

EVALUATING THE USE OF TWO DIFFERENT MODELS OF COLLABORATIVE
TESTS IN AN ONLINE INTRODUCTORY STATISTICS COURSE

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Audbjorg Bjornsdottir

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Joan Garfield, Adviser
Michelle Everson Co-adviser

April 2012

Acknowledgements

This dissertation would not have been completed without the support of numerous individuals. Foremost, I want to thank my adviser Joan Garfield for her endless support, encouragement and guidance in both writing this paper and through my graduate studies for the past 5 years. I cannot image completing this work without her. To my other committee members; my co-adviser Michelle Everson for providing me with continues support in my teaching and research, Robert delMas for guidance and support, and Aaron Doering for his help. Thanks to the two wonderful teaching assistants Ting and Cengiz that helped me with the course and collect data in fall 2011.

Thanks to all my friends that believed in me and supported me through my graduate work. I would further like to thank my family, especially my parents for supporting and encouraging me to pursue my studies. Lastly and most importantly to my husband and kids, thanks Gísli for revising this paper numerous times, for supporting and enduring me throughout this paper and my graduate studies. And to my three kids Björn, Ágústa and Helgi: thanks for being exactly the way you are: I could not have done this without you all. “Núna er mamma búin í skólanum.”

Dedication

This dissertation is dedicated to the power that propels the Universe and to all present and future students and instructors of introductory statistics. Do not despair; there is a light at the end of the tunnel.

Abstract

The purpose of this study was to explore how collaborative tests could be implemented successfully in online introductory statistics courses. The research questions set forth were (1) What is the impact of using collaborative tests in an online statistics course on students' learning? (2) What is the effect of using collaborative tests on students' attitudes towards statistics? and (3) How does using a required consensus on collaborative tests vs. a nonconsensus approach affect group discussions?

Three collaborative tests were implemented in two online sections of the EPSY-3264 Basic and Applied Statistics course offered at the University of Minnesota. The two sections were identical in terms of the instructor, assignments, assessments, and lecture notes used. The only difference between the two sections was in terms of the format of the collaborative tests that were used. In the consensus section, students worked together in groups and submitted one answer per group. In the nonconsensus section, students worked on the test together in groups but submitted tests individually. Students were randomly assigned to a consensus ($n=32$) or a nonconsensus ($n=27$) section of the course.

The Comprehensive Assessment of Important Outcomes in Statistics (CAOS) test was used to measure students' learning, both at the beginning and at the end of the course. The Survey Of Attitudes Toward Statistics (SATS-36) instrument was used to measure students' change in attitudes towards statistics. Another instrument designed by the instructor to measure students' perspective towards collaborative testing was also used. Students' discussions during the three collaborative tests were reviewed using the Pozzi, Manca, Persico, & Sarti, (2007) framework to evaluate and monitor computer-supported collaborative learning. Discussions were coded using three dimensions,

(Social, Teaching and Cognitive) and their indicators from the framework and then converted to quantitative variables that were used in the data analysis.

No significant relationship was found between different sections and students' scores on the CAOS. There was no significant difference in students' attitudes towards statistics between the two sections. However, for both sections, students' attitudes increased in terms of their intellectual knowledge, skills, and interest towards statistics after taking the three collaborative tests. The effects of using a required consensus on collaborative tests vs. a nonconsensus approach on group discussions did not seem to be significantly different. The two formats of the collaborative tests that were used seemed to support students' discussion more in terms of the Cognitive dimension compared to the Social and Teaching dimensions.

Overall, the results suggest that the difference between using two different formats of collaborative tests is not significant. However, the results support what research on collaborative tests in face-to-face courses have demonstrated before such as an increase in students' attitudes towards learning (e.g., Giraud & Enders, 2000; Ioannou & Artion, 2010). Instructors and researchers should continue to use and experiment with collaborative tests in online introductory statistics courses. The study here is just the beginning in terms of conducting empirical research into what teaching methods and assessments should be used in an effort to create quality and effective online statistics courses.

Table of Contents

Acknowledgements	i
Dedication	ii
Abstract	iii
List of Tables	x
List of Figures	xiii
Chapter 1 Introduction	1
The Value of Using Collaborative Tests	2
Description of the Study	3
Structure of the Dissertation	4
Chapter 2 Review of the Literature	5
Collaborative Learning	6
Cooperative Learning Theory	7
Research on the Effectiveness of Cooperative Learning	9
Collaborative Testing	10
Summary and Discussion	32
Online Instruction of College Courses	34
Collaborative Learning Online	35
Collaborative Testing Online	42
Summary and Discussion	43
Collaborative Learning in Introductory College Statistics Courses	44
Implementing Collaborative Learning in Statistics Courses	47

Studies on Using Collaborative Testing in Introductory Statistics Courses	48
Summary and Discussion	49
Online Instruction in Introductory Statistics Courses	51
Research on Online/Hybrid Statistics Course vs. Face-to-Face Statistics Courses	51
Suggestions on Teaching Statistics Courses Online	53
Summary and Discussion	56
Summary and Implications of the Literature Review	58
Implications for Research	63
Chapter 3 Methods	65
Overview of the Study	65
Subjects/Settings	66
The Course	68
Pedagogical Model	69
Learning Environment	70
The Experimental Conditions	71
Instruments and Measurements	72
Required Assessments	73
Instruments Offered as Extra Credit	76
Reliability Analysis of the Research Instruments	80
Timeline for Instruments and Test Administration	81
Analysis of Data	82

Qualitative Analysis	82
Qualitative Data Analysis Framework	83
Quantitative Analyses	90
Chapter 4 Results	96
Examining the Two Online Sections	96
Reliability and Scale Scores	99
Examining the First Research Question: What is the impact of using collaborative tests in an online statistics course on students learning?	101
Model 1: Midterm Exam as the Dependent Variable	101
Model 2: Final Exam as the Dependent Variable	112
Examining the Second Research Question: What is the effect of using collaborative tests on students' attitudes towards statistics?	121
Examining the Third Research Question: How does using a required consensus on collaborative tests vs. a nonconsensus approach effect group discussions?	124
Additional Analysis	135
Summary	138
Chapter 5 Discussion	139
Research Question 1. What is the impact of using collaborative tests in an online statistics course on students' learning?	142
Research Question 2. What is the effect of using collaborative tests on students' attitudes towards statistics?	144

Research Question 3. How does using a required consensus on collaborative tests vs. a nonconsensus approach affect group discussions?	145
Limitations of the Study	147
Implications for Teaching Online Statistics	148
Implications for Future Research	149
References	151
Appendix A Instruments	162
A-1 Comprehensive Assessment of Important Outcomes in Statistics (CAOS)	162
A-2 Midterm	177
A-3 Directions on Collaborative Test	189
Consensus Section	189
Nonconsensus	190
A-4 Collaborative Tests	191
A-5 Students Perception on Collaborative Tests (SPCT)	207
A-6 The Survey Of Attitudes Toward Statistics (SATS-36)	210
Pre survey	210
Post Survey	213
Appendix B Syllabus	217
Appendix C Correspondence to Students	231
C-1 Initial Email to Students	231
C-2 Consent Form	232

C-3 Invite Email for the The Survey Of Attitudes Toward Statistics (Pre-SATS-36)	233
C-4 Invite Email for the The Survey Of Attitudes Toward Statistics (Post-SATS-36)	234
C-5 Invite Email for the Students Perception on Collaborative Tests (SPCT)	235
C-6 Thank you Email for The Survey Of Attitudes Toward Statistics (SATS-36).	236
C-7 Thank you Email for The Students Perception on Collaborative Tests (SPCT)	237

List of Tables

Table 1 <i>Summary of Formats Used in Studies Regarding Collaborative Exams/tests</i>	16
Table 2 <i>Midterm Exam Item Numbers Based on Format and Source</i>	74
Table 3 <i>Frequency of Item Types and Level of Difficulty on the Collaborative Tests</i>	75
Table 4 <i>Original Statements and Statements Used on Pre- and Post-SATS-36 Instruments</i>	77
Table 5 <i>Questions Added to the Pre-SATS-36</i>	78
Table 6 <i>Different Statements on the SPCT between the Two Course Sections</i>	79
Table 7 <i>Coefficient Alpha for Sample Scores and Responses</i>	80
Table 8 <i>Weeks & Instruments Administered</i>	82
Table 9 <i>Indicators for All Dimensions and Examples from the Collaborative Tests (Inspired by Persico et al., 2010)</i>	87
Table 10 <i>Proportions of Students' Academic Levels between Sections</i>	98
Table 11 <i>Proportion and Frequency of Students that Have Been Enrolled in Online Courses Before</i>	99
Table 12. <i>Reliability and Descriptive Summary for the Average Scale Scores</i>	100
Table 13 <i>Correlation for CLT1 & 2 Discussion (Cognitive, Teaching and Social) and the Midterm Exam Variables</i>	102
Table 14 <i>Hierarchical Multiple Regression Reports for Model 1, Midterm Exam as Dependent Variable</i>	104

Table 15 <i>Multiple Regression Reports for the Final Model with Midterm Exam as a Dependent Variable</i>	106
Table 16 <i>Model 1, the Final Model with the Interaction Terms</i>	110
Table 17 <i>Multiple Regression Results for the Interaction Model, Midterm Exam as Dependent Variable</i>	111
Table 18 <i>Correlation for CLTDiscussion (Teaching, Cognitive and Social) and the Final Exam Variable</i>	114
Table 19 <i>Hierarchical Multiple Regression Reports for Model 2, Final Exam as the Dependent Variable</i>	115
Table 20 <i>Multiple Regression Reports for the Final Model with Final Exam as a Dependent Variable</i>	117
Table 21 <i>Model 2, the Model with the Interaction Terms, Final Exam as a Dependent Variable</i>	120
Table 22 <i>Mean Preresponses on the SATS-36 Subscales by Section</i>	121
Table 23 <i>Mean Postresponses on the SATS-36 Subscales by Section</i>	122
Table 24 <i>Mean Difference Scores on the SATS-36 Subscales by Section</i>	123
Table 25 <i>Test of Mean Difference Scores on SATS-36 Subscales within Section</i>	124
Table 26 <i>Tests of Mean Proportions of Different Dimensions between the Two Sections</i>	126
Table 27 <i>Tests of Mean Proportions for the 10 Indicators between Sections</i>	128
Table 28 <i>Tests of Mean Proportions for the Participative Dimension between Sections</i>	129

Table 29 <i>Test of Mean Proportions for the Cognitive Dimension Indicators between Sections</i>	130
Table 30 <i>Test of Mean Proportions for the Social Dimension Indicators between Sections</i>	131
Table 31 <i>Test of Mean Proportions for the Teaching Dimension Indicators between Sections</i>	133
Table 32 <i>Correlations between the 10 Indicators</i>	134

List of Figures

<i>Figure 1.</i> Scatterplot of the standardized residuals against the standardized predicted values for Model 1.	107
<i>Figure 2.</i> Normal P-P Plot of regression standardized residuals.	108
<i>Figure 3.</i> Interaction relationship between Midterm exam score and sections (Nonconsensus=0, Consensus=1), controlling for CLT1&2Discussion-Social.	112
<i>Figure 4.</i> Scatterplot of the standardized residuals against the standardized predicted values for Model 2.	118
<i>Figure 5.</i> Normal P-P Plot of the regression standardized residuals.	119
<i>Figure 6.</i> Mean proportions of three dimensions on the three group tests.	125
<i>Figure 7.</i> Mean proportion of the three dimensions between sections.	126
<i>Figure 8.</i> Mean proportions of indicators on the cognitive dimension between the two sections.	129
<i>Figure 9.</i> Mean proportions of indicators on the social dimension between sections.	131
<i>Figure 10.</i> Mean proportions of indicators on the teaching dimension between sections.	132

Chapter 1

Introduction

In the last two decades, there have been calls for change in statistics education that have been influenced by a movement to reform teaching of the mathematical sciences in general (Moore, 1997). When it comes to statistics education, this reform movement has focused heavily on the teaching of introductory statistics at the college-level, with emphasis on content (e.g., more data analysis and less probability), pedagogy (e.g., altering teaching styles), and the use of technology for data analysis and simulation. In regards to pedagogy, the focus has been on altering teaching styles by having instructors move away from lectures to more active learning approaches (e.g., Moore, 1997; Garfield, Hogg, Schau, & Whittinghill, 2002). More recently, the Guidelines for Assessment and Instruction in Statistics Education (GAISE, 2005) that were endorsed by the American Statistical Association (ASA) outline areas for change in the introductory statistics class:

1. Emphasize statistical literacy and develop statistical thinking.
2. Use real data.
3. Stress conceptual understanding rather than mere knowledge of procedures.
4. Foster active learning in the classroom.
5. Use technology for developing conceptual understanding and analyzing data.
6. Integrate assessments that are aligned with course goals to improve as well as evaluate student learning.

Some ways to emphasize active learning in the classroom include using group problem solving activities and group discussions (GAISE, 2005). Well-designed and effective active-learning activities allow students to construct their own knowledge. This approach is quite different from long held traditional teaching styles that are mostly made up of lectures where teachers tell students information they are to remember (Garfield, 1993).

A way for statistics teachers to incorporate active learning in their classes is to structure opportunities for students to learn together in small groups (GAISE, 2005; Garfield, 1993). These suggestions have been supported by research on cooperative learning that has documented the effectiveness of cooperative learning activities in classrooms (see Johnson, Johnson, & Smith, 1991).

The Value of Using Collaborative Tests

By implementing well-structured cooperative learning activities, students can be actively involved cognitively, physically, emotionally and psychologically in constructing their own knowledge (Johnson et al., 1991). In particular, exams, administered collaboratively, can be used as teaching and learning tools that enhance the construction of knowledge (Giuliodori, Ljuan, & DiCarlo, 2008). However, no research currently exists about using collaborative tests in statistics courses, online or face-to-face. The main purpose of this paper is to explore how collaborative tests could be implemented in an online introductory statistics course by reviewing different areas of literature related to this topic.

Collaborative tests have been used successfully in the classroom setting (e.g., Ioannou & Artino, 2010, Zimbardo, Butler, & Wolfe, 2003) and the effective use of

collaborative learning methods in online courses has been reported in the literature on online education (Roberts, 2004). Research suggests that using collaborative learning in statistics courses has been successful in terms of increasing students' test scores (e.g., Giraud, 1997; Magel, 1998; Keller & Steinhorst, 1995; Potthast, 1999), improving students' attitude toward learning statistics (Potthast, 1999), and increasing student engagement in class (e.g., Giraud, 1997; Magel, 1998; Keller & Steinhorst, 1995). Statistics courses offered online have been shown to be as effective as traditional statistics courses offered in classroom settings (e.g., Gunnarsson, 2001; Hong, Lai, & Holton, 2003; Ward, 2004), and the use of collaborative learning in online statistics courses has been described (Everson, 2006; Everson, & Garfield, 2008). However, the effects of using collaborative learning in online statistics courses have not been established through research.

Description of the Study

The main goal of this study is to explore the impact of using collaborative tests on students' learning in online introductory statistics courses. The literature review revealed, among other things, that by using collaborative tests there was an increase in positive student attitudes (e.g., Giraud & Enders, 2000; Ioannou & Artino, 2010) and an increase in student interaction (e.g., Ioannou & Artino, 2010). However, the effect of using different formats (e.g. consensus, nonconsensus, multiple choice, essays) of collaborative tests remains unknown. Because of this, one of the objectives of this study will be to investigate the effects of using different formats of collaborative tests on students' attitudes towards statistics and group discussion in online courses. This will be done in

order to explore whether the positive effects of using collaborative tests reported in the literature apply to using different formats of collaborative tests in online courses.

More specifically, the goal of this study is to answer the following three research questions:

1. What is the impact of using collaborative tests in an online statistics course on students' learning?
2. What is the effect of using collaborative tests on students' attitudes towards statistics?
3. How does using a required consensus on collaborative tests vs. a nonconsensus approach affect group discussions?

