction will take place. To investigate this issue, Hite (1996) and Yokomoto and Ware (1997) had groups rate their quality of interaction and/or the contribution of group members. These ratings were used as motivators and were also sometimes factored into grades. Other researchers have video- or audio-taped discussions to identify categories of communication and patterns of decision making (e.g., Castor, 2004; Ewald, 2005). Using audio-recorded discussions of seventh grade mathematics students, Webb (1993) concluded that the "profile" of group interaction (for example, making an effort to understand explanations rather than simply copying answers) was a better predictor of students' achievement than their prior performance. Analyses of video-taped group discussions in eighth grade science classes (Webb et al., 1998) indicated that the quality of group interactions correlated well with the performance of below-average ability students, but not with that of above-average ability students who usually perform well in most situations.

Jensen, Moore, and Hatch (2005) studied both quantitative and qualitative aspects of the "chat" taking place during computerized quizzes. They found that students who were graded as a group versus individually not only scored better, but their communications were also more extensive and cooperative. Using a posttest survey, Kapitanoff (2009) found that high scores earned under the collaborative condition were related to students' perceptions of the quality of their groups' discussions. Rating collaborative groups' quality of interaction on a three-point scale, Pandey and Kapitanoff (2011) found that students with higher interaction scores earned better scores on collaborative exams than did those with lower interaction scores. Interaction scores also correlated with lower test anxiety at both pre- and posttesting. All these studies point to the importance of the quality of group discussions as a critical factor in successful collaborative testing.

1.3. ANXIETY

One of the benefits most often claimed to result from collaborative testing is anxiety reduction (e.g., Johnson, Johnson, & Smith, 1991). It is possible that women, who are viewed by many as inherently weaker in mathematics than men, experience more anxiety than men in statistics classes because of stereotype threat—"being at risk of confirming, as self-characteristic, a negative stereotype of one's group" (Steele & Aronson, 1995, p. 797). Stereotype threat has been shown to significantly decrease the performance of those who belong to stereotyped groups including women's performance in mathematics (e.g., Marx & Roman, 2002; Stout, Dasgupta, Hunsinger, & McManus, 2011). Collaborative testing might, therefore, be especially useful for women.

Most of the collaborative testing studies which have examined the moderating effect of anxiety have used posttest surveys given either after each exam (e.g., Hendrickson, Brady, & Algozzine, 1987) or once at the end of the term (e.g., Phillips, 1988). Breedlove, Burkett, and Winfield (2004) had students complete a test anxiety scale just before taking two tests at different times. Both were either collaboratively- or individually-taken. Although they found no significant differences in test anxiety between the two conditions, these researchers discovered that collaborative students were less likely to experience an increase in test anxiety from the first to the second exam.

Using the Alpert-Haber Achievement Anxiety Test, Hanshaw (1982) found that its "test-debilitating" subscale was negatively correlated with collaborative testing performance. Employing Spielberger's State-Trait and Test Anxiety Inventories, Pandey and Kapitanoff (2011) found that collaborative testing reduced test anxiety in high, but not low, test-anxious students. These students' lowered anxiety was associated with higher quality-of-interaction scores and improved exam performance. Additionally, positive correlations were found between the amount of benefit derived from collaborative testing and pretest trait and test anxiety scores. Using the 10-item Mathematics

Anxiety Scale (Betz, 1978), Townsend, Moore, Tuck, and Wilton (1998) found, to the contrary, that group work did not reduce mathematics anxiety. One difficulty in summarizing the research in this area is that a variety of anxiety measures have been used. As different types of anxiety may reflect differing underlying processes, this may explain some of the disparate results noted above. It is important to examine which type of anxiety is most related to performance under collaborative testing conditions.

1.4. GROUP FORMATION

Differences in outcomes may also be the result of how groups were formed and testing was administered, factors that could affect the quality and quantity of interactions in as yet unspecified ways. Group size has ranged from pairs (e.g., Ley, Hodges, & Young, 1995) to as many as seven members (e.g., Castor, 2004). The groups themselves have been self-selected (e.g., Zimbardo, Butler, & Wolfe, 2003), assigned by the instructor on some basis such as past performance (e.g., Helmericks, 1993), or entirely at random (e.g., Giraud & Enders, 1998). Collaborative groups have sometimes been changed for each exam (e.g., Helmericks, 1993) and, at other times, held constant throughout the term (e.g., Giraud & Enders, 1998). Groupings have been announced prior to exams (e.g., Helmericks, 1993) or just before testing began (e.g., Pandey & Kapitanoff, 2011).

In some studies (e.g., Stearns, 1996), students have first taken an exam individually and then retaken all or part of it in groups (the two-stage method). In others, collaborative group exam scores have been compared to the scores of students who took the same exam individually in previous semesters (e.g., Helmericks, 1993), with a control group that took the exam concurrently but individually (e.g., Breedlove et al., 2004), or with their own scores on individually-taken exams (e.g., Hanshaw, 1982). Group members have either been required to reach consensus (e.g., Helmericks, 1993) or allowed to submit their own, possibly differing, answer forms (e.g., Pandey & Kapitanoff, 2011). Collaborative testing has been used from only once during the semester to continuously throughout the term. Instructors have also given differing weights to the impact of collaborative scores on overall course grades. Finally, in studies using the two-stage design, the intervals of time between the collaborative testing experience and subsequent individual re-testing have varied (e.g., Ives, 2014).

2. PLAN OF THE CURRENT STUDY

The current study was designed to investigate the role played by the quantity and quality of interaction in collaborative group testing sessions on performance outcomes in statistics classes and to identify which type(s) of anxiety correlate most with benefits from collaborative testing. The specific goals of this study were to investigate

1. whether taking midterm exams collaboratively as compared to individually improves scores on these exams as well as the individually-taken final exam;
2. whether the number of collaborative exams taken is related to scores on subsequent midterm exams or the individually-taken final exam;
3. which type of student (age, sex, etc.) benefits most from collaborative testing;
4. which types of pretest measures of anxiety: trait, test, mathematics, and specific-class-related anxiety, are related to benefits derived from the collaborative process;
5. whether mathematics anxiety specifically is reduced as a result of collaborative testing;
6. how the quantity and quality of collaborative group interactions relate to performance; and
7. how students' recollections of the level of test anxiety they experienced under the two conditions compare and relate to performance.