Structure of the Dissertation

The second chapter includes the literature review for the study in this dissertation. Previous research on collaborative learning, collaborative tests, online learning and introductory statistics courses is reviewed, specifically as it relates to using collaborative tests in online introductory statistics courses. The three research questions this study aims to answer are a direct result of the review of the literature in Chapter 2.

Chapter 3 covers the methodology used in the study and provides details on how data was gathered, the subjects involved in the study, and the setting in which the study took place. It describes the instruments and the data analysis that were used. Chapter 4 provides the results of the data analysis used to answer the research questions. The last and final chapter includes a discussion and summary of the results along with implications for future research and teaching.

Chapter 2

Review of the Literature

This literature review was conducted to explore the use of collaborative learning in online introductory statistics courses, and, in particular, the use of collaborative tests in those courses.

The review begins with a broad definition of collaborative learning and the difference between the terms collaborative and cooperative. A short overview of cooperative learning theory is provided, followed by discussion of the effectiveness and implementation of cooperative learning in classroom teaching. A definition of collaborative testing is provided along with a substantial review of studies that have investigated the benefits and disadvantages of using collaborative testing in face-to-face classroom settings. This is followed by a critique of the methods used and the inferences made from the results in these studies.

The next section focuses on the teaching of online college courses and the ways in which collaborative learning have been used in the online environment. The implementation and evaluation of collaborative learning in online courses will be examined, followed by a short review of studies that have attempted to implement collaborative testing in the online environment.

The third section focuses on ways in which collaborative learning has been used in introductory college statistics courses. Research studies that have employed collaborative learning and collaborative testing in statistics courses were explored, and major findings and implications are reviewed.

In the fourth section, studies pertaining to the teaching of online introductory statistics courses are presented. Here, the focus is on studies that have (a) compared online or hybrid statistics courses to face-to-face statistics courses, and (b) presented teachers' experiences with the teaching of online statistics courses and suggestions for ways to teach such courses. Included in this section is a review of ways in which collaborative learning has been used in online statistics courses. Finally, a summary is given of what the literature suggests regarding what the use of collaborative tests in online statistics courses might look like.

Collaborative Learning

The term collaborative learning has been used in such a variety of forms that a single definition is not available. The broadest definition of collaborative learning would be when two or more people learn or attempt to learn something together (Dillenbourg, 1999).

Resta and Laferrière (2007) point out that there is a lack of universal meaning for the terms collaborative and cooperative learning and it is unclear what the exact differences and commonalities between the two terms are. According to Resta and Laferrière, a clear distinction of the two terms might be hindered because of researchers' different purposes, goals and perspectives. Johnson and Johnson (1996) stated that a clear definition does exist for cooperative learning, but there is ambiguity regarding the meaning for collaborative learning. According to Johnson and Johnson, historically, when compared to cooperative learning, collaborative learning has been seen as less structured and more student-directed, with limited teacher direction. The ambiguity in the definition of collaborative learning is the result of the vagueness of the teachers' and

students' roles. This results in the terms cooperative and collaborative learning often being used interchangeably and synonymously.

In this paper, the term *collaborative* will be used as an umbrella term to encompass learning that involves peer and group learning where students work together to maximize learning. The term cooperative learning will be used when there is a reference to the *cooperative learning theory* defined by Johnson, Johnson, and Holubec (2008).

Cooperative Learning Theory

Cooperative learning takes place in groups, and it is where individuals work together to accomplish their shared learning goals (Johnson, Johnson, & Holubec, 2008). The desired outcomes are seen as beneficial to the individuals themselves and also to their group members (Johnson et al., 2008). Cooperative learning derives from three general theoretical perspectives: social interdependence, cognitive-development and behavioral learning theories.

Cooperative learning is made up of five basic elements: positive interdependence, individual accountability, promotive interaction, interpersonal and small group skills, and group processing (Johnson et al., 2008). Positive interdependence is the most important element; it “exists when group members perceive that they are linked with each other in a way that one cannot succeed unless everyone succeeds. If one fails everyone fails ...” (Johnson et al., 2008, p. 14). Individual accountability refers to the idea that every member in the group is accountable to the group goals and the work each individual needs to contribute in order to achieve them. Every member is accountable in contributing to the work in order for the group to achieve its goals. Individual

accountability eliminates free riders in-group work. Promotive interaction is where group members help each other, by sharing resources, providing support, encouragement and praise to each other in their effort to learn. Interpersonal and small group skills are about students learning the skills needed to function as a part of a group and how to achieve their goals as a group. It involves, among other things, each group member knowing and learning effective leadership, decision-making, trust-building, communication and conflict-management. The fifth element is group processing, and this is where group members review their work as it relates to them achieving their goals and maintaining effective working relationships (Johnson et al., 2008). Cooperative learning requires considerable planning from the teachers in order to make sure all of the five basic elements are in place (Johnson et al., 2008).

Students in cooperative learning groups differ from students in other learning groups because they perceive that they cannot achieve their learning goals if other group members cannot also reach their goals; thus, the students are linked together (Johnson & Johnson, 1996). According to Slavin (1991), if cooperative learning is to be successful, group goals and individual accountability must be present. Group members must work and share the same goals and the group success is dependent on the individual learning of each group member. While there are different cooperative learning methods available they all “share the idea that students work together to learn and are responsible for one another’s learning as well as their own” (Slavin, 1991, p. 73). According to Johnson et al. (2008), the size of cooperative learning groups normally ranges from 2 to 4 students; smaller groups are considered better. No ideal size for a cooperative learning group exists, however. Group size is dependent on several factors: how long the group will be

working together, students' ages and experiences with group work, and the material and equipment available (Johnson et al., 2008).

Research on the Effectiveness of Cooperative Learning

Numerous studies have been conducted to measure the effects of using cooperative learning compared to competitive and individualistic learning (see more in Johnson et al., 2008). Results from these studies have been categorized into three major components: efforts to achieve, positive relationships and psychological health.

Cooperative learning has been shown to improve factors that count toward all of these three major categories. Examples of these include higher scholastic achievement, more caring and committed relationships, and increased self-esteem for those students who engage in cooperative learning (Johnson et al., 2008).

In a meta-analysis of 164 studies related to cooperative learning methods, Johnson, Johnson and Stanne (2000) found that there was a significant positive impact on student achievement when eight different cooperative learning methods (Learning together (LT), Academic Controversy (AC), Student-Team-Achievement-Division (STAD), Teams-Games-Tournaments (TGT), Group Investigation (GI), Jigsaw, Teams-Assisted-Individualization (TAI) and Cooperative Integrated Reading and Composition (CIRC) were compared to competitive and individualistic learning methods. When compared to competitive learning methods the largest effect size (0.82) was for the LT method and the smallest effect size (0.18) was for the CIRC method. When compared to the individual learning method the largest effect size (1.03) was again for the LT method and the smallest effect size (0.13) was for the Jigsaw method.

Roseth, Johnson, and Johnson (2008) conducted a meta-analysis of 148 studies comparing the relative effectiveness of cooperative, competitive and individualistic goal structures on promoting early adolescents achievement and positive peer relationship. Results from the meta-analysis showed that higher achievement was associated with cooperative goal structures compared to competitive (effect size .46) and individualistic (effect size .55) goal structures. Positive peer relationships were also associated with cooperative goal structure compare to competitive (effect size .48) and individualist (effect size .42) goal structure.

Collaborative Testing

Testing is a type of formal assessment where data is gathered systematically to give both information and guide decisions making regarding the learning progress (Eggen & Kauchak, 2006). Collaborative testing occurs when students work together on an exam or assessment (Lusk & Conklin, 2003). Different formats of collaborative testing have been used and reported in the literature, with the most common formats being to have students work together in pairs or groups on a test and turn in either individual nonconsensus answers (e.g., Lusk & Conklin, 2003; Breedlove, Burkett & Winfield, 2004; Kapitanoff, 2009) or group consensus answer sheets (e.g., Haberyan & Barnett, 2010; Helmericks, 1993; Hick, 2007), or to have students take the same test twice--first individually and then as a group—and turn in answers for both sections (e.g., Giuliadori et al., 2008; Ioannou & Artino, 2010; Rao, Collins & DiCarlo, 2002). The difference between the nonconsensus versus the consensus format is that in the former, students need not agree on their answers. Although much research has been conducted on the

effectiveness of cooperative learning (as the aforementioned meta-analyses make clear), evidence of the effectiveness of collaborative testing is sparse.

Use of collaborative testing. In a literature review on collaborative testing, Sandahl (2009) located nine studies from the field of nursing that involved the use of collaborative tests. Five of the studies reviewed by Sandahl did not require a consensus among students regarding answers on the collaborative tests. One of the nonconsensus studies focused on the effectiveness of group testing and individual testing for short-and long-term retention by comparing scores from a pretest to the midterm (short-term) and to the final (long-term). By using scores on a pretest and midterms and final exams as outcome measures, the results revealed significantly better retention of the material for the collaborative testing method. Further, the students favored the group testing strategy. The second nonconsensus study examined the effects of collaborative testing on student anxiety and learning using scores from group and individual quizzes, final exam scores, and students' responses to a questionnaire designed to rate anxiety. Data came from three classes from a course that was taught by the same instructor in fall, winter and spring. The results suggested that there was an increase in learning and retention of material for students in the winter and spring classes when comparing quiz averages and final exam scores. From all classes combined 97% of the students reported that their anxiety levels decreased during the course. The third nonconsensus study (Lusk & Conklin, 2003) explored the effects of collaborative testing on student learning and test-taking skills, where learning style and test taking skills were measured by a National nursing entrance test. Exam scores served as outcome measures in this study and scores were compared to exam scores from previous semester where students were tested individually. The results

showed a significant difference between unit exam scores with students testing better using collaborative testing when compare to unit exam scores from previous semester. However, there was no significant difference between scores on final exams, there were some indicators of improved test taking skills for students who took the collaborative tests, or as the authors stated, “Collaborative testing provided students with the opportunity to become more proficient with critical thinking and collaboration skills” (Lusk & Conklin, 2003, p. 124). In addition, it was observed that all students reported a decrease in test anxiety for students that took the collaborative tests. The fourth nonconsensus study observed the effect of a group test on nursing students’ learning of fluid and electrolyte content, again using exam scores as the outcome measure. The results showed that perceptions towards the group testing format were positive, but there was concern among students that unprepared students might be able to earn a higher exam score than they deserved. The fifth nonconsensus study incorporated group quizzes to explore the impact of active, student-centered learning strategies on nursing students using qualitative data from a course evaluation. The results showed that grades were higher for students in the student-centered approach compare to grades from previous semesters where lecture approach was used and that course evaluations were positive for the student-centered approach (which included small-groups, role-play, take-home quizzes and group quizzes) (Sandahl, 2009).

Of the four studies reviewed by Sandahl (2009) where consensus was required, three of the studies explored required consensus on collaborative testing using exam scores as outcome measures. Students completed the same exams two times in those studies: an individual exam and then a group exam where a consensus was required.

Individual and group exam scores were used as an outcome measure. The results showed higher mean scores for group exams and students' perception of the collaborative testing experience was overwhelmingly positive. In particular, students felt they experienced less anxiety, increased learning, and improved peer relationships and thinking skills. The final consensus study reviewed by Sandahl looked at nursing students' responses to group consensus testing using both exam scores and a questionnaire about students' experiences with collaborative testing as outcome measures. Students completed an individual exam and then they repeated the same exam in groups. No difference was found in student grades when compared to previous semesters where collaborative testing had not been used. Results from the questionnaire showed that students were positive in regards to the collaborative testing and it did encourage students to communicate and provide rationale for their answers to their group members. However majority of the students reported that their anxiety was not reduced because of the testing procedure.

All the studies in Sandahl's (2009) literature review revealed positive results in favor of collaborative learning, and increased exam scores. Further, according to Sandahl, future research on collaborative testing could be improved by randomly assigning students to groups and including a control group (since none of the nine studies discussed above did include a control group). The focus should also be on group size, group formation, and the stability of groups over time, in addition to the effect of these characteristics on student learning, critical thinking skills, and group processing skills.

Use of collaborative tests at the university level. In addition to Sandahl's (2009) literature review, 12 other studies were located that utilized collaborative exams at the university level (see Table 1). Different formats of collaborative testing become apparent

when looking at Table 1. The size of the student groups working on the collaborative exams ranged from 2 to 6 students per group; students working in pairs or groups of 3 were most common. Four studies (Haberyan & Barnett, 2010; Helmericks, 1993; Hick, 2007; Zimbardo et al., 2003) required students to reach a consensus regarding answers to questions on the exams or tests. In those studies, only one answer sheet was submitted and everybody in the same group received the same grade. Four studies (Giuliodori et al., 2008; Ioannou & Artino, 2010; Rao et al., 2002; Simkin, 2005) had students turn in both individual and group answers for the same test. Nine of the studies (Breedlove et al., 2004; Giraud & Enders, 2000; Giuliodori et al., 2008; Haberyan & Barnett, 2010; Ioannou & Artino, 2010; Kapitanoff, 2009; Rao et al., 2002; Simkin, 2005; Zimbardo et al., 2003) used multiple-choice exams, and of those, 4 (Haberyan & Barnett, 2010; Kapitanoff, 2009; Rao et al., 2002; Simkin, 2005) also used other types of questions such as short-answer, essay or constructed response.

Two studies (Helmericks, 1993; Hick, 2007) did not provide information about the exam type used, and one study (Shindler, 2004) used only essay exam questions. Five studies (Breedlove et al., 2004; Giraud & Enders, 2000; Haberyan & Barnett, 2010; Helmericks, 1993; Kapitanoff, 2009) involved students being randomly assigned to groups, and six studies (Giuliodori et al., 2008; Haberyan & Barnett, 2010; Ioannou & Artino, 2010; Shindler, 2004; Simkin, 2005; Zimbardo et al., 2003) had students self-select their groups. One study (Hick, 2007) assigned students to groups based on their grades on a content knowledge exam, and one study (Rao et al., 2002) did not clarify how students were assigned to groups. Ten of the studies (Breedlove et al., 2004; Giraud & Enders, 2000; Giuliodori et al., 2008; Haberyan & Barnett, 2010; Helmericks, 1993;

Ioannou & Artino, 2010; Kapitanoff, 2009; Simkin, 2005; Rao et al., 2002; Zimbardo et al., 2003) used test scores from either individual and group exams (or both types of exams) as outcome measures. Surveys measuring students' attitudes, personality factors, perceptions, willingness to participate in collaborative testing, and anxiety were used in 8 studies (Giraud & Enders, 2000; Haberyan & Barnett, 2010; Hick, 2007; Ioannou & Artino, 2010; Kapitanoff, 2009; Shindler, 2004; Simkin, 2005; Zimbardo et al., 2003) and students' evaluations were used in 3 studies (Helmericks, 1993; Giuliadori et al., 2008; Kapitanoff, 2009).

Table 1

Summary of Formats Used in Studies Regarding Collaborative Exams/tests

<i>Authors/ year</i>	<i>Research Design</i>	<i>Exam Format</i>	<i>Exam Type</i>	<i>Control Group</i>	<i>Selection into Groups</i>	<i>Group Size</i>	<i>n</i>	<i>Outcome Measures</i>
Helmericks, 1993	Quasi- experimental	One exam turned in	NA	Yes, prior class	Stratified sampling, to control for gender	3-4	58	Test scores, student evaluations
Giraud & Enders, 2000	Quasi- experimental	Individual exam	Multiple- choice, students received items stems	two consecutive classes compared	Randomly	3-4	53	Individual test score, Questionnaire measuring test anxiety, Questionnaire measuring student attitude towards collaborative testing experience
Rao, Collins & DiCarlo, 2002	Nonexperime ntal design	Individual and group answers	Fill in blank, multiple- choice, short essay & true and false statements	No	Students assigned to groups (random?)	2-3	16	Test scores
Zimbardo, Butler & Wolfe, 2003	Quasi- experiment	One test turned in	Multiple- choice,	Yes, solo testers. Same class	Students selected partners	2	300 & 276	Test scores, self-reports of student attitudes and perceptions towards collaborative exams
Breedlove, Burkett & Winfield, 2004	Quasi- experiment	Individual exams	Multiple- choice,	Another section of the same course	Randomly- assigned same sex partner	2	67	Test scores from both individual and collaborative exams

(cont.)