3. METHOD

3.1. PARTICIPANTS

At the beginning of the semester, the concept of collaborative testing was explained to the 85 students in two introductory Statistics for the Social Sciences classes (both taught using the same lectures and materials by one of the researchers) offered at a large urban U.S. community college. Students not wishing to participate in the study were told that they would receive help in enrolling in one of the eight other sections of the course offered throughout the day and evening. To ensure that

students' grades would not be negatively affected if they were not selected for collaborative testing as often as others, it was written in the syllabus and described in class that their final class grade would be calculated in two ways, using the scores of their tests taken individually and taken collaboratively. Whichever result was higher would be the percentage used in assigning the class grade. All 85 students (58% female, 42% male) enrolled in the classes agreed to participate in the study. Their mean age was 22.78 years (*SD* = 8.94). They had completed a mean of 34.93 credits (*SD* = 25.57) and had earned a mean GPA of 2.7 (*SD* = 0.63) on a scale of 0 to 4.

All 85 students completed the pretest measures. However, as is common in this community college, a number of students did not remain in the class by the time of the first exam. The total number of students used in this analysis was 71. A comparison of the demographic variables, (GPAs, units completed, age, or any anxiety measure) indicated no significant difference between those individuals who dropped and the students who remained in the class. Thus, 14 students (16.5% of the sample) took all pretest measures but dropped the class before taking the first midterm exam. Another fifteen students (17.5% of the original sample) took at least one exam but eventually dropped the class before taking the final examination. Of these, three had taken only one midterm, nine had taken two, one had taken three, one had taken five, and one had taken six. Together, these students had taken 19 of their midterms individually and 16 collaboratively.

3.2. INSTRUMENTS

Pretest To measure Specific-Class-Related Anxiety, participants responded to a Likert-type item rating their anxiety about taking this specific class from 1 (*very anxious*) to 7 (*not at all anxious*). Test, trait, and mathematics anxiety were measured using Spielberger's Test Anxiety Inventory (TAI) (Spielberger et al., 1980), Spielberger's State-Trait Anxiety Inventory for Adults (STAI) (Spielberger, Gorsuch, Lushene, Vagg, & Jacobs, 1983), and the Abbreviated Mathematics Anxiety Scale (AMAS) (Hopko, Mahadevan, Bare, & Hunt, 2003), respectively. Both the TAI and STAI consist of 20 statements to which the student answers "almost never," "sometimes," "often," or "almost always." A sample item from the STAI (Cronbach $\alpha = 0.86$) is "I worry too much over something that really doesn't matter" and from the TAI (Cronbach $\alpha = 0.96$) "I feel confident and relaxed while taking tests." The AMAS (Cronbach $\alpha = 0.90$) consists of nine situations such as "Taking an exam in a mathematics course" which the student rates as causing 1 (*low*) to 5 (*high*) anxiety.

It was decided to use a mathematics anxiety scale rather than a statistics anxiety scale because students tend to think of statistics as a form of mathematics (e.g., Baloglu, 2004), and previous studies have found performance in statistics classes to be negatively correlated with scores on mathematics anxiety tests (e.g., Morris, Kellaway, & Smith, 1978; Zeidner, 1991). Because anxiety measurement was done at the very beginning of the semester before students had the opportunity to discover that statistics is different from mathematics, it was thought that their fear of mathematics would more accurately measure their initial level of anxiety about the subject.

Posttest The posttest consisted of a brief survey asking whether the participants would have preferred taking the final exam collaboratively and to rate the overall collaborative testing experience on a 7-point Likert-type scale ranging from 1 (*very negative*) to 7 (*very positive*).

3.3. PROCEDURE

After completing consent forms, the pretest survey, and the TAI, STAI, and AMAS, students took part in training which included practice working collaboratively with a partner on a challenging mathematics problem. Students were told to discuss the problem together, to show respect for each other, and not to interrupt while the other was speaking. Next, it was explained that when they entered the classroom to take each of the six midterm multiple-choice exams, it would be determined randomly whether they would be taking that particular exam individually or with a randomly selected partner or, when the total number of collaborative participants was odd, two partners. Thus, the number of exams taken collaboratively by any one student could vary from none to all six. This design allowed for a greater range of the possible number of collaborative exams taken, thus yielding a broader analysis of the relation between number of exams taken and outcomes. Because the chance of taking

any one exam collaboratively was always 50%, students were strongly advised to prepare for each exam as if they would be taking it individually. They were told they would submit their own individual answer sheets and that group consensus was not required. Because collaborative group interactions were expected to increase the time needed to complete an exam, testing periods were extended equally for both conditions. Those taking an exam individually remained in the classroom, while those assigned to take an exam collaboratively moved to a nearby room where the exam was proctored by two assistants.

Experience in teaching this class showed that students preferred to have more exams covering fewer chapters rather than fewer exams covering more chapters. The first exam (which covered fewer chapters than any of the others) had 50 multiple-choice questions. The remaining five exams had 70 questions each. The final exam consisted of 120 multiple-choice questions. Students were allowed 85 minutes for each of the six midterm exams, but they could submit their exams and leave the classroom sooner if they finished to their satisfaction. The final exam was scheduled for two hours but again, students could leave early if their exams were completed.

To measure the quantity of interaction within collaborative groups, every 3.5 minutes, two assistants individually recorded whether each group was currently interacting. Immediately after a group submitted their answer sheets, the group members filled out an evaluation form designed to measure the quality of interaction. On this form, students rated their own and their partner's contribution to the discussion and test performance on a 7-point Likert-type scale going from 1(*extremely harmful*) to 7 (*extremely helpful*). They also indicated the particular way(s) in which the partner had facilitated (e.g., "*filled in gaps in my knowledge*") and/or hindered (e.g., "*wasted my time*") their exam performance (see Table 5 for the checklist they were given). Students were told that their responses to these posttest evaluations would be kept confidential and would not affect their partners' grades. Aside from the collaborative testing, the classes were taught in a traditional lecture style, and students were not involved in any other form of organized collaborative learning.

Upon finishing the individually-taken comprehensive multiple-choice final exam, students responded to the posttest survey, retook the AMAS and completed the TAI twice, responding on the basis of how they had felt while taking exams collaboratively and individually. The sequence of these last two measurements was counterbalanced. Trait anxiety was not measured posttest because it was assumed that such a basic characteristic would not be altered by this brief and limited intervention. Specific-Class-Related Anxiety was also not measured posttest because it was thought to measure anticipation felt at the beginning of the class.