<i>Authors/ year</i>	<i>Research Design</i>	<i>Exam Format</i>	<i>Exam Type</i>	<i>Control Group</i>	<i>Selection into Groups</i>	<i>Group Size</i>	<i>n</i>	<i>Outcome Measures</i>
<i>(Table 1, cont.)</i>								
Shindler, 2004	Nonexperimental design	Two exam conditions individual & group	Essay exam, 3 items	No	Students selected groups	NA	46 & 382	Survey of student perception related to four dimensions of soundness (validity, reliability, efficiency and effect on the learner), focus group interviews
Simkin, 2005	Nonexperimental design	Individual and group quiz	Multiple-choice & constructed response	No	Students selected groups of max 3	2-3	25	Test scores, survey of students' responses about their attitudes regarding group exams.
Hick, 2007	Nonexperimental design	Only one quiz turned in	NA	No	Based on students' scores on a content knowledge exam	5-6	28	Pre- and post surveys to measure students' familiarity with the terms and willingness to participate in collaborative learning and group testing.
Giuliodori et al., 2008	Nonexperimental design	Individual and group tests	Multiple-choice,	No	Students selected a partner	2	65	Test scores, students evaluation on the collaborative testing
Kapitanoff, 2009	Nonexperimental design	Individual exams	Multiple-choice, short-answers	No	Randomly	2	33	Individual and group exam scores, a pretest questionnaire that asked about, text anxiety, study behavior, confident about doing well, how much they liked MC exams, attitude toward collaborative testing and an end of the semester evaluation. <i>(cont.)</i>

<i>Authors/ year</i>	<i>Research Design</i>	<i>Exam Format</i>	<i>Exam Type</i>	<i>Control Group</i>	<i>Selection into Groups</i>	<i>Group Size</i>	<i>n</i>	<i>Outcome Measures</i>
<i>(Table 1, cont.)</i>								
Haberyan & Barnett, 2010	Quasi-experiment	Consensus only one exam turned in	Multiple-choice & essay	Yes, those that tested individually in both studies	Study 1 students selected partners. Study 2. Students randomly assigned to two groups: test alone, test together	2	164 & 104	For both studies, test scores from individual and group test. For study 2 Scores from NEO-PIR (240 item personality inventory to measure the Big Five Personality Factors (Neuroticism, Extraversion, Openness To Experience, Agreeableness, and Conscientiousness), and the traits associated with each factor.
Ioannou & Artino, 2010	Nonexperimental design	Individual and group answers turned in	Multiple-choice,	No	Students selected groups	3-4	31	Individual and group exam scores and a collaborative assessment survey asking about administration, students perceive learning, enjoyment, anxiety, satisfaction and fairness in regards to the exam

Use of consensus in collaborative tests. In four studies (Haberyan & Barnett, 2010; Helmericks, 1993; Hick, 2007; Zimbardo et al., 2003) where consensus on collaborative tests was required, groups or pairs took one test together, discussed the answers to questions on the test, and then turned in just one test as a group (for which one grade was assigned).

Helmericks (1993) explored how a group exam could be implemented successfully in an undergraduate social statistics class. Students took three collaborative exams and scores from those were compared to individual exam scores from a previous class. Scores on each of the three collaborative exams were on average 13.46 % higher than the individual exams. However, the average score on the final exam for the collaborative class was 5.75 % lower compared to the individual class. According to Helmericks, one possible explanation for the difference in the final scores might be that students who completed individual exams were simply “better students.” Alternatively, taking collaborative exams might not have prepared students enough for the final exam, or might not have challenged them enough to apply the topic under study. Also, before taking the final exam, students were aware of their grade and what they needed to get on the final exam in order to get a good grade in the course. This might have led them to stop trying once they felt they had done enough to get the final course grade they wanted. Unfortunately, Helmericks’ study lacked detailed research questions, and as for the reasons why collaborative exams were used in the first place, Helmericks explained that the semester before, he had observed the use of collaborative groups to be successful and so wanted to see if collaborative exams might be successful as well. It should also be noted that although mean scores were computed and compared for the exams, no

significance tests were used, so it is unclear if the difference in grades between the two classes was statistically significant.

In a study of the effectiveness of collaborative team testing in an undergraduate introductory psychology course, students were given the option to take their midterm and/or final course examination with a partner in two consecutive comparable courses (Zimbardo, Butler, & Wolfe, 2003). In one course with 300 students, 62% chose to participate in a collaborative midterm and 48% chose to do a collaborative final exam. In the other course out of 276 students, 30% chose to take the collaborative midterm and 34% chose the collaborative final exam. Students were to select a classmate to take the test with and they had to do so 1 week before the exam. The exams consisted of multiple-choice items, with 90 items to be completed in 90 minutes. Grades from collaborative testing were significantly higher than for solo testing. The effect size was significant at 0.8. Students who participated in the collaborative testing reported decreased test anxiety during both studying and testing.

Zimbardo et al. (2003) also examined what expectations were at play to keep students from experimenting with collaborative testing by having a comparable group of students enrolled in another introductory psychology course evaluate the hypothetical option of selecting collaborative testing. These students did not participate in the testing method themselves. Students with higher grades reported favoring the solo approach. Students believed that a poor initial solo test performance to be the main reason for why a student would select collaborative testing. Data from the first class that participated in the collaborative testing did not support these two ideas. The authors arrived at the

conclusion that students hold the misconception that students choosing collaborative testing are less competitive and motivated.

Haberyan and Barnett (2010) tried to replicate the Zimbardo et al. (2003) study by using collaborative testing with a different sample and in different subject areas. They failed to replicate the effects of collaborative testing. The study differed from Zimbardo et al. in regards to the type of students recruited for the study. Zimbardo and colleagues worked with students who came from Stanford University whereas Haberyan and Barnett worked with students who came from a moderately selective regional university. Students' academic abilities, competitiveness and sophistication might have differed between the two schools. Students in the Haberyan and Barnett study were not used to studying cooperatively; they typically studied alone while many students in the Zimbardo et al. study reported using sophisticated collaborative study strategies. This led Haberyan and Barnett to question whether the benefits of collaborative testing might be due more to the preparation students get working cooperatively before the test than to the collaboration that occurs during the test. In addition, the tests used by Haberyan and Barnett consisted of half multiple-choice items and half essay items; Zimbardo et al. used tests that were entirely composed of multiple-choice items. Thus, "it is possible that test format plays a more role in the benefits of working with a peer" (Haberyan & Barnett, 2010, p. 37).

To better understand what factors facilitate test performance in a collaborative testing setting, Haberyan and Barnett (2010) tried to "tease apart the influence of collaboration at the time of studying, testing and the interaction between the two" (p. 38) by using a test that was based on a 2000 word text describing violence in the workplace.

The test itself consisted of 10 questions to measure comprehension of the passage: 4 factual multiple-choice items and 6 short essay questions. Students enrolled in a general psychology course were randomly assigned to two groups: study alone or study in pairs. Both groups got 20 minutes to study the text. After the 20 minutes, students who studied alone were again randomly assigned to two groups: a test alone group and a test with a partner group. The students who studied with a partner were also randomly assigned to two groups: a test alone group and a test with a partner group. The results showed that collaborative testing improved performance compared to individual testing, and the testing effect was independent of study conditioning; students studied with a partner or studied alone. This difference was statistically significant.

Hicks (2007) explored implementation of collaborative learning and group testing within a radiologic technology curriculum and to see if the terms *cooperative learning* and *group testing* were familiar to students. Hicks also wanted to determine if there was a change in students' perceptions after the students engaged in collaborative learning and participated in a group testing situation. Undergraduates enrolled in a radiologic technology program preparing for a certification exam (American Registry of Radiologic Technologists, ARRT) were surveyed at the beginning of the semester and at the end. Collaborative groups were formed based on students' scores from a content knowledge exam they took on the first day of class. Groups were formed based on content category; the group leader had the highest score in that category, and two students had low performance scores while other group members' scores were average. Each group had to present a 35-45 minute review of information from their particular content category to class. Results from the survey showed that a significant difference was found between

students familiarity and understanding of the term cooperative learning and group testing. The mean was higher for the postsurvey. All survey items showed an increase in mean score but not all were statistically significant. There was an increase in students' willingness to participate in or accept the active learning style. A majority of students were neutral regarding their overall experience with collaborative learning and group testing compared to a traditional lecture –based style. They were also neutral regarding prospective benefits of collaborative learning and group testing in preparing for their ARRT certification exam (Hicks, 2007). The author mentions that because students had not taken the certification exam when they took the postsurvey, they might have been neutral regarding the prospective benefits of collaborative learning and group testing in preparing for their ARRT certification; this might have changed had they taken the exam. Anecdotal evidence from the instructor's observations showed that students did take part in the collaborative learning process and there was a lot of interaction that took place in the classroom.

Use of nonconsensus in collaborative tests. In three studies (Breedlove et al., 2004; Giraud & Enders, 2000; Kapitanoff, 2009) where nonconsensus on collaborative tests was required, groups or pairs worked together either before or during the test, but each student turned in his or her own test and individual grades were assigned.

Giraud and Enders (2000) looked at the practice of using collaborative testing in an introductory statistics class by comparing test scores from two consecutive summer classes at the undergraduate level. Collaborative methods were used in the first class, and students received test forms with item stems but no answer choices. Students were given 15 minutes to discuss these items before being asked to complete the exam individually.

In the second class—taught the following summer--students received the same test to complete individually but did not engage in collaboration prior to completing the test. The only significant difference found between the two classes was in the students' attitudes toward collaborative testing; their attitudes became more positive after each test administration.

Breedlove, Burkett, and Winfield (2004) sought to determine if academic achievement is affected by collaborative testing alone without having students engage in collaborative learning. They also attempted to investigate whether a relationship exists between collaborative testing and performance on different types of questions. Breedlove et al. administered collaborative exams in five sections of the same introductory sociology undergraduate course offered Fall 2001 and Spring 2003. The authors used one control group per semester (i.e., one group who did not receive collaborative exams). The second exam was collaborative for the fall course, and exams 1 and 2 were collaborative for the spring course. Different instructors taught the courses, and tests were different except for 15 multiple-choice questions that were the same across the tests; these items were the ones used for analysis. Seven of the questions were classified as concept questions and seven as theory questions. Students worked in the same randomly assigned same-sex pairs while taking the exam. Exam scores from the collaborative and individual exams were used for analysis; they showed no significant difference between the experimental and the control groups in test performance on individual exams. There was however, a significant difference on scores on the concept questions between taking exams collaboratively and individually. Those that took the collaborative exam scored higher. No significant difference was found on scores for theory questions between

collaborative and individual exams. Breedlove et al. concluded that the relationship between collaborative testing and test performance is significant and positive, but it does not apply to questions that represent higher levels of cognitive processing. This study explored whether collaborative testing alone could influence test performance when controlling for collaborative learning. However, there is no mention of how the effects of collaborative learning were eliminated. The fact that students were obviously engaged in collaborative learning while working on the collaborative test begs the question of whether collaborative testing can exist without collaborative learning. Is collaborative testing nested within collaborative learning? These questions remain unanswered in the literature. Also, no information was given about the teaching methods used in the class. It is hard to determine if the observed results are based solely on collaborative testing or a combination of both collaborative testing and learning.

Kapitanoff (2009) studied the mechanisms that take place in collaborative testing.

The goals put forth in her study were to

(1) Replicate results demonstrating enhanced test performance using collaborative testing, (2) Examine the cognitive processes used by students in collaborative testing and determine how they are related to testing outcomes, and (3) Determine the relationship between self-reported anxiety, collaborative testing, and enhanced test performance. (p. 60)

Undergraduates from two psychology courses participated in the study. In one course—an introductory psychology course—the second exam was collaborative; in the other course—a cross-cultural psychology course—the first and the third exams were collaborative. Students first completed 50 multiple-choice items and 15 short-answer items individually. Twenty multiple-choice items were randomly selected and used for the collaborative exam. Students were randomly paired and worked on the collaborative

part straight after the individual part. Grading was based on a weighted formula: 60% from the individual portion and 40% from the group portion. Students turned in individual answers for both parts. Exam scores and a preexam questionnaire asking about test anxiety, study behavior, confidence, and attitudes toward multiple-choice exams and collaborative testing were used for the analysis. Results showed that there was an 8.4% mean score gain in the collaborative testing group. The effect size for change in exam scores was high (Cohen's $d=0.77$). There was a significant difference in terms of the reduction in test anxiety with the collaborative test, with a reported effect size of $d=1.14$. There was a general correlation present between the cognitive processes students used while taking the collaborative exam. But different patterns exist for example the cognitive processes; processing of information, helping to fill in gaps, remembering information, thinking through the question, and understanding what the questions were asking were all highly correlated with each other (Kapitanoff, 2009). This study was limited by its small sample size of 33 and the lack of a control group.

Use of both consensus and nonconsensus in collaborative tests. In four studies (Giuliodori et al., 2008; Ioannou & Artino, 2010; Rao et al., 2002; Simkin, 2005) where both consensus and nonconsensus on collaborative tests were used, students took the same test twice—first individually and then as a group. Two answer sheets were turned in—an individual and a group part. Grades were based on combination of those two parts. In one study, students worked collaboratively on exams, but one section turned in individual answers and the other turned in group answers (Shindler, 2004).

To see if students' understanding of the material would be enhanced by collaborative learning and by receiving immediate feedback, Rao, Collins, and DiCarlo

(2002) had students enrolled in a postbaccalaureate program in medicine take four collaborative quizzes. The questions on the quizzes varied from fill in the blank questions, single best response multiple-choice questions, short essay questions, and true/false statements. Students first answered quizzes individually for 30 minutes and then were assigned to groups of 2 to 3 that answered the same quiz together. The quiz grade was based on 80% individual and 20% group contributions. The results showed that there was a significant difference in scores between the quiz formats, with group scores being higher for all quizzes. The largest difference was for single best response multiple-choice items. Students liked the individual quiz followed by the group quiz better than the traditional method (which consisted of only an individual quiz). Students also reported that their understanding was enhanced along with the opportunity to improve their scores. One of the weaknesses of this study was its lack of description on how groups were formed for each quiz. Information was also lacking on how students were selected to groups. The researchers also did not include a control group in the study, which would have strengthened the design considerably.

Shindler (2004) examined the soundness of using a collaborative essay exam for students enrolled in a method courses offered by a graduate level teachers candidate preparation program. The courses were offered at two different institutions. Five different sections of methods courses were used. In one institution, students worked collaboratively on the exam but turned in individual answers. At the other institution, students worked collaboratively and only one answer was turned in per group. The exam consisted of three items that required synthesis and application of the course content. Results were based on (a) a survey that was implemented to measure the soundness of

using collaborative essays exams, and (b) focus group interviews with volunteers from the courses. A four-dimensional framework consisting of validity, reliability, efficiency and effect on the learner was used to measure the construct of soundness. Results showed that students perceived that using collaborative exams aligned favorably to the four dimensions of soundness. Students reported that the collaborative format was a more valid form of assessment compared to taking exams individually, and it was as fair and more efficient way to assess the content in the course. They also believed that it did “provide a context that was more aligned with actual effective teaching practice” (p. 273). The study did not mention the group sizes used, and it did not look at the difference in using nonconsensus or consensus on exams. Apparently, the reason students turned in only one exam per group was because the class sizes were large (over 120 students per class). It would have been interesting to see if there was a difference in using the two conditions in terms of students’ scores.