4. RESULTS

On the pretest survey, Specific-Class-Related Anxiety was measured using a scale of 1 to 7, 1 indicating (*not at all anxious*) and 7 indicating (*very anxious*). Students typically reported a medium level of Specific-Class-Related Anxiety (M = 4.14, SD = 1.80, n = 71). The distribution, which had a mode of 5.0, was negatively skewed, with the mean for women being significantly higher than that for men, women: M = 4.52, SD = 1.80, n = 40, and men: M = 3.65, SD = 1.70, n = 31; t (69) = 2.092, p = 0.04, d = 0.503. (Note: Effect sizes for mean differences reported throughout the Results section are Cohen's d.) The mean pretest STAI score which can range from 20 to 80, was 38.13 (SD = 9.71, n = 31) for the men and 37.51 (SD = 9.04, n = 37) for the women. This difference was not significant. The mean pretest TAI score which also can range from 20 to 80, was 36.65 (SD = 11.69, n = 31) for the men and 46.21 (SD = 15.68, n = 39) for the women, t (67.76) = 2.921, p = 0.005, d = 0.691 (estimated assuming unequal variances for men and women). The mean AMAS score, which can range from 9 to 45, was 20.74 (SD = 56.06, n = 31) for the men and 24.63 (SD = 7.53, n = 40) for the women, t (69) = 2.342, p = 0.022, d = 0.569. All these means were close to the published norms for college students and indicated the typical pattern of higher anxiety for women.

The three pretest standardized anxiety measures as well as Specific-Class-Related Anxiety were significantly inter-correlated. However, there was no relationship found between any of them and credits completed or cumulative GPA for the sexes combined. There was a significant correlation between age and Specific-Class-Related Anxiety, r = 0.372, n = 71, p = 0.001. Examining sex of student and anxiety, there were significant negative correlations for men between GPA and pretest Test Anxiety (r = -0.426, n = 28, p = 0.024) and GPA and Trait Anxiety (r = -0.522, n = 28, p = 0.004)

suggesting that male students with better academic records had less anxiety. For women, there was no relationship between GPA and the standardized anxiety measures. But, for women, there were positive correlations between age and credits completed ($r = 0.407$, $n = 44$, $p = 0.006$), age and Specific-Class-Related Anxiety ($r = 0.393$, $n = 39$, $p = 0.013$), and a negative correlation between credits completed and Trait Anxiety ($r = -0.436$, $n = 36$, $p = 0.008$). Thus, for women, anxiety was related to age and number of previous classes, but not to prior academic achievement. A per-test Type I Error rate of 0.004 was used to adjust for the large number of correlational tests performed. This more conservative approach led to a single significant correlation, that between GPA and Trait Anxiety in males.

4.1. MIDTERM EXAM PERFORMANCE

Students in the collaborative condition were assigned to pairs randomly before the class began. However, they did not know whether they were assigned to the collaborative condition or with whom they would be working until they entered the room. If a student who was supposed to be in the collaborative condition was absent, one student originally assigned to the individual condition was reassigned to the collaborative condition in order to complete the pair. As a result, more students ended up in the collaborative than individual condition. Despite the fact that the procedure used to assign students to conditions inadvertently resulted in unequal numbers, comparisons between the two conditions are valid because assignment remained random.

Of the 71 students who took at least one midterm exam, 59 took one or more exams under each of the two conditions. The percentage of students who took from zero to six exams collaboratively was 10%, 18.6%, 8.6%, 24.3%, 27.1%, 8.6%, and 2.9% respectively. Overall, 46 students earned higher scores under the collaborative condition, whereas 12 did better in the individual condition, and one did equally well under both. A chi-square test ("no preference" null hypothesis) calculated for the 58 students who did better in one condition than the other was significant ($X^2(1) = 19.93$, $p < 0.001$). Comparing the mean scores for the 59 students who took at least one exam collaboratively and one exam individually, the mean of the test scores earned under collaborative testing (65.38%, SD = 14.98%) was significantly higher than the mean score earned under individual testing (54.79%, SD = 15.91%), $t(58) = 5.06$, $p < 0.001$, $d = 0.685$. This difference of 9.9 percentage points between the groups actually constitutes an 18% differential over the individual condition group, which is within the range found in the literature.

The data for the midterm exams were analyzed as a set of six separate replications with each exam as one replicate. A Type I error rate of 0.05 was used in the analyses. For each of the exams, there were no differences in students' pretest scores for Mathematics Anxiety, Test Anxiety, Trait Anxiety, or Specific-Class-Related Anxiety, age, cumulative GPA, number of credits completed and sex between those in the collaborative and individual exam groups.

Although the students had been told that consensus was not required and that they would turn in their exams separately, scores for the collaborative group were not completely independent. Twenty-seven percent of the time, scores on the six midterm exams were the same, the percentage of groups earning the same sores being 22%, 26%, 20%, 28.5%, 23% and 42.8%, respectively. Because of this lack of total independence, the scores of the group members were averaged and used as the single score for the group. They were then compared with those taking the exam individually, thus more accurately reflecting the degrees of freedom for this comparison. Therefore, the sample size for the collaborative exams in Table 1 represents the number of groups.

The collaborative mean scores were higher than the individual mean scores for all six exams, with four of the differences being statistically significant. Using a Bonferroni correction for multiple tests, the per-test Type I Error rate was set to 0.008, resulting in three of the differences reaching significance. For both conditions, the means tended to decrease over the course of the semester as usually happens as the material becomes more difficult, see Table 1.

Previous research suggests that collaborative testing might be particularly helpful for students who have higher levels of anxiety, lower GPAs, or are female. No interaction was found between pretest measures of anxiety, GPA, or sex with exam score when each exam was analyzed separately.

Table 1. Midterm exam scores: Mean percentage correct

	Collaborative exam		Individual exam				
Exam	*n*	Mean (*SD*)	*n*	Mean (*SD*)	*t*-test (pooled)	df	*p*
1	16	82.54 (10.18)	32	69.28 (15.72)	3.058	46	.002*
2	18	68.65 (16.68)	28	63.37 (15.10)	1.111	44	.136
3	15	66.16 (16.32)	35	50.94 (20.00)	2.596	48	.006*
4	14	59.49 (14.88)	29	56.60 (16.43)	0.557	41	.290
5	13	56.28 (14.48)	29	47.11 (14.43)	1.902	40	.032
6	14	69.76 (14.13)	27	52.54 (22.90)	2.563	39	.007*

Note. The sample size for Collaborative Exams represents the number of groups; the sample size for Individual Exams represents the number of individuals. Because previous research has established that collaborative testing produces higher test results than individual testing, a one-tailed *t*-test was used.

*Using a Bonferroni correction for multiple tests, the Type I error rate was set to 0.008.

To see whether more experience with collaborative testing would lead to better performance on succeeding collaborative exams, scores on each collaboratively-taken exam were correlated with the number of such exams each student had taken to that point. There was no correlation found between the overall number of collaborative exams taken by a student and the mean score for all their collaboratively-taken exams combined. Examining each exam separately, there was also no significant trend in correlations between the number of exams taken collaboratively to that point and exam score.