To see if group testing would help students’ testing performance and more specifically, to see how students performed on group tests compared to individual tests. Simkin (2005) had undergraduates enrolled in his information systems class take one group test and one group exam. The difference between the group test and exam here is that the exam was longer. The test consisted of 15 multiple-choice questions. Students first took the test individually and then took it with a group, and they were free to form groups with a maximum of three members. Groups worked on the same test. The exam was in two parts, and two lessons were used for the exam. The first part was made up of 40 multiple-choice questions, and students took the exam first individually and then as a group. The second part was made up of 60 points in class programming (constructed

response items). The same method was used as before in that students first took the exam individually and then worked in groups. Guessing was allowed on both the test and the exam. Students turned in individual answers and groups returned one answer sheet for the test and the two parts for the exam. Simkin also explored students' attitudes towards the testing method by having students answer a survey. Test scores showed that there was a significant difference on the grades; students did better on group tests. For the exam, 90% of students did better on both group exams than on the individual exam. This difference was significant. Responses from the survey showed that all students wanted more collaborative assessment in the future, and, also most students reported that they had learned something from their group members. All but two students thought that everybody contributed to the discussion. Simkin claimed that the study was an experiment even though there was no control group or randomization present, and this was not addressed as a limitation. In addition, the author mentioned his lack of experience with collaborative testing; he wanted to utilize it, but, at the same time, he was reluctant to devote too much time on it.

Giuliodori, Ljuan, and DiCarlo (2008) looked at the effect of collaborative group testing on test performance for both high and low performing students taking a veterinary psychology course in Argentina. Students were given the option to participate in the collaborative test. They first completed an individual test and then answered the same questions in pairs. At the end, the instructor projected the same questions on the board and students provided answers to them out loud in class. There were three exams. Scores from the individual and group tests were compared, and this revealed that group scores were higher than individual scores 76.8% of the time, for all exams. Scores for low

performing students increased more than high performing students. The effect size for students' performance in collaborative testing for the three exams was 0.78; for high performing students, the effect size was 0.22 and for low performing students, the effect size was 1.38. Students' also completed an evaluation of the collaborative testing, and their responses were positive.

Giuliodori et al. (2009) used the same data to see whether students with correct or incorrect answers on the individual tests were more or less prone to change their answers on the group tests. They also wanted to determine whether low- or high-performing students had a greater impact on the group response. The results showed that it was more likely that students with incorrect answers on the individual exam would change their answers on the collaborative exam (compared to students who had correct answers on the individual exam). A total of 22% of the individual responses were changed in the collaborative testing, and 77% of those changes were from students who had incorrect responses on the individual test. It is more beneficial in collaborative testing to have two students with different answers instead of having the same answer because students with incorrect responses on the individual test were 7.58 times more likely to change answers when taking the collaborative test. Only 8% of the students who had correct responses on the individual test changed answers during the collaborative exam. According to these researchers this result demonstrates that it is more important to have a student with a correct answer on the individual part to convince their partner to select the right answer on collaborative test than it is to have a high performing student on the collaborative test. This study did not include a control group or any randomization. The classification of high and low students was based on individual test scores and there was not a cut-off

point for low or high scores; instead, it was relative to how their partners scored. Students with higher scores were classified as high performing, and if partners had the same scores, they were not included in the analysis. Students in a duo could both have had relatively high scores, but one person would still be classified as “low.” This does not give reliable information regarding actual low or high performing students.

In a more recent study, Ioannou and Artino (2010) implemented one collaborative exam in their undergraduate introductory educational psychology course in an effort to endorse collaborative learning, decrease test anxiety, and make classroom testing a more positive educational experience for their students. Each question on a multiple-choice exam was projected to the class and students were given 1 minute to answer the question individually. After this, 1 minute was given for groups to collaborate on answers. Both individual and group answers were reported but a consensus between the two was not required. The individual answers were 40% of the grade and the group answers made up the remaining 60%. The mean scores from the group answers were statistically higher than the individual scores, with an effect size of 1.06. A majority of the students reported not studying differently for the collaborative exam, while 23% of students reported studying less because they felt that the group would help them. Students’ attitudes towards the collaborative exam were found to be positive and students’ reported anxiety levels were generally low. Based on the results from both the survey that asked about the collaborative assessment and the exam scores, the authors concluded that the experience supported the idea that collaborative assessment is “beneficial for learning, ...more enjoyable and generally less stressful than a regular test. And ...probably ‘as fair’ as a regular exam” (p. 195). The study lacked a control group, groups were not randomized,

and there was no mention of students' familiarity with working in collaborative groups prior to the exam.

Summary and Discussion

The effectiveness of cooperative learning has been well established by researchers and confirmed by meta-analysis (e.g., Johnson et al., 2008; Johnson et al., 2000; Roseth et al., 2008). The studies reviewed in this paper on collaborative testing illustrate that different test formats exist and that collaborative tests can be implemented in a variety of ways. In some studies, collaborative tests required a consensus among group members regarding answers; in other studies, the nonconsensus method was used, or a mix of consensus and nonconsensus formats was explored. Group size varied from 2 to 6 students, and in some cases, students took both an individual and a collaborative test. Assignment to groups was sometimes random, but, in some studies, students selected their own groups or were placed in groups in a purposeful manner by the instructor. Tests often consisted of either multiple-choice or essay-type questions. Despite the difference in test formats, a majority of the studies reported an increase in test scores for students taking collaborative tests (e.g., Breedlove et al., 2004; Helmericks, 1993; Sandhal, 2009; Zimbardo et al., 2003), a decrease in test anxiety (e.g., Ioannou & Artino, 2010; Zimbardo et al., 2003), an increase in student attitudes (e.g., Giraud & Enders, 2000; Ioannou & Artino, 2010), and an increase in student interaction (e.g., Ioannou & Artino, 2010). The effectiveness of collaborative tests is evident based on the studies reviewed here. However, it is also apparent that many questions regarding the implementation and validity of the research methods used in these studies remains uncertain. For example, the quality of these studies differed from not including any research questions and

significance tests (e.g., Helmericks, 1993) to not including a control group (e.g., Giuliodori et al., 2008; Hicks, 2007; Ioannou & Artino, 2010; Kapitanoff, 2009; Rao et al., 2002; Shindler, 2004; Simkin, 2005) and not randomly assigning students to groups (e.g., Giuliodori et al., 2008; Ioannou & Artino, 2010; Shindler, 2004; Simkin, 2005; Zimbardo et al., 2003).

Based on the review of studies on collaborative testing, it becomes apparent that improved ways to study collaborative testing are needed, and future studies should at least use control groups and randomly assign students to groups in order to be able to make inferences about the effectiveness of this testing method. Although the effectiveness of using collaborative methods has been established, improvement is needed in terms of applying experimental design principles to the study of collaborative testing. Further, most of the studies on collaborative testing have involved comparing collaborative to individual testing. Now, the focus should be more on the format of the collaborative testing procedure in order to explore the efficiency of using different exam formats. For example, what differences might we see if we require a consensus or not during the collaborative process, and how might different formats (i.e., consensus or nonconsensus) affect group processes? In addition, the effect prior experience working in collaborative groups might have on collaborative tests has not been studied and remains unclear. Also, the types of quizzes or exams best suited for use in collaborative testing situations (e.g., multiple-choice, essay, open-ended, etc.) are not well-established. More studies on collaborative testing that address the areas of test format, test type used, and group size are warranted to better understand how this testing method might benefit student learning.

It should also be noted that the research reported thus far on collaborative learning has focused on face-to-face classroom instruction at the university level. Collaborative learning methods have also been used in online instruction, and the next section will review the research on online instruction and, more specifically, the use of collaborative learning methods in online college courses.

Online Instruction of College Courses

Distance learning has been conceptualized “as the deliberate organization and coordination of distributed forms of interaction and learning activities to achieve a shared goal” (Dabbagh & Bannan-Ritland, 2005, p. 12). Online education is one form of distance education. Online education uses the Internet as the medium of delivery, while distance learning uses the Internet or other forms of delivery (e.g., television, satellite, video, correspondence, etc.). In both forms, education can take place at the same time but in a different place, and this is called synchronous distance learning. This occurs when students meet at the same time but in different locations, either online or through satellite, television or other available technology. In asynchronous distance learning, students learn at different times in different places. Asynchronous distance learning has been referred to as the “cleanest” or “purest” form of distance education. It is where students choose when and where to learn and access the instructional materials (Simonson, Smaldino, Albright, & Zvacek, 2000). E-learning is a term that has been used synonymously with online learning in the literature (Dabbagh & Bannan-Ritland, 2005). In this paper, online learning will be used as an umbrella term for all education that takes place online.

In the 2010 report *Learning on Demand: Online Education in the United States* (Allen & Seaman, 2010), an online course is defined as “a course where most or all of the

content is delivered online” (p. 4). These courses typically have no face-to-face meetings, and 80% or more of the content is delivered online. A hybrid or a blended course is defined as a course that has from 30-79% of its content online, and it meets partially online and partially in face-to-face environments (Allen & Seaman, 2010).

Dabbagh and Bannan-Ritland (2005) define online learning as “an open and distributed learning environment that uses pedagogical tools, enabled by Internet and Web-based technologies, to facilitate learning and knowledge building through meaningful action and interaction” (p. 15).

Online education became much more accessible with the large-scale introduction of the World Wide Web in 1992. It allowed easy use of features and capabilities to present multimedia, which resulted in an expanded range of disciplines offering courses online (Harasim, 2000). Enrollment in online courses in higher education in the USA has been increasing since 2003. In the fall of 2008, over 4.6 million students were taking at least one online course. It is estimated that more than 1 in every 4 students in higher education now takes at least one online course as a part of their curriculum (Allen & Seaman, 2010).

Collaborative Learning Online

After reviewing the literature on interaction in distance education, Mahle (2007) concluded that interaction is the most important component of effective distance education. Interaction can both be instructor-to-student and student-to-student interaction. In these terms, Simonson et al. (2000) state that “interaction is *needed and should be available*” (p. 82), and it should not be viewed as “the ‘end all and be all’ of learning” (p. 82). The nature and quality of the interaction is what matters. For example, effective

learning can be diminished as much with forced interaction as with its absence. The term social presence is important in this sense. Social presence has been defined as how much “a person is perceived to be ‘real’ in communication that is conducted via the use of some form of media” (Williams & Christie, 1976, as cited in Palloff and Pratt, 2007, p. 30). In terms of interactions, Palloff and Pratt (2007) acknowledge the importance of using collaboration learning in online instruction when they write that the “key to the learning process are the interactions among students themselves, the interactions between faculty and students, and the collaboration in learning that results from these interactions” (p. 4). According to Palloff and Pratt, building a community of learners is one of the key features of successful online learning. This notion has been supported by a large body of research (see in Palloff and Pratt; Garrison, n.d.; Rovai, 2002; Rovai and Jordan, 2004; Shea, Swan, and Pickett, 2004 and Wegner, 1999). Learners gain a deeper level of knowledge-generation with the help of collaborative efforts while moving from independence to interdependence, and this strengthens the foundations of online learning communities.

Collaborative learning has been used successfully in online education (Roberts, 2004). Factors such as characteristics of group members, group member preparation, discussion topics, and the quality of discussion have been shown to be related to students’ satisfaction with the collaborative group process (Jianxia, Durrington, & Mathews, 2007). Students’ familiarity and the ease of using the online medium are also important in creating an effective collaborative online environment (Curtis & Lawson, 2001). It has also been shown that when the impact of structure on the collaborative learning process is explored, more highly structured activities help enhance high levels of participation and

develop students' ability to work in groups and to reflect with others (Pozzi, 2010). Shen, Hiltz, and Bieber (2006) found that by including small group activities in the online learning process, interaction was significantly increased among students and their perception of the online learning community was improved.

Terms used for collaborative learning online. In the literature on online education, terms such as online group work and computer-supported collaborative learning (CSCL) are often used interchangeably with the terms collaborative or cooperative learning. In CSCL, learning is believed to take place through social interaction using a computer or through the Internet and it can be implemented both online and in a classroom (Stahl, Koschmann, & Suthers, 2006). As stated before the term collaborative learning will be used, in this paper, as an umbrella term to encompass learning that involves peer and group learning, where students work together to maximize both their and each other's learning (both online or in a face-to-face classroom).

Collaborative learning in online courses. The practice of effective collaborative learning in face-to-face classrooms does transfer into the online classroom. However, because of the difference between the learning environments, it is more challenging to incorporate collaborative learning assignments in online courses (Palloff & Pratt, 2007). Because of the differences between the collaborative learning environments in online and traditional classrooms, other instructional design considerations are necessary (Kieser & Golden, 2009). Different frameworks have been suggested for how to create a successful collaborative online learning environment.

According to Dabbagh and Bannan-Ritland (2005), for a successful collaborative online environment to exist, students need or should develop abilities in four skill areas:

social learning, discursive/dialogic, self-and group evaluation, and reflection. The aforementioned skills help establish a truly collaborative online learning environment and support equal construction of knowledge and negotiation of alternative perspectives. Students possessing these skills are active in group work and they rely on the process of shared knowledge construction with their teachers and peers in order to reach a meaningful understanding of the concept under study. Students gain these skills through the design and implementation of suitable instructional strategies and learning activities. In online learning environments, collaborative learning is mostly achieved by using asynchronous and synchronous communication, along with document-sharing and groupware¹.

Palloff and Pratt (2004) provide practical guidelines for online instructors to design, implement, facilitate, and evaluate online collaborative learning activities in what they call the Stages of Collaboration. Because “collaboration does not just happen” (p. 19), it is important for instructors to plan well when they want to incorporate successful collaborative learning in their courses. Instructors should first *set the stage*, which involves preparing students for the collaborative work that will be expected of them and inform them about the importance of it. Secondly, they should *create the environment*, or the place where the students will meet online (e.g., a discussion board). Instructors inform students on where they should meet and how they should communicate. The third stage is to *model the process*; it is important that the instructor shows commitment to the learning process by modeling collaborative behavior in the course. The fourth stage is to

¹ “multiuser software that enables synchronous and asynchronous communication and document sharing and production” (Dabbagh & Bannan-Ritland, 2005. p.15).

guide the process, and this is where the instructor shows his or her responsibility towards the collaborative learning process taking place in the course. In this sense, it is important that students are made aware beforehand about how the instructor is going to be involved in the collaborative activities taking place in the course. Finally, the instructor should *evaluate the process* by providing students with evaluation upon the completion of a collaborative activity or an event.

Implementing collaborative activities in an online course does require advanced planning and designing from the instructor, but the instructor's work does not end once the course begins and students start working. The instructor is responsible for guiding the collaborative activities throughout the course in order for successful collaborative learning processes to take place (Palloff & Pratt, 2004).

Methods of evaluating collaborative learning online. Different models have been used to evaluate collaborative learning in online courses (Garrison & Anderson, 2003; Henri, 1992; Pozzi, Manca, Persico, & Sarti, 2007; Weinberger & Fischer, 2006). It is important that more than one indicator be used to evaluate the learning and that the "evaluation occur in contexts that are as rich and complex as the instructional environments" (see Jonassen, 1991, as cited in in Dabbagh & Bannan-Ritland, 2005, p. 239). In this paper, two models will be explored: the Pozzi et al. (2007) model that builds on older models from Garrison and Anderson (2003) and Henri (1992), and the Weinberger and Fischer (2006) model.