4.2. FINAL EXAM PERFORMANCE

The scores on the individually-taken cumulative final exam ranged from 22% to 77% with a mean of 50.88% (*SD* = 12.65%, *n* = 56), testifying to the difficulty of this course for students. The correlation between the number of collaborative tests taken and the final score was nearly zero, suggesting that the benefits of collaborative testing did not transfer to the cumulative final. (This conclusion is qualified by the analysis of the quantity and quality of interaction in the Group Interaction section below.) Final exam scores were positively correlated with total scores earned on both the individually-taken, $r = 0.740$, $n = 53$, $p < 0.001$, and collaboratively-taken exams, $r = 0.617$, $n = 54$, $p < 0.001$, but negatively correlated with pretest Specific-Class-Related Anxiety, $r = -0.347$, $n = 56$, $p = 0.009$.

Sixteen-and-a-half percent of those who had taken the first exam did not complete the final, an attrition rate typical for this course at this college. Of the 56 students who took the final, 41 (75.5%) said they would have preferred to take it collaboratively. These students tended to do more poorly on the final compared with those who did not prefer to take it collaboratively, earning a mean score of 48.81% (*SD* = 11.82%), compared to 55.56% (*SD* = 14.30%). This difference was not significant. There were no differences in pretest measures between students who preferred to take the final exam collaboratively and those who did not. However, unlike the 14 students who preferred taking the final exam individually, the students who would have preferred taking it collaboratively had scored significantly better on collaboratively-taken (*M* = 65.63%, *SD* = 14.19%) than on individually-taken (*M* = 52.66%, *SD* = 14.45%) midterm exams, $t(39) = 5.72$, $p < 0.001$, $d = 0.951$.

4.3. RATINGS OF THE COLLABORATIVE EXPERIENCE

Student ratings of the overall collaborative experience on a scale of 1 to 7 yielded a mean rating of 5.35 (*SD* = 1.60, *n* = 54). This evaluation was negatively correlated with age, $r = -0.305$, $n = 54$, $p = 0.025$, and positively correlated with recalled Test Anxiety when tested individually, $r = 0.284$, $n = 51$, $p = 0.044$, but not when tested collaboratively. It appears that students who were more anxious when tested individually favored the collaborative experience most highly.

Among all participants, there was a significant correlation between ratings of the collaborative experience and mean collaborative score, $r = 0.292$, $n = 54$, $p = 0.032$. There was also a significant difference in ratings of the collaborative experience between those who did (*M* = 6.00, *SD* = 0.95, *n* = 41) and did not (*M* = 3.31, *SD* = 1.55, *n* = 13) prefer to take the final collaboratively, $t(52) = 7.579$, $p < 0.001$, $d = 2.093$.

4.4. GROUP INTERACTION

Quantity of interaction Quantity of interaction was rated at set intervals by two raters. A student's quantity of interaction score was the percentage of observations in which some interaction was recorded. Inter-rater reliability was significantly correlated except for the first exam (see Table 2). Although the inter-rater reliability for the sixth exam ($r = 0.415$) was statistically significant, it did not approach the customarily accepted value for inter-rater reliability which is usually 0.7 or above (Cohen, 1988). Although most ratings between the two observers were quite similar, a small number of ratings were very different. To prevent the analyses from being distorted by these questionable ratings, *disparate ratings* were operationally defined as those where the rating given by one observer was more than twice the other. These disparate pairs of ratings (19 out of 189, or 10.05%) were considered unreliable and therefore not included in the analysis. This resulted in above criterion inter-rater reliability for Exams 2 through 6 and much improved inter-rater reliability for Exam 1, see Table 2.

Table 2. Inter-Rater reliability of quantity of interaction with and without disparate scores

	All scores		With disparate scores removed	
Exam	*n*	Reliability correlation	*n*	Reliability correlation
1	35	.195	28	.590
2	38	.779	38	.779
3	31	.904	27	.928
4	26	.858	24	.821
5	29	.867	22	.868
6	30	.415	24	.792

Using non-disparate data only, in other words where the ratings given by one observer did not differ from that given by the other observer by a factor of 2 or more, a significant positive correlation was found between students' mean quantity of interaction scores and ratings of the overall experience, $r = 0.391$, $n = 53$, $p = 0.004$ and a negative correlation with the number of tests taken collaboratively, $r = -0.350$, $n = 61$, $p = 0.006$.

For men and women combined, quantity of interaction for Exam 6, was related to performance, $r = 0.388$, $n = 30$, $p = 0.034$. and the final exam, $r = 0.289$, $n = 53$, $p = 0.036$. These correlations were not significant when the sexes were examined separately. But, when each sex was divided into those with above- and below-mean (mean for their sex) quantity of interaction scores, among the women only, the two groups differed significantly on their mean final exam scores. Women with above mean quantity of interaction scores earned higher final exam scores ($M = 53.53\%$, $SD = 8.28\%$, $n = 15$) than did the below-mean women ($M = 44.84\%$, $SD = 11.11\%$, $n = 18$), $t(29) = 2.37$, $p = 0.024$, $d = 0.887$.

Students' GPAs also interacted with quantity of interaction and final exam scores. Among men and women combined, those with below-mean GPAs (GPA < 2.727 on a 0 to 4-point scale) showed significant correlations between quantity of interaction and both total collaborative scores ($r = 0.417$, $n = 28$, $p = 0.027$) and final exam scores ($r = 0.576$, $n = 24$, $p = 0.003$). No such correlations were found for those with above-mean GPAs. The amount of discussion engaged in by high-GPA students did not relate to their final exam scores. But among lower GPA students, final exam performance was different for those above (Final $M = 50.78\%$, $SD = 10.86\%$, $n = 13$) and below (Final $M = 38.38\%$, $SD = 8.97\%$, $n = 11$) the mean on quantity of interaction, $t(22) = 3.06$, $p = 0.006$, $d = 1.245$.

Quality of interaction Three measures of the quality of interaction were used based on a 7-point scale from 1 (*extremely harmful*) to 7 (*extremely helpful*). For a pair of collaborating students consisting of discussants "student" and "partner," one measure of quality of interaction was partner-rating-of-student. Another was self-rating, and the third was student-rating-of-partner. For triads, partners-ratings-of-student were averaged. The mean partner-ratings-of-student averaged over all six midterm exams was 5.70 ($SD = 1.23$, $n = 63$). The mean of the self-ratings was 5.59 ($SD = 0.92$, $n = 62$), and the mean of the student-ratings-of-partner was 5.88 ($SD = 0.87$, $n = 63$). For all three ratings, students judged interactions positively, between "slightly" and "moderately helpful." The measures of mean quality and quantity of interaction were generally highly inter-correlated (see Table 3). It is

recognized that the mean of student-ratings-of-partner and partner-ratings-of-student would be equal when all individuals are designated as "student." However, to tease out relationships between partners, we arbitrarily designated one student as the "student" and the other as the" partner" and looked at the relationships between them in one direction only.

Table 3. Inter-correlations of mean measures of interaction for the six midterm exams combined

	1	2	3	4
1. Mean quantity of interaction	1.000	0.172	0.440**	0.554**
2. Mean quality of interaction, self-rating		1.000	0.219	0.291*
3. Mean quality of interaction, student-ratings-of-partner			1.000	0.279*
4. Mean quality of interaction,				1.000

*two-tailed $p < 0.05$

**two-tailed $p < 0.01$

For all students examined together, the mean quantity of interaction was related to quality of interaction as measured by student-ratings-of-partners, $r = 0.440$, $n = 61$, $p < 0.001$, and partner-ratings-of-students, $r = 0.544$, $n = 61$, $p < 0.001$. Student-ratings-of-partner were significantly correlated with partner-ratings-of student, $r = 0.279$, $n = 63$, $p = 0.027$, suggesting that when there is a satisfying interaction, it is judged that way by both partners. Self-ratings were positively correlated with partner ratings-of-student, $r = 0.291$, $n = 62$, $p = 0.022$, and negatively correlated with recalled Test Anxiety under the individual condition, $r = -0.329$, $n = 50$, $p = 0.020$. Partner-ratings-of-student were related to performance on Exam 6, $r = 0.451$, $n = 31$, $p = 0.011$, and tended to be related to the student's performance on the final exam, $r = 0.262$, $n = 54$, $p = 0.056$. Several correlations were significant for women only. Among women, partner-ratings-of-student were correlated with performance on collaboratively-taken midterm exams, $r = 0.395$, $n = 34$, $p = 0.021$, and performance on the final exam, $r = 0.397$, $n = 31$, $p = 0.027$. Women's self-ratings were also correlated with final exam scores, $r = 0.370$, $n = 30$, $p = 0.044$.

Measures of student-rated quality and observer-rated quantity of interaction were assessed for each exam separately, with sexes combined. Some trends emerged in these associations. For five of the exams, quantity of interaction was significantly related to partner-ratings-of-student; for three of the exams, quantity of interaction was related to student-ratings-of-partner; and for three exams, quantity of interaction was related to self-ratings. The number of individuals responding for each exam differed depending upon which students were selected for the collaborative condition (see Table 4).

Table 4. Relationship of quantity and quality of collaboration by exam

		Quantity and partner ratings-of-student		Quantity and student-ratings-of-partner		Quantity and self-ratings	
Exam	*n*	*r*	*p*	*r*	*p*	*r*	*p*
1	28	0.498**	0.007	0.230	0.240	0.328	0.095
2	38	0.455**	0.004	0.554**	0.000	0.403*	0.013
3	27	0.592**	0.001	0.299	0.129	-0.265	0.181
4	24	0.593**	0.002	0.444*	0.030	0.165	0.441
5	22	0.744**	0.000	0.165	0.464	0.660**	0.001
6	24	0.342	0.102	0.598**	0.002	0.405*	0.050

*two-tailed $p < 0.05$

**two-tailed $p < 0.01$

Ratings of the Collaborative Experience Students were asked after each exam to indicate on a checklist in what ways their partners had helped or detracted from the collaborative experience. Many students found the collaboration to be helpful in processing information, for example, filling in gaps in knowledge or helping them think through the information better. The emotional benefits such as making the experience "fun" and relaxing were less consistent. A minority of students reported some negative

reactions: that their partners were distracting, that the process was a waste of time, it made me nervous, and it made me feel taken advantage of (see Table 5).

Table 5. Student ratings of the collaborative experience

	Percentage of students agreeing with each statement					
My partner…	Exam 1	Exam 2	Exam 3	Exam 4	Exam 5	Exam 6
helped me think through information better	70.3	81.6	66.7	71.9	58.6	62.5
helped me remember information I had forgotten	64.9	78.9	54.5	56.3	58.6	65.6
filled in gaps in my knowledge	70.3	63.2	72.7	59.4	55.2	56.3
helped me understand what was being asked	67.6	76.3	54.5	53.1	55.2	53.1
helped increase my confidence	67.6	57.9	42.4	53.1	48.3	53.1
helped me understand why answers were correct	51.4	55.3	48.5	46.9	41.4	34.4
turned testing into a learning experience	48.6	52.6	48.5	34.4	37.9	37.5
made it fun and relaxing	67.6	55.3	33.3	28.1	41.4	37.5
made me feel nervous	2.7	10.5	18.2	6.3	10.3	12.5
was distracting	5.4	13.2	9.1	12.5	10.3	12.9
made me feel taken advantage of	8.1	7.9	15.2	6.3	10.3	9.7
wasted my time	2.7	2.6	9.1	0.0	13.8	9.7

4.5. POSTTEST MEASURES OF ANXIETY

For the sample as a whole, as well as for males and females examined separately, Mathematics Anxiety remained unchanged from pretesting to posttesting. To avoid the distorting effects of frequent retesting, the amount of Test Anxiety experienced during collaborative and individual exams was measured only once, at the end of the semester rather than after each exam. Recalled Test Anxiety under the collaborative condition (M = 37.08, SD = 12.59, n = 51) was significantly lower than under the individual condition (M = 43.88, SD = 16.74, n = 51), paired $t(50) = 4.50$, $p < 0.001$, $d = 1.134$. Students' posttest recalled Test Anxiety under the collaborative condition (M = 36.87, SD = 12.48, n = 53) was also significantly lower than their pretest Test Anxiety (M = 42.32, SD = 15.21, n = 53), paired $t(52) = 3.25$, $p = 0.002$, $d = 0.392$. This decrease was not found for the individually-taken exam condition.

5. DISCUSSION

5.1. GENERAL FINDINGS

The benefits of collaborative testing were confirmed in this study using six replicates of classroom exams. The results suggest that collaboration is particularly beneficial for students with lower GPAs and high test anxiety and are similar to Giuliodori, Lujan, and DeCarlo's (2008) and Dahlström's (2012) findings that the "collaborative testing effect" was greater for low-performing students than for high-performing students. They can be contrasted, however, with the conclusions drawn by Fuchs et al. (1998) that collaboration might benefit above-average students more than their at- or below-grade level peers. They also differ in part from the results of Gilley and Clarkston (2014) who, when working with undergraduates taking a course in Earth and Ocean Science, found that the benefits of collaborative testing on subsequent individual testing were the same regardless of the students' initial individual performance. These inconsistencies may be due to any number of factors including the possibility of a ceiling effect when lower-performing students do better under collaborative conditions, none of which can be determined using the available data. Although many analyses were performed in this study,

which may concern some readers, results supporting the following findings were clear and consistent: 1) the difference in scores between collaboratively- and individually-taken tests, and 2) for women, the relationships between quality of interaction and performance.