After testing and modifying a framework that builds on older models (Garrison & Anderson, 2003; Henri, 1992), Pozzi et al. (2007) proposed a model to evaluate and monitor the CSCL processes. The framework is a four-dimensional approach that

includes participative, social, cognitive and teaching dimensions that take place in a learning community. To express these four dimensions, indicators consisting of both qualitative and quantitative variables have been identified for each dimension in the framework. For the participative dimension, the indicators are the frequency of active action per student (e.g., how many sent messages, uploaded documents, etc.), the frequency of reactive actions per student (e.g., reading messages, downloading documents, etc.) and the level of continuity in participation across time. The framework consists of interaction analysis techniques and content analysis of messages posted by students in the online environment. One of the benefits of the course management systems used for online classes is the ability to collect quantitative data such as how many times students log in, number of posts per student, and time spent online, using nonintrusive methods. Depending on the purpose of the evaluation, the model can be used to focus on some or all of the dimensions at the same time. The indicators in that sense are not seen as stable; they may in fact vary in weight depending on the context and goals of the analysis. For example, when exploring the collaborative activity in a course, more focus would be on indicators related to the participation and the social dimensions (Manca, Persico, Pozzi, & Sarti, NA).

Weinberger and Fischer (2006) propose a multi-dimensional framework to analyze knowledge construction in CSCL. The framework consists of the four dimensions of participation, epistemic, argument, and social modes of co-construction. The framework is based on analysis of discourse of collaborative learners that can take place in asynchronous discussion boards and it “is guided by an explicit or implicit theoretical framework on what processes and outcomes are seen as relevant for

collaborative learning to be beneficial for the group and the individual” (p. 72). The theoretical background depends on the assumption that through argumentative knowledge construction, learners participate in discourse activities, and that the acquisition of knowledge is related to the frequency of these multi-dimensional discourse activities. The four dimensions provide different kinds of qualitative and quantitative information. The participation dimension seeks to gather information regarding learners’ participation measured by quantity of participation and also to see if the participation is on an equal basis measured by the heterogeneity of participation. In the epistemic dimension, the analysis is geared towards the content of the learners’ contribution, for example, by examining if learners discourse is on-task, which is when learners attempt to solve the task with their contributions. In the argument dimension, the discourse is analyzed, for example, by exploring the construction of arguments and sequences of arguments. Nonargumentative statements can also be identified and differentiated. The social mode of co-construction dimension is about how much learners refer to the contributions of the learning partners. Examples of that would be asking questions of learning partners and reaching a consensus regarding the contribution of all learning members in order to continue with the discourse.

The difference between these two frameworks is that the Pozzi et al. (2007) framework is designed to evaluate and monitor the CSCL process, making its application broader than the Weinberger and Fischer (2006) framework, which explores the knowledge construction in CSCL. In the Pozzi et al. framework, the focus can be on some or all of the dimensions at the same time depending on the research question under

study, while in the latter framework, all four dimensions should be explored when determining knowledge construction in the CSCL.

Collaborative Testing Online

A literature search regarding online collaborative testing was conducted using the following keywords: collaborative testing online, collaborative assessment online, collaborative examination online, collaborative exams/quiz(zes) online, cooperative testing online, cooperative assessment online, cooperative examination online, cooperative exams/quiz(zes) online, CSCL testing, CSCL assessment, CSCL exams/quiz(zes), CSCL examination, group exams/quiz(zes) online, group assessment online, group examination online. This search revealed four articles and three conference proceedings by the same group of authors: Shen, Hiltz, Bieber, and Swan. The format of the collaborative exams reported in these publications differs from the kinds of collaborative exams typically used in the face-to-face classroom (as previously reviewed in this paper) in that groups of students worked on the question design and grading phases, but students completed and turned in exam answers individually (Shen, Hiltz, & Bieber, 2006; Shen, Hiltz, & Bieber, 2008; Swan, Shen, & Hiltz, 2006).

Shen et al. (2006, 2008) and Swan et al. (2006) reported that as a result of using collaborative examinations online, surface learning was significantly reduced, there was an increase in students' perceived learning and interactions, and the perception of the online learning community for students was improved. The collaborative examination consisted of having students first design questions and grading rubrics individually and then in groups of 3 to 5 a consensus was reached regarding reviewing, revising and grading the questions. The instructor also reviewed and revised the questions made by the

students and then assigned them to students. At the end, the instructor assigned the final grades.

Based on the literature search no articles were found that involved having students work together in groups or pairs to complete exams or quizzes in online courses. In addition, a literature review from 2006 that summarized research on online teaching and learning located 76 studies on online courses that were categorized into the topics of course environment, learners' outcomes, learners' characteristics and institutional and administrative factors (Tallent-Runnels et al., 2006). Of those 76 studies, none appeared to be related to collaborative assessment or testing.

Summary and Discussion

Online education has evolved immensely in the last two decades and has led to increased enrollment in online courses in higher education in the United States. This is a trend that is only expected to increase in the future. Numerous researchers (e.g., Palloff & Pratt, 2007; Roberts, 2004) have pointed out the importance and effectiveness of using collaborative learning in online education. To implement successful collaborative activities in online courses, instructors need to carefully plan and design the activities along with monitoring them as they take place (Palloff & Pratt, 2004). Frameworks to evaluate different aspects of collaborative learning in online courses have been designed and used (Pozzi et al., 2007; Weinberger & Fischer, 2006).

The only research that has been reported regarding the use of collaborative testing in an online environment does not involve students working collaboratively on completing tests; instead, students work independently and in groups on designing tests and grading rubrics. This has led to positive outcomes such as reducing surface learning

and increasing students' perceived learning and interactions (Shen et al., 2006, 2008). It is apparent that further research regarding collaborative assessment and collaborative testing within the online environment is sorely needed. There currently appears to be no available information on using collaborative tests in online courses, and thus any discussion regarding collaborative testing in the online environment is purely hypothetical. Bearing in mind that (a) this testing method has been shown, in the face-to-face classroom environment, to have a positive impact on students' grades, attitudes, and anxiety levels (e.g., Breedlove et al., 2004; Giraud & Enders, 2000; Helmericks, 1993; Ioannou & Artino, 2010; Lusk & Conklin, 2003; Sandhal, 2009; Zimbardo et al., 2003) and (b) the importance of using collaborative learning in the online course has been pointed out and affirmed by research (e.g., Palloff and Pratt, 2007; Roberts, 2004; Shen, Hiltz, and Bieber, 2006), the next step would be to see if the benefits of using collaborative testing in face-to-face classrooms also apply to the online classroom. Because of the unique learning environment in the online course, it is important to see how collaborative exams would be implemented successfully, and how they might differ from face-to-face classroom settings.

The next section will focus on collaborative learning with respect to the college-level introductory statistics course and, more specifically, on the online college-level introductory statistics course.

Collaborative Learning in Introductory College Statistics Courses

In an early paper on this topic, Garfield (1993) describes three studies that examined the use of collaborative learning in college statistics courses. One study (see Shaughnessy, 1977 in Garfield 1993) found that a small group model appeared to help

students overcome some misconceptions about probability, and it also enhanced students' learning of statistics concepts. The second study (see Dietz, 1993 in Garfield, 1993) demonstrated that using collaborative learning activities about methods of selecting a sample allowed students to invent for themselves a standard sampling method, which resulted in students' better understanding of these methods. The third study (see Jones, 1991 in Garfield, 1993) showed that when collaborative learning techniques were introduced in several statistics courses, researchers observed dramatic increases in attendance, class participation, office visits and student attitudes.

Since that time, six additional studies examined the effects of using collaborative learning in statistics courses at the undergraduate level. Two studies (Girard, 1997; Potthast, 1999) compared test scores from different sections of the same course, where collaborative learning was used in one section and individual learning and lectures were used in another. Both of these studies found that students who took the collaborative sections performed better than students in the traditional sections. Girard (1997) also concluded the collaborative classes were especially beneficial for the students who had the least statistics background.

Three studies (Magel, 1998; Keller & Steinhorst, 1995; Perkins & Saris, 2001) compared tests scores from a course using collaborative learning to a previously taught traditional course that did not include collaborative learning. Magel (1998) and Keller and Steinhorst (1995) found that students in the collaborative learning course received higher final grades than students in the traditional courses. These two studies also reported that the collaborative classes were more active and students were more engaged in those courses. Keeler and Steinhorst (1995) noted the completion rate was higher in

the collaborative learning course compared to the previous course where collaborative learning was not used. Perkins and Saris (2001) used one specific collaborative learning technique, the Jigsaw method. Scores from four exam items that were related to the Jigsaw activities were compared to scores from previous years when the Jigsaw method was not used in the course. The results revealed no difference. However, overall scores on exams were higher for the Jigsaw course. Students' evaluation of the Jigsaw was also positive; students especially liked the alternative teaching method compared to lectures.

Delucchi (2006) wanted to see if using collaboratively designed group projects could enhance students' learning of statistics. A statistically significant association was found between a group project and the final exam, which indicated that students who received a higher grade on the group project "increase their learning of the material and score more points on the final examination than students earning lower group projects grades" (p. 246). However, students also took individual quizzes over the semester, and these quiz scores were more strongly related to the final exam score than the group project. Because of this, Delucchi remained skeptical about whether collaborative groups do in fact enhance student learning.

These six studies that utilized collaborative learning in statistics courses lacked some fundamental principles of experimental design such as a control group (Delucchi, 2006; Magel, 1998; Keller & Steinhorst, 1995; Perkins & Saris, 2001) and random assignment to classes or groups (e.g., Delucchi, 2006; Keller & Steinhorst, 1995; Perkins & Saris, 2001; Potthast, 1999). Also, formal group structure based on collaborative learning techniques, which enhances positive interaction and individual accountability,

were often missing, or there was a lack of information regarding the group structure (Delucchi, 2006; Giraud, 1997).

Implementing Collaborative Learning in Statistics Courses

Roseth, Garfield, and Ben-Zvi (2008) offer practical ways to apply a collaborative framework in an introductory statistics classroom. It resides on the notion that “statistics instruction ought to resemble statistical practice, an inherently cooperative enterprise” (p. 1).

When implementing a collaborative lesson, the following four steps should be involved (Roseth et al., 2008). The first step is *making preinstructional decisions*, by deciding on the academic objectives of the lesson. Social skills objectives should also be determined by deciding on which interpersonal and small group skills are to be emphasized in the lesson. The second step is *explaining the task and cooperative structure to students*. In this step, students are assigned to groups or pairs, and the criteria for how to successfully complete the activity is discussed (e.g., by explaining to students how their performance will be evaluated and by emphasizing the individual accountability of each student in the group). This also involves structuring the cooperation goals of the activity, which adds to the positive interdependence of students. An example would be when students are taking a group quiz, they have to make sure that each member understands the problems and that if each member turns in one copy, only one will be graded for the entire group and that will be the whole grade for the group. When implementing collaborative assessment it is important that all group members understand the concepts and topics addressed in the assessment (Garfield et al., 2011). The third step is *monitoring and intervening during the cooperative activity*, where

instructors should observe interactions among group members and intervene when appropriate. An example of this would be asking the group questions if they are not interacting, and also assessing the learning progress and use of interpersonal and group skills. The fourth and final step is *group processing* where group members process, reflect on, and evaluate how the work on the assignment went. In a classroom setting, the instructor might ask students to identify things that were helpful.

These general steps are put forth for use in the face-to-face classroom, but they might be used in online courses as well. The process involved in the first two steps would be similar, but the last two steps would need to be adjusted to the online environment, especially in a course that is asynchronous. The group processing would be different since it would not be synchronous, but students could still be asked to identify or evaluate their group work after each assignment.

The use of Wikis, or websites where users can add, edit, remove and change content, has also been suggested as a way to promote successful collaborative learning in a statistics classroom (Ben-Zvi, 2007). This is a format that would work both in a face-to-face and an online classroom environment.

Studies on Using Collaborative Testing in Introductory Statistics Courses

As noted earlier, Helmericks (1993) and Giraud (2000) were the only references found related to using collaborative testing in introductory statistics courses at the undergraduate level. These two studies differed in terms of how groups were selected and whether tests/exams were turned in individually or as a group. Both studies compared test scores in one class with scores in previously offered sections of the same course, but no significant differences were observed between scores for the classes using collaborative

methods and those using traditional teaching methods (like lecture). However, Giraud reported a significant difference in students' attitudes towards collaborative testing for the collaborative course. In the Helmericks' study, students' evaluations of the collaborative testing showed that 90% of the students strongly favored that method. The author also stated that using collaborative testing in the classroom generated "an open and relaxed atmosphere for learning in what has traditionally been a rather unyielding environment" (Helmericks, 1993, p. 296).

As mentioned in a previous section of this paper, collaborative testing has been reported to be effective in terms of increasing test scores (e.g., Breedlove et al., 2004; Helmericks, 1993; Sandhal, 2009; Zimbardo et al., 2003). However, using collaborative testing in introductory statistics courses has not been shown to be effective in this way, and this may be due to the lack of research on collaborative testing in these types of courses. It is therefore important that more studies using collaborative testing in introductory statistics courses be conducted to determine if this method can be beneficial for this environment.

Summary and Discussion

The use of active learning has been encouraged when it comes to the teaching of statistics (GAISE, 2005). One way to incorporate active learning is the use of collaborative learning in the classroom. Collaborative learning has been used in the teaching of statistics with positive outcomes such as enhancing students' learning of statistics concepts, increasing students' understanding, increasing attendance and class participation, improving students' attitudes (see more in Garfield, 1993), and higher student achievement (Giraud, 1997; Magel, 1998; Keller & Steinhorst, 1995; Potthast,

1999). Statistics education researchers have recommended practical ways to apply collaborative learning methods in statistics classrooms (Roseth et al., 2008).

Only two studies were found that have used collaborative tests in their college introductory statistics courses, and both of those studies used scores from collaborative tests to compare to individual scores and found no significant difference between the two testing methods (Giraud, 1997; Helmericks, 1993). However, there was a significant difference in students positive attitudes towards the collaborative testing method in one study (Giraud, 2000), and in the other study, students strongly favored the method (Helmericks, 1993).

In the Giraud (2000) study, students received item stems without answer choices to discuss for 15 minutes, and they then received the complete test individually. By not providing the answer choices, students might have spent valuable time trying to figure out what the answer choices were instead of discussing the correct answer. Giraud did not address this and did not indicate why he chose to use this format. The Helmericks' (1993) study lacked clear research questions, in addition to information regarding the test types used in the study (e.g., if they were multiple-choice, short-answer, etc.). Also, Helmericks did not use the same test for the control class that was taught the previous semester; the tests covered the same content, but did not have the same questions, and this makes it hard to justify a comparison of the two tests. The results from these two are not enough to determine if the benefits of using collaborative testing that have been reported by others (within other types of courses) apply also to the introductory statistics course. More research involving the use of collaborative testing within the introductory statistics

classroom is needed in order to better determine if this is a suitable testing method to be used in that environment.

Online Instruction in Introductory Statistics Courses

Research in online statistics education has mostly focused on two different areas: (1) comparisons of online or hybrid statistics course to face-to-face statistics courses, (2) researchers/instructors sharing of experiences and suggestions on teaching statistics courses in the online environment. A review of this research follows.

Research on Online/Hybrid Statistics Course vs. Face-to-Face Statistics Courses

In his study with MBA students, Gunnarsson (2001) designed a web-based graduate-level statistics course and explored students' attitudes and achievements (which were then compared to students taking the same class in a face-to-face classroom setting). No significant difference was found between students in the two sections in terms of performance. Students in the online course reported having a more positive attitude towards the course compared to their counterparts taking the in-class course.

In a study of postgraduate students in a Malaysian university who were taking an online statistics course for social science that supported problem-based learning, students' responses and reactions towards the course were explored (Hong, Lai, & Holton, 2003). A majority of students were satisfied with the course and their learning outcomes were comparable to students taking the face-to-face version of the course. Students also expressed their satisfaction with the flexibility the online course offered them. The authors reported some difficulties getting students to participate in discussion and group activities. This lack of participation was explained by students "inadequate mathematical or computer knowledge, anxiety about using computers, fear of

embarrassment and the discomfort felt when peers commented on their work” (p. 1443). Findings from this study suggest that more structure and guidance might be appropriate to enhance students’ learning when it comes to asynchronous interactions and group activities in online statistics courses.

Utts, Sommer, Acredolo, Maher and Matthews (2003) compared two sections of the same introductory statistics course, a hybrid Internet based course and a traditional course. Students’ performance in both sections was the same. However, students’ evaluations from the hybrid course were less positive in terms of how the course was organized, expectations of students, pace of course material, and the quality of the course. The authors concluded that this difference might be because students taking a traditional course are more familiar with that format and therefore more satisfied with it.

Other studies have reported no significant difference in students’ learning. Ward (2004) compared a hybrid and a traditional course in elementary statistics offered for first year business students. The only statistical difference found was in students’ attitudes toward the course, but students in the hybrid section appeared to be more positive towards the course (Ward, 2004). Schou (2007) compared students taking an online and a traditional introductory business statistics course and found no difference in students learning outcomes. There was, however, a difference in students’ attitudes towards statistics after instruction with improvement for students taking the online section (Schou, 2007). Bakker (2009) studied 4 sections of the same community college statistics course, 2 face-to-face and 2 online. The face-to-face sections were compared to the online sections. Results from using the CAOS (Comprehensive Assessment of Outcomes in a First Statistics course (see more in delMas, Garfield, Ooms, & Chance, 2007)) test both

as a pre- and posttest and a departmental final exam “indicated that students in web-based statistics courses can have levels of average achievement comparable to that of their classroom-instructed counterparts” (Bakker, 2009, p. ii).

Dutton and Dutton (2005) found that students taking an online business class performed better than students taking the same class face-to-face, when the two courses were compared. This difference was significant even when variables such as GPA and computer experience were controlled for. The study also showed that the online students tended to be older and more mature, and had higher academic performance with a range of 3 to 8% higher grades compared to student in the face-to-face course (Dutton & Dutton, 2005).

DeVaney (2010) compared the levels of anxiety and attitudes towards statistics for graduate students taking the same statistics course, either online or on-campus. At the beginning of the semester students taking the online section reported higher levels of anxiety and less positive attitudes towards statistics; however, at the end of the semester, students in both sections reported similar levels of statistical anxiety and attitudes toward statistics. Since the study did not incorporate any strategy to reduce anxiety or improve attitudes toward statistics, this noticeable decrease with the online students “provides promise that faculty and others who design statistics courses for online delivery can incorporate materials and techniques that will lessen the anxiety of students and hopefully lead to better performance” (p. 12).

Suggestions on Teaching Statistics Courses Online

In the earliest paper found on this topic, Zhang (2002) reported his own experiences teaching online elementary statistics course through the course managements

system WebCT. Even though it has been almost 10 years since Zhang presented his paper at the International Conference on Teaching Statistics (ICOTS), some of the advantages and disadvantages he described are still relevant to today's teacher of the online statistics course. Examples of advantages of online statistics courses are: convenience for both students and instructors, the absence of time and location barriers, and the ability of the instructor to "perform complicated computations and construct sophisticated graphical illustrations" (p. 3). In addition, online courses offer more flexibility in administering exams and collecting homework since they can be delivered through the course management system. Some disadvantages noted by Zhang included: students' computer knowledge, students' ability to learn to use statistical software, difficulty in motivating students, students' and teachers' communication and lack of community in the online setting. According to Tudor (2006), the two key components for a successful online course are organization and involvement. Instructor/student interaction is important to avoid the course becoming an independent study course. Small group discussions, feedback on exams, and weekly online communication with the instructors are suggested to maintain adequate interaction in an online course. Tudor reported an increase in students' satisfaction in regards to the amount of interaction they had with the professor as the author increased the variety and quantity of interactions in an online statistics entry-level masters course. Students' satisfaction increased from 75% to 99%, from the first years the course was offered until the fifth year it was offered. For the fifth year, the author added emailed exams with comments, weekly emails to students, posting weekly announcements, and having students participate in small group discussions (Tudor, 2006).

Based on his experience teaching two online sections and numerous in-class sections of the same graduate level introductory statistics course, Wisenbaker (2003) suggested clear organization in terms of weekly schedules in covering course content and setting up some kind of structure to encourage students to work collaboratively on homework. Wisenbaker concluded that the most effective technology to be used in online introductory statistics courses “is probably that which helps keep students engaged with their own efforts to learn and instructors engaged with them as they try to do so” (p. 9).

Everson (2006) described how she used small group discussion assignments in her graduate level introductory statistics course. Small group discussions were part of the whole assessment in the course. Everson provided advice about implementing small-group discussion in online courses, including giving students enough time to work on the assignments and designing discussion assignments that do not lead to just one correct answer (so students can discuss and respond to different answers). She also suggests that assignments should include clear guidelines so students know what is expected of them, and the instructor should encourage students and offer them incentives to respond to other group members’ posts, in addition to providing them with examples of how to respond to others. Finally, Everson suggests that instructors should have a strong presence in the discussion groups to let students know if they are on the right track, ask questions, comment, and help move the discussion along.

Everson and Garfield (2008) describe how collaborative discussion groups are used in their online introductory statistics courses to meet the six GAISE recommendations (GAISE, 2005). They describe students being assigned to small discussion groups where they have to complete a series of discussion assignment over the

semester, with each graded discussion assignment taking 1 week to wrap-up. These discussion assignments are designed to emphasize statistical literacy and develop students' statistical thinking by having students use real data sets to explore and discuss and by using different technologies to solve problems (with more focus on interpreting results). The authors point out that one of the main advantages of having students work in groups where they have to write their comments is that the instructors can monitor what every student is thinking; this is something that is impossible in a face-to-face classroom setting. This way, the instructor can observe how the whole collaborative process takes place in each group, making it easier for the instructor or group member to address students' misconceptions or misunderstandings.

Several educators have described the teaching of introductory statistics courses online and in some of them; the focus has been on interaction both between students and the instructor and students. Discussion assignments, chat rooms and group projects have been used for engaging interactions in these online courses (see more in Everson and Garfield, 2008).

No studies were found on the effectiveness of using collaborative learning in online introductory statistics courses. In addition, no studies were found pertaining to using or designing collaborative exams in online statistics courses. There seems to be a lack of research in the field of statistics education that explores collaborative learning in online introductory statistics courses.

Summary and Discussion

Publications on online statistics education have mostly focused on comparing online or hybrid courses to face-to-face courses. Results indicate no significant difference

in students' achievement between the two different formats (Gunnarsson, 2001; Hong et al., 2003; Utts et al., 2003; Ward, 2004; Bakker, 2009), with the exception of one study where students taking the online course performed better than students taking the same course in the face-to-face environment (Dutton & Dutton, 2005). In comparison to students taking face-to-face courses, students taking online statistics courses have reported a more positive attitude towards the course (Gunnarsson, 2001; Ward, 2004). However, another study showed that students taking a hybrid course were less positive towards the course (Utts et al., 2003). None of the studies reported here explored the effectiveness of other important characteristics of the online format, such as assessments, assignments, and the learning and teaching methods used. Researchers should focus on how to teach statistics online effectively instead of comparing online to face-to-face courses since students' achievement in the two settings has been shown to be similar or better in the online settings (e.g., Gunnarsson, 2001; Utts et al., 2003; Ward, 2004).

Several papers (Everson, 2006; Tudor, 2006; Wisenbaker, 2003; Zhang, 2002) have suggested successful ways to teach online statistics courses and have stressed issues such as the importance of establishing instructors' and students' interactions, using small-groups discussions, providing detailed instruction regarding technology, assignments, and assessment, and organization of course content. These ideas are based on researchers' and instructors' own experiences, however, rather than on empirical investigation. More research is clearly needed to provide stronger evidence for how statistics should be taught online, especially in regards to assessment, assignments, and the learning and teaching methods that might be most effective to use in an online introductory statistics course.

Summary and Implications of the Literature Review

The effectiveness of using cooperative learning has been established through research (Johnson et al., 2008; Johnson et al., 2000; Roseth et al., 2008). The usefulness of using collaborative tests has also been demonstrated, despite the existence of different formats that have been used in implementing collaborative tests (e.g., Breedlove et al., 2004; Giraud & Enders, 2000; Helmericks, 1993; Ioannou & Artino, 2010; Sandhal, 2009; Zimbardo et al., 2003). A majority of the research demonstrates an increase in test scores for students taking collaborative tests compared to individual tests (e.g., Helmericks, 1993; Sandhal, 2009; Zimbardo et al., 2003; Breedlove et al., 2004), a decrease in test anxiety (Zimbardo et al., 2003; Ioannou & Artino, 2010), an increase in students' positive attitudes (Giraud & Enders, 2000; Ioannou & Artino, 2010), and more student interactions (Ioannou & Artino, 2010). The quality of the studies on collaborative testing vary, from a lack of clear research questions and significance tests (Helmericks, 1993), to the absence of a control group (Rao et al., 2002; Shindler, 2004; Simkin, 2005; Hicks, 2007; Giuliiodori et al., 2008; Kapitanoff, 2009; Ioannou & Artino, 2010) and the absence of random assignment to groups (Zimbardo et al., 2003; Shindler, 2004; Simkin, 2005; Giuliiodori et al., 2008; Ioannou & Artino, 2010). Based on these limitations, it becomes apparent that improvement is needed in terms of how to study collaborative testing, and future studies should try when possible to use control groups and random assignment into classes and groups in order to be able to make inferences about the effectiveness of this testing method.

With continuing increase in courses offered online, and the evidence demonstrating the importance of incorporating collaborative learning in order to design

effective online courses, there comes the need to conduct research on the use of collaborative learning in online statistics courses. In addition, the lack of research on using collaborative testing online—both in online education in general and statistics education in particular—highlights the need to systematically study collaborative testing in the online statistics course.

The importance of using collaborative learning, both in face-to-face classroom introductory statistics courses (e.g., Giraud, 1997; Magel, 1998; Keller & Steinhorst, 1995) and in general online courses (e.g., Palloff & Pratt, 2007; Roberts, 2004), has been well-established, and collaborative learning methods have been encouraged as a way to incorporate active learning in statistics courses in an effort to help students better learn important statistical concepts and ideas (Garfield, 1993). The next step would be to see if the benefits found in using collaborative testing in face-to-face classrooms also apply to the online format, especially in regard to the teaching of statistics. In addition, because of the unique learning environment in the online course, it is important to see how collaborative exams can be implemented successfully, and how these would differ from the classroom setting. The research on online statistics courses has involved comparing online and hybrid face-to-face courses, and the results have shown that the online or hybrid format is as effective as the traditional face-to-face format (e.g., Gunnarsson, 2001; Utts et al., 2003; Ward, 2004). The research in statistics education now needs to move on to explore how online collaborative learning and testing can be implemented successfully in the online statistics course. This could be assessed by using the available resources and research findings from studies conducted in the field of online education that have shown the benefits of using collaborative learning in online courses.

The literature presented here provides some clear indications of the effectiveness of using both collaborative learning in teaching and in utilizing collaborative tests. However, the significant limitations of the present review demonstrate that further research is necessary regarding the effectiveness of using collaborative tests in introductory statistics courses and the effectiveness of using collaborative testing in terms of online education is warranted.

Collaborative test formats, as mentioned earlier in this review, are different in many ways. For example, some tests require group consensus and some do not. Having students reach a consensus regarding the answers could add to the positive interdependence within the group or pair taking the test. The difference between requiring and not requiring group consensus on collaborative tests has not been systematically examined by researchers.

The importance of improving research methods in terms of the ways in which collaborative testing is studied in the face-to-face classroom has already been discussed in this review, but we also need to consider how this applies to the online classroom (e.g., including control groups, randomly assigning students to groups, considering the effects of group size along with the formation and stability of the group process skills (see Sandahl, 2009)). Students' prior experience working in collaborative groups should also be considered. The guidelines regarding collaborative tests should be made explicit, and students should know what is expected of them. Palloff and Pratt's (2007) *Stages of Collaboration* are one set of the guidelines instructors should use to help successfully implement collaborative testing in an online course. The Roseth et al. (2008) guidelines

on how a collaborative framework can be applied to a statistics classroom should also be used and modified to fit the online environment.

The effectiveness of using collaborative testing has mostly been measured using test scores and student surveys focusing on student attitudes, experience, perception, willingness to participate, anxiety, and evaluations of the collaborative testing process. These measures, along with others (such as Pozzi's (2007) four-dimensional framework), could be used together to more holistically evaluate the collaborative process. Using these different measures could provide researchers with important information regarding the use of collaboratively tests in online statistics courses.

Research on collaborative tests has shown that they have positive effects on students (e.g., in regard to increased test scores, decreased test anxiety, and positive attitudes towards the test method) (e.g., Breedlove et al., 2004; Giraud & Enders, 2000; Helmericks, 1993; Ioannou & Artino, 2010; Sandhal, 2009; Zimbardo et al., 2003). Considering the long standing idea held by many students about statistics courses being both difficult and unpleasant experiences (Garfield & Ben-Zvi, 2008), why not use collaborative tests in an introductory statistics course when the results of using collaborative learning and collaborative tests in statistics courses might make students' experiences a bit more positive?

Garfield and Franklin (2011) suggest that the purpose and use of assessment utilized in statistics course be that of: assessment *of* learning, assessment *for* learning and assessment *as* learning. Where:

Statistics teachers have traditionally used summative assessment to provide information *of* student learning, while using some types of formative assessments as agents *for* student learning, i.e., to provide

feedback to students to help them better learn statistics. The use of assessment *as learning*, which could encompass both summative and formative methods, situates the student at the integral junction between learning and assessment. In this unique purpose of assessment, students engage in new learning by monitoring and adapting their own understanding via the assessment process (Garfield & Franklin, 2011, p. 3).

Using collaborative testing in statistics courses would encompass all three of these aspects of assessment: *of learning* by providing information about student achievement, *for learning* by providing students with feedback on their learning, and *as learning* by helping students reflect on their own statistical knowledge and discuss it with other students while working on the collaborative test.

Integrating discussion and active learning in statistics classrooms has been shown to help students learn to think and reason about statistical concepts (Cobb & McClain, 2004; Garfield, 1995; Garfield, 2007), so why not apply this to statistics courses taught online and, more specifically, to the assessment used in these online courses?

Shaughnessy (2007) recommends that researchers in statistics education explore classroom discourse in statistics, and he asks, “Are we doing enough to identify and promote the types of teaching that will enhance our students discourse skills in statistics?” (p. 1001). Shaughnessy points out that skills such as analyzing, critiquing, communicating and representing have been identified as critical in models of statistical thinking (see Wild and Pfannkuch, 1999, in Shaughnessy, 2007) and statistical literacy (see Watson, 1997, in (Shaughnessy, 2007)). These critical skills can all be incorporated into collaborative exams. In addition, by implementing collaborative exams in the online environment, the discourse would be in the form of written communication between students, and this is something that has not yet been explored in the field of statistics

education. This could provide valuable information regarding students' thinking, and analyzing when it come to the learning of statistics. The online format offers new and exciting ways to explore teaching, assessment, and students' thinking, and these opportunities should be taken advantage of.

Implications for Research

Considering the increase that has taken place in online enrollment in recent years (Allen & Seaman, 2010), one can expect that more and more courses will be offered online and more and more students will elect to take these courses. Statistics courses would not be excluded in this trend. It is therefore important that more research takes place on using effective teaching methods and assessment in online statistics courses so that high quality courses grounded in empirical research on how to teach statistics and assess students' learning can be offered. Research on using collaborative testing in online statistics courses would likely help us better understand how we can structure and teach such quality courses.