As found in some other studies, collaboration did not necessarily relate to better performance on the individually-taken cumulative final exam. Simply being offered the opportunity to discuss exam questions in a group does not guarantee that productive discussions will take place. The variables that appeared to be relevant to positive transfer among our participants were the quantity and quality of group discussions. That a high level of engagement is one of the factors that make collaboration beneficial is suggested by the positive relationship between partner-ratings-of-student and final exam performance.

Two other factors which might account for the lack of positive transfer of learning from the midterms to the final exam in this study are the low level of intrinsic motivation in the learners and the difficulty of the subject. First, an involved student might benefit more from collaboration than others for whom the course is of little interest. The overwhelming majority of students in the current study were not social science majors. Many of them took this course to fulfill a non-social science requirement such as critical thinking rather than as a subject judged to be important to their future academic and professional careers. Second, the observers in this study noted that at times the "interaction" taking place in groups amounted to one student completing a page of the exam and then giving it to his/her partner for copying. As noted above, this occurred for a minority of students. Compared to the excitement and high levels of energy and enthusiasm reported by many instructors in their studies of collaboration, the energy levels apparent in the statistics groups were sometimes low and subdued. This is most likely due to the difficulty of the subject for most students. Even if students are not very well prepared, they can still contribute something to the discussion in a class such as introductory psychology because many of its topics relate to things they have experienced in their own lives. But mathematics, in general, and statistics, in particular, may be so challenging for some students that they cannot talk about it easily. Despite these reservations, however, these data are important in understanding how collaboration helps students in real world, typical community college classrooms.

One unexpected finding of this study was that as the number of exams taken collaboratively increased, the quantity of interaction decreased. It is possible that although the overall quantity of interaction declined, some aspect(s) of the discussion might have increased. For example, the discussion may have become more focused. Although the two participants tended to agree on the quality of their interactions, it is not clear what aspects of those interactions underlay these ratings. A summary of the ratings of the benefits of discussions indicates that many students found collaboration helpful in understanding the material. For some, collaboration made the experience more relaxing and increased their confidence.

The amount of interaction engaged in by students with above-mean GPAs probably did not correlate with their final exam scores because, being better students, they were prepared to do well regardless of the amount of their collaboration. The below-mean GPA students probably needed and, therefore, benefitted more from collaborative discussions. For these students, the more they participated, the better they did.

Interesting differences between the sexes emerged in this research. These are important to understand as they may contribute to the persistent underrepresentation of women in mathematics-related fields. The women in this study scored higher on all the measures of anxiety except Trait Anxiety, the type of anxiety that proved to be least related to performance. Unlike the higher-GPA men who showed less anxiety than their lower-GPA peers, higher-GPA women showed as much anxiety as lower-GPA women.

Both the quantity and quality of collaborative interactions had a greater impact on women's performance than men's. Women who interacted more during collaborative testing earned higher final exam scores. Those who were rated as good discussants by their partners earned higher scores on the final exam and the midterm exams they had taken collaboratively than had the women who had earned poorer partner-ratings. And women's, but not men's, self-ratings of their quality of interaction also correlated with their final exam scores.

Previous research has been inconsistent in determining which type of anxiety is related to collaborative testing. Although the types of anxiety examined in this study were intercorrelated, they were differentially related to outcomes. Trait Anxiety correlated with a few demographic variables but

with neither overall performance nor differences in performance between the two conditions. Similar to Trait Anxiety, Mathematics Anxiety did not relate to any performance variables or change from pre- to posttesting for all students combined. Specific-Class-Related Anxiety was correlated with some demographic variables but, more importantly, was negatively correlated with final exam scores for all students combined.

Besides being related to demographic variables, as was true of other forms of anxiety, Test anxiety turned out to be more related to performance outcomes than did the other types of anxiety. The amount of test anxiety students recalled experiencing under the collaborative condition was lower than their initial pretest levels. It was also less than recalled test anxiety under the individual condition. Recalled test anxiety under the individual condition was positively correlated with overall ratings of the collaborative process but negatively correlated with self-ratings of the quality of discussion. Thus, it appears that test anxiety is the most relevant anxiety measure for studies of collaborative testing.

5.2. LIMITATIONS OF THE CURRENT STUDY AND RECOMMENDATIONS FOR FUTURE RESEARCH

Johnson, Johnson, and Smith (1991) postulate that a key feature of collaborative testing is *mutual interdependency*, that is, students' grades are raised or lowered depending upon how well their partners do. Future research could extend the current study by introducing *mutual interdependency* as a second independent variable.

Although this study focused on students' perceptions of the quality of their interactions, stronger support for the relationships uncovered in this study might have resulted if the quality of interaction had been rated using analyses of recorded session transcripts rather than subjective, retrospective reports by the participants.

Future work in this area might include the measurement of students' statistics anxiety, attitudes towards statistics, and/or approaches to its learning, as well as mathematics anxiety as was done here. Several standardized tests of statistics anxiety are available, including the SAS (Pretorius & Norman, 1992) and the STARS (Cruise, Cash, & Bolton, 1985). Four surveys of students' attitudes towards statistics which they believe have been demonstrated to have adequate validity and reliability are recommended by Nolan, Beran, and Hecker (2012). Identifying students' approaches to learning (Deep, Surface, and Strategic) using the abbreviated Approaches and Study Skills Inventory for Students (ASSIST; Tait, Entwistle, & McCune, 1998) might also be valuable as this scale has recently been found to be invariant across different languages and educational contexts (Chiesi et al., 2016).

A question that was not addressed in the current study is, what is the optimal collaborative group size? A survey of past studies relating the amount of benefit from collaboration with the size of collaborative groups used in that study could possibly answer this question. Another approach would be to conduct a new study with group size as one of the independent variables. It is possible, of course, that optimal group size would vary with the age of the students, the sexual composition of the groups, and the difficulty of the subject.

Whereas the current study examined whether the number of midterm exams taken collaboratively correlated with cumulative final exam performance, it might be informative to examine whether the sequence of taking the exams collaboratively was influential. Because the *number* of exams students took collaboratively differed as well as *when* in the series of six tests they took them, the benefits of collaboration might have been affected by the interaction between number and sequencing. One way to address this issue in future studies would be to have all students take an equal number of midterm exams individually and collaboratively, but vary the sequencing of the exams according to a preplanned schedule. For example, one student might take his/her collaborative exams at the beginning of the semester, another in the second half of the semester and another interspersed throughout the semester. The sequences would be randomly assigned to students who would not know ahead of time which particular exam they would be taking collaboratively or individually.