This literature review has shown that collaborative testing in face-to-face classrooms can be beneficial for students' learning, but when it comes to online settings, little is known about using collaborative testing. The only reported online use of collaborative testing showed positive effects on students' learning, but in that study, students worked together designing the test but completed the test individually (Shen et al., 2006 & 2008). The effects of using collaborative tests in online courses, where students work together on a test, remain unknown. If we bear in mind that the use of collaborative methods in online courses has been encouraged and shown to be successful, one might expect positive effects on students' learning when using collaborative tests in

online classes. This literature review has raised many questions that still remain unanswered regarding the implementation of collaborative tests in the face-to-face classroom, such as which test format (consensus or nonconsensus) works better. One way of exploring how to implement collaborative tests successfully in an online course would be to explore the use of different test formats and their effects on student learning. The online format offers exciting ways to explore this. For example, students can easily be randomized into classes receiving different formats of collaborative tests, and students' participation and discussions in the course can be monitored through the course management system.

The goal of this study will be to explore the impact of using different formats of collaborative tests in an online statistics course on students learning and attitudes towards statistics, by answering these three research questions: (1) What is the impact of using collaborative tests in an online statistics course on students learning?; (2) What is the effect of using collaborative tests on students' attitudes towards statistics?; and (3) How does using a required consensus on collaborative tests vs. a nonconsensus approach affect group discussions? By attempting to answer these questions, we will ideally arrive at a better understanding of how to assess student learning in collaborative ways that are aligned with the ways in which students actually learn statistics. This will also help us gain more insight and information on how online statistics courses can be taught effectively using collaborative testing methods.

Chapter 3

Methods

The literature review in the previous chapter pointed out the lack of research on using collaborative tests in online introductory statistics courses. The aim of this study was to explore the impact of using collaborative tests in an online statistics course on students' learning. This chapter describes the procedure and data gathering that took place for the study, starting with a general overview of the study. The second section describes the subjects who participated in the study and the setting where data was collected. The third section outlines the instruments and the measures that were used, the reliability analyses of the instruments, their timeline and administration, and how variables were computed from the instruments. Finally, the fourth section describes the methodology that was used to analyze data in the study; it explains the procedures used to construct variables that were used to answer the research questions in this study.

Overview of the Study

This study took place in fall semester 2011. It included two online sections of a one-semester 3 credit, introductory statistics course (EPSY-3264 Basic & Applied Statistics) offered by the Department of Educational Psychology at the University of Minnesota. The same instructor taught both sections. Both sections were taught entirely online using version 8 of the course management system WebVista.

There were three main research questions:

1. What is the impact of using collaborative tests in an online statistics course on students' learning?

2. What is the effect of using collaborative tests on students' attitudes towards statistics?
3. How does using a required consensus on collaborative tests vs. a nonconsensus approach affect group discussions?

The study employed descriptive statistics, t-tests, and multiple regression to explore the impact of using collaborative tests in an online statistics course on students' learning. Examining students' test scores using multiple regression allowed for the first research question to be answered.

To explore the effects of both using collaborative tests on students' attitudes towards statistics and the effect of using a required consensus on collaborative tests vs. nonconsensus approach on group discussions, two different data analyses were used. The analyses consisted of descriptive statistics and t-tests, with the addition of a qualitative analysis that employed the aforementioned Pozzi et al. (2007) (see chapter 2) four-dimensional framework that measured participative, social, cognitive and teaching dimensions present in group discussion during the collaborative tests.

Subjects/Settings

Participants in the study consisted of 59 undergraduate students who were enrolled in one of the two sections of the online introductory statistics course (EPSY-3264 Basic & Applied Statistics), offered by the Department of Educational Psychology at the University of Minnesota. Of the students, 66% were female and 34% male. Students were undergraduate students enrolled in the course to complete the mathematical reasoning requirements for a Liberal Arts degree or a requirement for their particular major. Most of the students came from the College of Liberal Arts (47.5%),

College Continuing Education (23.7%) or College of Educational and Human Development (8.5%). There were also a few undergraduate students from Nursing (6.8%). Other students (13.5%) came from Dentistry, the College of Science & Engineering, Pharmacy and the College of Design and non-degree students. Students were randomly assigned to two different sections of the course, and each section received a different treatment. In the consensus section, students turned in one collaborative test per group, while in nonconsensus section, students discussed the tests with their group members but turned them in individually. Random assignment was used to protect against confounding. The instructor randomly assigned all students who were enrolled in one of two online sections of EPSY-3264 Basic & Applied Statistics to two different sections, using a random number generator.

Total enrollment in the two sections was 72 students, 36 in each section. The number of students enrolled after the deadline to drop from the course was $n=72$; however, six of those students did not participate at all, one withdrew later from the course, and six other students did not participate in all of the three collaborative tests. To be included in the data analysis, students had to participate in all three collaborative tests; therefore data from these 13 students was not included. The final analysis used data from $n=59$ students. Of the 13 students who were excluded from the analysis, four were in the consensus section and nine were enrolled in nonconsensus section. Of those four students who were not included in the data for the consensus section, two of them did not participate in all three collaborative tests, and two did not participate at all in the course. The consensus section had $n=32$ student and nonconsensus section had $n=27$ students.

The two course sections were offered entirely online except for one optional face-to-face introductory meeting that was held at the start of the semester. In this face-to-face meeting, the instructor went over the logistics of the course such as the syllabus, assignments, requirements, and how to navigate the WebVista course site. Slides from this meeting were made available to all enrolled students after the meeting.

The Course

The original, face-to-face version of the EPSY-3264 course was developed based on the Guidelines for the Assessment and Instruction in Statistics Education (GAISE, 2005). This course was designed to develop students' statistical literacy and their statistical thinking. It included collaborative learning activities such as discussion assignments and collaborative quizzes. Real data was used and students' used technology where the focus was on conceptual understanding instead of learning a set of procedures. The course followed the Adapting and Implementing Innovative Material in Statistics (AIMS) curriculum that was designed through a NSF-funded project. That project developed lesson plans and activities based on innovative materials for introductory statistics courses aligned with the GAISE for teaching introductory statistics courses (AIMS Project, NA; Garfield, & Ben-Zvi, 2008 & Garfield, delMas, & Zieffler, 2008).

The EPSY-3264 online course was modified from the face-to-face version of the course to be offered first online in the year 2004. Most of the activities and assessments (labs, midterm, collaborative tests and final) used in the face-to-face course were adapted to the online environment and lecture notes were created. Assessments traditionally used in the face-to-face course came from the AIMS curriculum. Many of the items used for the midterm and on the collaborative tests came from the Assessment Resource Tools for

Improving Statistical Thinking (ARTIST) online item database. The ARTIST online item database was created through a NSF-funded project where the goal was to design a variety of online assessment resources aimed at improving statistical thinking in the teaching of a first courses in statistics (Web Artists, 2006).

Pedagogical Model

The pedagogical model that was used in this study is influenced by the epistemological assumption of constructivism: the idea that the learner constructs knowledge; that he or she is not merely a recipient of information. Additionally, in social constructivism, the construction of knowledge is believed to take place among learners. It is the shared experience rather than individual experience of the learner (Eggen & Kauchak, 2006). According to this theory, students are seen as bringing their own ideas, experiences and beliefs to the classroom, effecting how they understand and learn new material. Students do not receive material presented in the classroom without questions; they receive and restructure it, so it fits into their own cognitive framework. They are seen as active participants in constructing their own knowledge, rather than simply receiving the knowledge that was transmitted to them (Garfield, 1993).

Furthermore, the pedagogical model used resides on the situated cognition learning theory influenced by social constructivism. It suggests that learning is socially constructed, depended on and cannot be separated from the context in which it takes place (Eggen & Kauchak, 2006). This idea emphasizes the importance of creating a learning environment where a community of practice can form. “Communities of practice are groups of people who share a concern or a passion for something they do and learn how to do it better as they interact regularly” (Wenger, 2006, p. 1). Communities of

practice (CoP) take place where members of the community are joined by their common purpose and take part in shared activities, when the common purpose is learning they are known as learning communities. The goal of the CoP is to help develop the abilities and skills of its members and, in a meaningful way, construct, and share knowledge in a relevant context in a supportive learning environment (Dabbagh & Bannan-Ritland, 2005).

Learning Environment

The online course was taught using an asynchronous format. It was offered completely online using the course management system (CMS), WebVista, version 8. It allows students to access learning materials and participate in learning activities through the Internet. All assignments, assessment and communication in the course were administrated through WebVista. A learners-centered learning environment was created in WebVista, where learners were responsible for their own learning by creating and organizing information available to them in the CMS (Dabbagh & Bannan-Ritland, 2005). This learning environment was aligned with the before mentioned constructivist paradigm, where

the learning environment is largely learners centered, providing multiple opportunities for the student to synthesize, organize, and restructure information, and to create and contribute resources to the virtual space of the course. Students can select and sequence educational activities as well as create their own learning opportunities to satisfy their learning needs. (Hooper & Hannafin, 1991, as cited in Dabbagh & Bannan-Ritland, 2005, p. 39)

Self-directed learning is cherished and supported in this learning environment and self-directed learning skills are seen as crucial for supporting the learner to take responsibility for his or her own learning process. These aforementioned instructional

characteristics of the learning environment along with others (e.g., assigning students to groups with shared goals and problems to work on, distributing the control of learning among students and not just the instructor, including flexible and negotiated learning activities, encouraging and supporting innovation and creativity in group assignments) were believed to promote the formation of CoPs in the courses (Dabbagh & Bannan-Ritland, 2005).

The Experimental Conditions

There were two sections of the EPSY-3264 online course offered in fall 2011. Two different conditions were designed, each assigned to one version of the course. The same instructor taught both sections of the course. The instructor had taught the course 12 times in the past 2.5 years.

The only difference in the two sections, the treatment, was the type of collaborative assessment used. The two sections of the course were constructed to be exactly the same in all other aspects: assigned readings, topics covered, assignments (discussion assignments, labs, activities, article and graph critiques, midterm and final) (see Appendix B). The only difference between the two sections was in the format of the collaborative tests administered.

In the consensus section, student groups were required to work together on the tests and to submit one group test, to be graded. In the nonconsensus section, students worked collaboratively in groups on tests but then turned in their own, individual test to be graded. The consensus section included 6 groups, and 5 groups were in the nonconsensus section. Group size varied from 4 to 6 students. Students were randomly assigned to groups within the course to protect against confounding.

During the optional face-to-face meeting, students were told that there would be two different formats of collaborative tests used between the two sections and the formats were briefly described. Then two weeks before each collaborative test, an announcement was sent out explaining the procedure for each test. Students were graded on the collaborative tests based on correctness and participation. In both sections, students were required to post their answers to the test in their discussion forum. In addition to the first post, each member was to provide at least two meaningful answers or comments in his or her discussion group. A meaningful comment was considered a comment related to the concept under study. Each post was worth 33.33% of the individual's grade on the test. For example, if the score on the individual test was 15 points and a student only contributed two times to the discussion, he received at most a score of 10 points (66.67% of 15). The three posts (the initial post and two comments/questions/answers) were the minimum, and the initial post with individual answers was required to receive a full grade. In the nonconsensus section, students were to submit their individual versions of the test for grading. For the consensus section, one student was to summarize the responses for each group and turn them in for grading. Only one answer per question was allowed and groups had to reach a consensus on answers.

Instruments and Measurements

Five different assessment instruments (see Appendix A) were used to gather data to explore the impact of using different collaborative tests in an online statistics course on students' learning. The instruments used in the study were either required (tests, midterm exam, final exam), or optional surveys measuring students' attitudes towards statistics and perspective towards taking collaborative tests. Students received extra credit for

completing the optional surveys; the required instruments were all part of the assessment used in the course.

Required Assessments

Students were required to take three different types of exams (Midterm exam, Final exam and collaborative tests) as part of the assessment in the course. All of these exams were used to measure students' knowledge in statistics at different times in the course. The first exam was the pretest, which was the Comprehensive Assessment of Important Outcomes in Statistics (CAOS), and this same exam was used as the final exam. Two collaborative tests were administered before the Midterm exam and then one collaborative test was administered after that. Each instrument is described more below.

Comprehensive Assessment of Important Outcomes in Statistics (CAOS). The CAOS test was used as a pre- and posttest to measure students' prior knowledge in statistics and students' learning at the end of the course. The CAOS test was designed to measure students' statistical reasoning after completing a first course in statistics. The test focuses on statistical literacy and conceptual understanding (delMas, Garfield, Ooms, & Chance, 2007). The CAOS test includes 40 multiple-choice questions. Only 33 items were used on the pre and post test; the other 7 were used on the Midterm exam. The CAOS posttest was used as the final exam in the course (see Appendix). Students received 10 points for completing the pretests regardless of their score; students' score on the final exam was computed based on number of correct items.

Midterm exam. The Midterm exam was made up of 7 items from the CAOS test (items that were not included on the pre- and final tests), 14 items from the ARTIST

online item database, and 8 items previously used in the course. The Midterm exam was worth 35 points it included 29 items (19 multiple-choice and 10 open-ended items).

Table 2

Midterm Exam Item Numbers Based on Format and Source

<i>Item Format</i>	<i>Item Source</i>		
	CAOS	ARTIST online item database	Old EPSY-3264 items
MC*	8,9,10,11,12,14,15,	1,2,4,5,18,19,20,21,22,23,24	16
OE**		3,13, 26	6,7, 17,25,26,27,29

*MC= Multiple-choice, OE**= Open-ended.

The 8 other items come from the AIMS curriculum that had been used frequently for assessment items in the EPSY-3264 courses. All the items were used at one time in the face-to-face version of the course and they have all been used in the online course since it was offered. Seven of these items were open-ended. A detailed rubric for these items exists and the instructor used that for grading the Midterm exam. The Midterm exam covered material from the first 8 weeks of the course.

Collaborative tests. The instructor used selected items from the ARTIST online item database to develop the three collaborative tests. The collaborative tests accounted for 20% of the final grade in the course and each test was worth 20 points out of 290 available points. Each test consisted of 15 items representing three different difficulty levels, weighing from 1 to 2 points each. The three tests varied in terms of topic covered. They all include similar frequency of both multiple-choice and open-ended questions and

levels of difficulty defined by items measuring either statistical literacy, statistical reasoning or statistical thinking (see Table 3).

Table 3

Frequency of Item Types and Level of Difficulty on the Collaborative Tests

<i>Test</i>	<i>Item Format</i>	<i>Number of Items Selected to Assess Statistical:</i>		
		<i>Literacy</i>	<i>Reasoning</i>	<i>Thinking</i>
1	MC*	4	3	
	OE**	2	3	3
2	MC	3	6	
	OE	1	4	1
3	MC	2	7	
	OE	2	1	3

*MC= Multiple-choice, OE**= Open-ended.

Items that measure statistical literacy, reasoning and thinking were used to reflect difficulty of items on the collaborative tests. The three concepts have been defined as:

- Statistical literacy: “understanding and using the basic language and tools of statistics” (Garfield & Franklin, 2011, p. 4).
- Statistical reasoning: “reasoning with statistical ideas and make sense of statistical information” (Garfield & Franklin, 2011, pp. 4-5).
- Statistical thinking: “recognizing the importance of examining and trying to explain variability and knowing where the data came from, as well as

connecting data analysis to the larger context of a statistical investigation”
(Garfield & Franklin, 2011, p. 5).

Items measuring these concepts are ranked in hierarchical order in terms of difficulty, starting with statistical literacy items as less difficult and ending with items measuring statistical thinking as most difficult.

The instructor designed a rubric and coded the items based on which statistical learning outcome (literacy, reasoning or thinking) they were measuring. The rubric and the item coding were reviewed by three faculty members, all of whom are experts in the field of statistics education. The rubric was used for grading by the instructor and the two teaching assistants who helped with the course. The teaching assistants graded one section each. The instructor and the teaching assistant graded each collaborative test separately. The instructor then compared the two graded tests and when inconsistency was present, she discussed them with the teaching assistants to reach a consensus regarding the grading.

Instruments Offered as Extra Credit

Students were offered the opportunity of taking three surveys for extra credit. The pre- and post-Survey Of Attitudes Toward Statistics (SATS-36) were used to measure changes in students’ attitudes towards statistics over the semester, and the Students Perception on Collaborative Tests (SPCT) instrument was used to explore students’ perspective towards taking the collaborative tests over the semester. Five extra credit points were offered for each instrument completed. Each instrument is described more below.