Although there were more students who took collaborative than individual exams, there was no systematic bias in the assignment of student to conditions. The students were assigned randomly before the testing period began. A preferable approach might have been to assign students to conditions as they walked in the door, thus assuring a more equal distribution. The issue of multiple tests—increasing the probability of Type I error—is recognized. However, as an exploratory study, all hypotheses of

interest were examined to give direction to future research. Finally, for some analyses of subgroups, a small *n* may have masked significant findings that only appeared as trends in this data.

Any speculation about the cause(s) of the sex differences found in this study would be premature. It would be helpful in future research to explore the relationships among gender, anxiety, and group interaction. The gender composition of collaborative groups, a possible moderating factor, should also be explored by comparing the performance of all-male, all-female, and mixed-sex groups.

It is also possible that the women experienced so much anxiety related to this class because of stereotype threat. Their high level of involvement in collaborative discussions and their beliefs that their contributions had been successful may have helped alleviate this stress. Men's exam performance, on the other hand, may not have been affected by the amount of their involvement in collaborative discussions or their ratings of the success of their own collaborative contributions because they did not experience stereotype threat in this situation and felt more confident than women as a result. In order to examine the idea that collaborative discussions might be especially helpful to statistics students experiencing stereotype threat, a future study might purposefully induce stereotype threat in some participants to see whether they show more benefit from collaborative discussion than do those in whom no stereotype threat had been evoked. Researchers could also survey minority and female students as to whether or not they accept (and to what extent) the negative stereotypes asserting their groups' lack of ability in mathematics. How strongly they agree with the stereotype could then be related to how much minority and female students seem to benefit from collaborative testing compared to non-majority and male students.

As noted in the ratings of how having a partner helped, collaboration was most often described by students as a good experience. This, in itself, may be a worthwhile outcome. Regardless of exam performance, because their first encounter with statistics was pleasant, students might have more positive attitudes towards additional class work in statistics and related classes

5.3. RECOMMENDATIONS FOR THE CLASSROOM

Group learning works best when the members of the group have had practice interacting. The training session which occurred at least one week before testing began in the current study may not have been extensive enough to adequately judge the potential of collaborative testing. It may be particularly important, as Wilson (1999) notes, that students know each other, have worked together, and respect each other. In this study, the emphasis was on examining only the facilitative effects of collaborative testing. It is possible that collaborative testing would have proven to be of greater benefit to students if they had worked collaboratively in study groups or in classwork regularly throughout the course. It is therefore recommended that collaborative testing be imbedded in a more general collaborative learning environment. It is also recommended that instructors attempt to prevent "social loafing" during testing by having ratings of partner contributions to discussions be part of the exam grade.

In order to maximize the benefits of collaborative testing, the groups themselves should be assigned so that weaker students would be paired with stronger ones. Greater gains would probably result than if two well-prepared or two poorly-prepared students were teamed as sometimes happened in this study due to the fact that pairs were randomly selected.

REFERENCES

Baloglu, M. (2004). Statistics anxiety and mathematics anxiety: Some interesting differences. *Educational Research Quarterly, 27*(3), 38–48.

Berry, J., & Nyman, M. A. (2002). Small-group assessment methods in mathematics. *International Journal of Mathematics Education in Science and Technology, 33*(5), 641–649.

Betz, N. E. (1978). Prevalence, distribution, and correlates of mathematics anxiety in college students. *Journal of Counseling Psychology, 25*(5), 441–448.

Björnsdóttir, A., Garfield, J., & Everson, M. (2015). Evaluating two models of collaborative tests in an online introductory statistics course. *Statistics Education Research Journal, 14*(1), 36–59. [Online: https://iase-web.org/documents/SERJ/SERJ14(1)_Bjornsdottir.pdf]

Breedlove, W., Burkett, T., & Winfield, I. (2004). Collaborative testing and test anxiety. *Journal of Scholarship of Teaching and Learning, 4*(2), 33–42.

Campbell, J. D. (2004). *Handbook of mathematics cognition*. Philadelphia, PA: Psychology Press.

Castor, T. (2004). Making student thinking visible by examining discussion during group testing. *New Directions for Teaching and Learning, 2004*(100), 95–99.

Chiesi, F., Primi, C., Bilgin, A. A., Lopez, M. W., del Carmen Fabrizio, M., Gozlu, S., & Tuan, N. M. (2016). Measuring university students' approaches to learning statistics: An invariance study. *Journal of Psychoeducational Assessment, 34*(3), 256–268.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cortright, R. N., Collins, H. L., & DiCarlo, S. E. (2005). Peer instruction enhanced meaningful learning: Ability to solve novel problems. *Advances in Physiology Education, 29*(2), 107–111.

Cruise, R. J., Cash, R. W., & Bolton, D. L. (1985, August). *Development and validation of an instrument to measure statistical anxiety.* Paper presented at the annual meeting of the Statistical Education Section, American Statistical Association, Washington, DC.

Dahlström, Ö. (2012). Learning during a collaborative final exam. *Educational Research and Evaluation, 18*(4), 321–332.

Dallmer, D. (2004). Collaborative test taking with adult learners. *Adult Learning, 15*(3–4), 4–7.

Ewald, J. D. (2005). Language-related episodes in an assessment context: A "small-group quiz." *Canadian Modern Language Review, 61*(4), 565–586.

Fitzgerald, S. M. (1997). The relationship between anxiety and statistic achievement: A meta-analysis. *Dissertation Abstracts, International Section A: Humanities and Social Sciences, 58,* (2–A), 0383.

Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C., Katzaroff, M., & Dutka, S. (1998). Comparisons among individual and cooperative performance assessments and other measures of mathematics competence. *The Elementary School Journal, 99*(1), 23–51.

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

Giuliodori, M. J., Lujan, H. L., & DiCarlo, S. E. (2008). Collaborative group testing benefits high- and low-performing students. *Advances in Physiology Education, 32*(4), 274–278.

Giraud, G., & Enders, C. (1998, April). *The effects of repeated cooperative testing in an introductory statistics course.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA. [On-line]. ERIC Document Reproduction Service, ED445103.

Hanshaw, L. G. (1982). Test anxiety, self-concept, and the test performance of students paired for testing and the same students working alone. *Science Education, 66*(1), 15–24.

Helmericks, S. G. (1993). Collaborative testing in social statistics: Toward gemeinstat. *Teaching Sociology, 21*(3), 287–297.

Hendrickson, J. M., Brady, M. P., & Algozzine, B. (1987). Peer-mediated testing: The effects of an alternative testing procedure in higher education**.** *Educational and Psychological Research, 7*(2), 91–101.

Hite, P. A. (1996). An experimental study of the effectiveness of group exams in an individual income tax class. *Issues in Accounting Education, 11*(1), 61–76.

Hopko, D.R., Mahadevan, R., Bare, R. L., & Hunt, M. K. (2003). The Abbreviated Mathematics Anxiety Scale (AMAS): Construction, validity, and reliability. *Assessment, 10*(2), 178–182.

Ives, J. (2014). Measuring the learning from two-stage collaborative group exams. In P. Engelhardt, A. Churukian, & D. Jones (Eds.), *2014 Physics Education Research Conference Proceedings* (pp. 12–126). Minneapolis, MN: American Association of Physics Teachers.

Jensen, M., Moore, R., & Hatch, J. (2002). Cooperative learning – Part 3: Electronic cooperative quizzes. *American Biology Teacher, 63*(3), 29–34.

Johnson, D.W., Johnson, R. T., & Smith, K. A. (1991). *Active learning: Cooperation in the college classroom.* Edina, MN: Interaction.

Kapitanoff, S. (2009). Collaborative testing: Cognitive and interpersonal processes related to enhanced test performance. *Active Learning in Higher Education, 10*(1), 56–70.

Klugman, S. F. (1944). Comparative versus individual efficiency in problem solving. *Educational Psychology, 35,* 91–100.

Ley, K., Hodges, R., & Young, D. (1995). Partner testing. *Research and Teaching in Developmental Education, 12*(1), 23–30.

LoGiudice, A. B., Pachai, A. A., & Kim, J. A. (2015). Testing together: When do students learn more through collaborative tests? *Scholarship of Teaching and Learning in Psychology, 1*(4), 377–389.

Marx, D. M., & Roman, J. S. (2002). Female role models: Protecting women's math test performance. *Journal Personality and Social Psychology Bulletin, 25*(9), 1183–1193.

Morris, L. W., Kellaway, D. –S., & Smith, D. H. (1978). Mathematics Anxiety Rating Scale: Predicting anxiety experiences and academic performance in two groups of students. *Journal of Educational Psychology, 70*(4), 589–594.

National Science Foundation, National Center for Science and Engineering Statistics. (2017). *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2017(* Special Report NSF 17-310). Arlington, VA: NSF.
[Online: http://www.nsf.gov/statistics/wmpd/]

Nolan, M. M., Beran, T., & Hecker, K. G. (2012). Surveys assessing students' attitudes towards statistics: A systematic review of validity and reliability. *Statistics Education Research Journal, 11*(2), 103–123.
[Online: https://iase-web.org/documents/SERJ/SERJ11(2)_Nolan.pdf]

Nowak, L. I., Miller, S. W., & Washburn, J. (1996). Team testing increases performance. *Journal of Education for Business, 71*(5), 253–256.

Onwuegbuzie, A. J., & Wilson, V. A. (2003). Statistics anxiety: Nature, etiology, antecedents, effects and treatment – A comprehensive review of the literature. *Teaching in Higher Education, 8*(2), 195–209.

Pandey, C., & Kapitanoff, S. (2011). The influence of anxiety and quality of interaction on collaborative test performance. *Active Learning in Higher Education, 12*(3), 163–174.

Phillips, A. P. (1988). Reducing nursing students' anxiety level and increasing retention of materials**.** *Journal of Nursing Education, 27*(1), 35–41.

Pretorius, T. B., & Norman, A. M. (1992). Psychometric data on the Statistics Anxiety Scale from a sample of South African Students. *Educational and Psychological Measurement, 52*, 933–937.

Singer, R. S. (1990). *The effects of collaborative testing on the test scores and the classroom attitudes of junior high school pre-algebra students* (ERIC Document Reproduction Services No. TM016072).

Spielberger, C. D., Gonzalez, H. P., Taylor, C.J., Anton, E. D., Algaze, B., Ross, G. R., & Westberry, L. G. (1980). *Test Anxiety Inventory: Sample set, manual, test booklet, and scoring key.* Redwood City, CA: Mind Garden.

Spielberger, C. D., Gorsuch, R.L., Lushene, R., Vagg, P. R., & Jacobs, G.A. (1983). S*tate-Trait Anxiety Inventory for Adults: Sample set, manual, test, scoring key.* Redwood City, CA: Mind Garden.

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

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance and African Americans, *Journal of Personality and Social Psychology, 69*(5), 797–811.

Stout, J. A., Dasgupta, N., Hunsinger, M., & McManus, M. A. (2011). STEMing the tide: Using ingroup experts to inoculate women's self-concept in science, technology, engineering, and mathematics (STEM). *Journal of Personality and Social Psychology, 100*(2), 255–270.

Tait, H., Entwistle, N., & McCune, V. (1998). ASSIST: A re-conceptualization of the approaches to studying inventory. In C. Rust (Ed.), *Improving student learning: Improving students as learners.* (pp. 262–271). Oxford, UK: The Oxford Centre for Staff and Learning Development.

Townsend, M. A. R., Moore, D. W., Tuck, B. F., & Wilson, K. M. (1998). Self-concept and anxiety in university students studying social science statistics within a co- operative learning structure. *Educational Psychology, 18*(1), 41–54.

Watson, J. (1988). Student characteristics and prediction of success in a conventional university mathematics course. *Journal of Experimental Education, 56*, 203212.

Webb, N. M. (1993). Collaborative group versus individual assessment in mathematics: Processes and outcomes. *Educational Assessment, 1*(2), 131–152.

Webb, N. M., Nemer, K. M., Chizhik, A. W., & Sugrue, B. (1998). Equity issues in collaborative group assessment: Group composition and performance. *American Educational Research Journal, 35*(4), 607–651.

Wilson, V. A. (1999, November). *Reducing statistics anxiety: A ranking of sixteen specific strategies*. Paper presented at the annual meeting of the Mid-south Educational Research Association. Point Clear, AL.

Yokomoto, C. F., & Ware, R. (1997). Variations of the group quiz that promote collaborative learning. *Proceedings of the 1997 Frontiers in Education Conference* (pp. 552–557). Pittsburgh, PA: Institute of Electrical and Electronic Engineers.

Zeidner, M. (1991). Statistics and mathematics anxiety in social science students: Some interesting parallels. *British Journal of Educational Psychology, 61*(Pt 3), 319–328.

Zimbardo, P. G., Butler, L. D., & Wolfe, V. A. (2003). Cooperative college examinations: More gain, less pain when students share information and grades**.** *Journal of Experimental Education, 71*(2), 101–125**.**

SUSAN KAPITANOFF
American Jewish University
15600 Mulholland Drive
Los Angeles, CA 90077
USA