The Survey of Attitudes Toward Statistics (SATS-36). The SATS-36 pre- and post instruments to measure students' attitude towards statistics were implemented. The SATS-36 measures the six attitudes components: Affect, Cognitive component, Value, Difficulty, Interest and Effort towards statistics. Scores from the SATS-36 have been carefully validated on postsecondary students taking statistics with a wide variety of characteristics in a large number of institutions both within and outside of the US (Schau, 2005). The instruments used included 53 items for the presurvey and 46 items for the postsurvey; 36 statements on both instruments included a 7-point response scale (strongly disagree, disagree, somewhat disagree, neither agree nor disagree, somewhat agree, agree, strongly agree). For the pre- and postsurveys, five statements were modified to better reflect an online course. See Table 4. Other items on the SATS included academic and demographic items.

Table 4

Original Statements and Statements Used on Pre- and Post-SATS-36 Instruments

<i>Original statements</i>	<i>Statements used</i>
I will get frustrated going over statistics tests in class (presurvey)	I will get frustrated going over statistics tests in this course.
I plan to attend every statistics class session (presurvey)	I plan to log into the course website two times a week
I get frustrated going over statistics tests in class (postsurvey)	I get frustrated going over statistics tests in this course
I am under stress during statistics class (postsurvey)	I am under stress when I am logged into the course
I tried to attend every statistics class session (postsurvey)	I tried to log into the course website two times a week

On the pretest version of the instrument, two academic items were changed to better-fit students taking the course. For the question “What is your major?” the answer option Nursing was added, Statistics and Mathematics were combined into one answer option, and the option Medicine/Premedicine was deleted. For the question “Degree you are currently seeking;” the following answer options were deleted: Associate, Certification, Postbachelor's Licensure and Specialist. Two new questions were added to the academic part (see Table 5).

Table 5

Questions Added to the Pre-SATS-36

<i>Items</i>	<i>Answer options</i>
Have you been enrolled in an online course before?	Yes No
Number of online courses completed:	_____

On the posttest version of the instrument, the academic question, “In a usual week, how many hours did you spend outside of class studying statistics?, ” was changed to: “In a usual week, how many hours did you spend studying statistics?”

Negatively worded items on the SATS-36 were reversed before the reliability and the data analyses. A score of 1 became 7, 2 became 6, and 3 became 5. The six aforementioned components were created from 36 items on the pre- and post-SATS-36 instrument. Each component had the same possible range of scores between 1 and 7, which corresponded with the 7-point response scale on the instrument. A higher score reflected a more positive attitude on each component.

Students' Perception on Collaborative Tests (SPCT). The SPCT instrument was used for students to evaluate their own learning and test taking styles, test anxiety, preparation, perceptions of freeloading, fairness of grading, and retention of information in regards to their experience taking the collaborative tests. The instructor developed the survey based on a literature review on collaborative testing. The survey included 20 items; 18 statement items on a 4 point-likert scale (*strongly disagree* to *strongly agree*) and 2 open-ended items. Between the two course sections, two statements were modified based on how the collaborative tests were administrated in each section. See Table 6.

Table 6

Different Statements on the SPCT between the Two Course Sections

<i>Consensus section</i>	<i>Nonconsensus section</i>
I would have preferred to take individual tests.	I would have preferred to take only individual tests.
I would have preferred being able to discuss group tests with my group but submit my own individual answers.	I would have preferred being able to discuss group tests with my group and submit one set of answers as a group.

Item 15 was negatively worded so those responses were reversed before the reliability and the data analysis. A score of 1 became 4, and 2 became 3.

Average scale scores were created from the SPCT instrument to be used in the data analysis. Sum scores were computed by adding up scores for each answer option selected. For example, items on the SPCT instrument had a 4-point Likert-type answer scale with the options *strongly disagree*, *disagree*, *agree* and *strongly agree*. The *strongly agree* answer option equals 4 points while the *strongly disagree* option is equal to 1

point. The average scale scores were computed by dividing the sum scores by the number of items used. For example, for the SPCT score 18 items were used so in that case, sum scores were divided by 18 resulting in an average score that corresponds to the 4-point Likert-type scale that was used for the 18 items on the SPCT instrument. A higher score reflects a more positive perception towards collaborative tests.

Reliability Analysis of the Research Instruments

A reliability analyses was used to estimate the internal consistency of each instrument. Coefficient alpha (Cronbach, 1951) was used as a measure of reliability. Coefficient alpha for the scores on CAOS, SATS-36, the Midterm and both versions of SPCT are reported in Table 7.

Table 7

Coefficient Alpha for Sample Scores and Responses

<i>Assessment</i>	<i>Weeks 1-2</i>	<i>Week 8</i>	<i>Weeks 13-14</i>	<i>Week 15</i>
1. CAOS	.442 (57)	--	--	.713 (59)
2. SATS-36	.930 (47)	--	.914 (47)	--
3. Midterm	--	.770 (59)	--	--
4. SPCT, Consensus section	--	--	.871 (28)	--
5. SPCT, Nonconsensus section			.870 (25)	

Note. Sample size is reported in parentheses.

These coefficient alphas indicate a satisfactory level of internal consistency between items on the instruments. The Pre-CAOS which had the lowest coefficient alpha .442. Students were required to take the Pre-CAOS at the beginning of the course.

Regardless of their score on it they received 10 points, which might have effected how they took the instrument, for example time spent and amount of guessing. However, because the psychometrics of the CAOS have been well established in the literature (delMas, Garfield, Ooms, & Chance, 2007), the Pre-CAOS will be used here despite its low coefficient alpha.

Timeline for Instruments and Test Administration

The first instruments, the CAOS test that was used as a pretest and the SATS-36 presurvey, were administered in the first 2 weeks of class. Students had 12 days to complete both instruments. Students had 5 days to work on the three collaborative tests and 8 days to complete the Midterm and the Final exam. The SATS-36 postsurvey and the SPCT evaluation were available for 14 days for students to complete. The CAOS pretest, the three collaborative tests the Midterm and the final were administrated through WebVista on the course website. The other instruments, SATS-36 pre- and post- and the SPCT were sent out using an online survey design program called *Qualtrics* (<http://www.qualtrics.com/>). Students were sent an email invitation with a link to each survey. A thank you email was sent to all those that complete a survey. See Appendix C for this correspondence to students. Table 8 shows the timeline for when instruments were administrated in the course.

Table 8

Weeks & Instruments Administered

<i>Weeks 1-2</i>	<i>Week 5</i>	<i>Week 8</i>	<i>Week 9</i>	<i>Week 12</i>	<i>Weeks 13-14</i>	<i>Week 15</i>
CAOS pretest	Collaborative test #1	Collaborative test #2	Midterm exam	Collaborative test #3	SPCT evaluation	Final exam (CAOS)
SATS-26 presurvey					SATS-36 postsurvey	

Analysis of Data

The data analysis and results are divided into two parts: a qualitative and a quantitative part. For the qualitative analysis, a specific framework (Pozzi et al., 2007) was used to evaluate and monitor computer-supported collaborative learning that took place in the collaborative tests in the course. The quantitative analyses include the use of descriptive statistics; some of them generated from the qualitative framework as well as multiple regression.

Qualitative Analysis

Qualitative data were collected and analyzed to explore the nature or the quality of the discussion while taking the collaborative tests. A qualitative analysis (see framework below) was used to gather information regarding the collaborative tests. This was done in an effort to help answer the research questions and to give insight into the nature of the discussions taking place. Qualitative variables were converted to quantitative variables that were used as exploratory data and in the regression analysis. Below is a description of the framework and how it was used to create the quantitative variables that were used for the quantitative analysis.

Qualitative Data Analysis Framework

To evaluate the collaborative learning process in the online course a framework proposed by Pozzi et al. (2007) was used. The framework was used in this study because it was designed to evaluate and monitor computer-supported collaborative learning (CSCL) processes. The framework consisted of interaction analysis techniques and content analysis of messages posted by students in the online environment. The framework was a four-dimensional approach that included participative, social, cognitive and teaching dimensions that take place in a learning community. Indicators consisting of both qualitative and quantitative variables have been identified to express each of these four dimensions (Pozzi et al., 2007). In the Pozzi et al. (2007) framework, the focus can be on some or all of the dimensions at the same time depending on the research question under study. The indicators in that sense were not seen as stable; they may in fact vary in weight depending on the context and goals of the analysis. For example, when exploring the collaborative activity in a course, more focus would be on indicators related to the participation and the social dimensions (Manca, Persico, Pozzi, & Sarti, NA). Below, each dimension will be described.

The participative dimension. This dimension was an important part of the monitoring process. It is a good indicator of students' involvement in the course, and it provides information about who is participating and how much (Pozzi et al., 2007; Persico, Pozzi & Sarti, 2010). Quantitative data in the form of frequency of posts for each test was gathered for this dimension.

The social dimension. This dimension measured the social presence of students in the course. To what extent participants were able to be and be perceived by others as

“real” people in the medium being used. Which in this case was the online course format (Pozzi et al., 2007; Persico et al., 2010).

The cognitive dimension. This dimension is defined as “the extent to which learners are able to construct and confirm meaning through sustained reflection and discourse in a critical community of inquiry” (Garrison et al., 2001 in Persico et al., 2010, p. 9). The cognitive dimension involved both individual and group knowledge building. First, by a personal explanation of content and expression of individual points, which lead to a collaborative discussion and negotiation where collective meaning and interpretations of reality were constructed (Manca, Persico, Pozzi, & Sarti, NA).

The teaching dimension. The definition for teaching presence is “the design, facilitation, and direction of cognitive and social processes for the purpose of realizing personally meaningful and educationally worthwhile learning outcomes” (Anderson et al., 2001 in Pozzi et al., 2007, p. 174). It is seen as the binding element of building a learning community. Messages that provide guidance and instruction, facilitate discourse, and manage organizational matters are seen as including teaching presence. Teaching presence does not only relate to instructors, it may also apply for students, for example, through group leadership (Persico et al., 2010; Pozzi et al., 2007).

How dimension variables where created. In this study, the focus was on the three dimensions, cognitive, teaching and social. Indicators from these dimensions were used since they were believed to be relevant to the research questions put forth. Because students were required to participate in discussion during the collaborative tests, only one indicator was used for the participative dimension. The other three dimensions were explored more to create quantitative variables that were used in the quantitative analysis.

The unit of analysis was each post or message during the collaborative test. A maximum of three indicators could be identified for each message from all three dimensions.

The instructor coded all the messages for the three collaborative tests for the two sections, which was a total of 33 tests including 753 messages. One of the teaching assistants helped validate the process by coding messages. The instructor met with the teaching assistant and explained the coding process, and together they went through two collaborative tests. Then, the teaching assistant coded three tests for a total of 46 messages individually. When the coding from the teaching assistant and the instructor were compared, it showed 77.1% agreement. The instructor and teaching assistant discussed their discrepancies and came to agreement with the coding.

Quantitative variables were created for the three dimensions—cognitive, teaching, and social—and their 10 indicators. Proportions out of total frequency of indicators for each student were computed, a total of 13 variables, 1 variable for each dimension and 10 variables for each indicator. Each indicator weighted the same. Table 9 includes the four dimensions, the indicators and message examples that were used in this study. The number of indicator per message varied from 0 to 3 indicators, when more than 3 indicators could be identified; the more apparent 3 indicators were selected. Each post could have up to three indicators from all the three dimensions, cognitive, teaching and social. The examples in Table 9 show the appearance of a strong indicator for each dimension. However, in some cases indicators are intertwined with other indicators from the same or another dimension; for example, the post that shows the *Exploration* indicator for the cognitive dimension ends with “Your comments, please – this is a 2 pointer and we want to get it right.” This is a *Cohesion* indicator for the social dimension.

The *Metareflection* post for the cognitive dimension also includes the *facilitating discourse* indicator for the teaching dimension when it ends with “Can you offer any insight into your answer. I am not very confident in my choice. How did you decide on C? thanks”

Table 9

Indicators for All Dimensions and Examples from the Collaborative Tests (Inspired by Persico et al., 2010)

<i>Dimension</i>	<i>Indicator</i>	<i>Data</i>	<i>Example</i>
Participative	Active participation	Frequency of post per student for each collaborative test	Count of posts for each discussion from the collaborative tests
Social	Affection	Expression of emotions, intimacy and personal anecdotes	"... I didn't see anything that stuck out as far as changing goes ☺ Good job!"
	Cohesion	Vocatives, references to the group, and salutations	"Thanks Mary!"
Cognitive	Revelation	Recognizing a problem, demonstrating a sense of puzzlement and explaining a point of view	".. I too am confused about #8. I think the correct answer is actually B, I originally thought C. But, the correlation of +.8 is just as strong as the correlation of -.8."
	Exploration	Expressing agreement or disagreement, sharing ideas and information, brainstorming and negotiating	"#15-The median is the equal point. The mean is the equal area point. The median and mean are very close together. Half of the fish – 39 – are between the median size of 25.295 in. and the 33.4 in., about 2 std dev. I am beginning to think that Stephanie and Emily are correct. Your comments, please – this is a 2 pointer and we want to get it right."
	Integration	Linking ideas together, making synthesis and constructing solutions	"It's A. Each time you flip a coin, you have 50/50 chance for heads or tails, so the mean of all coin flips is 50, and the shape of all coin flips is normal. The more times you flip a coin, the greater the "sample" of coin flips. Don't forget the central Limit Theorem which states that the larger the sample size, the closer to the population mean the mean of the sample will be." (cont.)

<i>Dimension</i>	<i>Indicator</i>	<i>Data</i>	<i>Example</i>
<i>(Table 9, cont.)</i>			
	Resolution	Connecting to real-life applications and testing solutions.	“I agree that this survey is biased towards those who watch CNN or go to CNN.com, but this survey is definitely directed towards those who have internet access. There are quite a few Americans who cannot afford internet access or computers, and therefore would not be included in this study. Also, older citizens in the American population may not use the internet solely on the fact that they just don’t like newer technology. These people would be excluded as well”
	Metareflection	Evaluating own knowledge, skills, limits, cognitive processes and planning, monitoring, or adjusting own cognitive processes	“... Thank you for contribution your answers. On question 4, I was torn between A or C. I ended up going with A, manly because the shape mimicked the original population. C also mimicked that shape but I ultimately went with A because the spread seemed to the smaller than in C. My line of thought is that the spread in the sample population would $6.404/2=3.202$. Can you offer any insight into your answer. I am not very confident in my choice. How did you decide on C? thanks”
Teaching	Direct instruction	Recommending activities, pointing out misconceptions, providing feedback and assessment that confirm understanding	“... I think you might have the definition of parameter and statistics a bit off in your example of height for Q2. Your parameter would be the average height of all buildings in Minnesota, whereas your statistics would be the average height of a sample of 10 buildings in Minnesota.”
	Facilitating discourse	Identifying areas of agreements/ disagreement to achieve consensus, encouraging, acknowledging or reinforcing participants contribution, setting the climate for learning	“I like your answer to #12. That is an interesting example. Also, nice lurking variable. That could definitely throw things off.”

(cont.)

<i>Dimension</i>	<i>Indicator</i>	<i>Data</i>	<i>Example</i>
<i>(Table 9, cont.)</i>			
	Organizational matters	Introducing topics, providing explanations for methods and letting students know of deadlines	“I just had a quick question about who is going to be the group leader this week, just so we don’t scramble like last week last minute to get someone to cover for the group lead position? Andrew and I (Mary) both had a shot at being group lead; does anybody else want to volunteer for this week’s Group Test assignment?”

Information gathered from the framework described above was used to help answer the third research question (3) How does using a required consensus on collaborative tests vs. a nonconsensus approach affect group discussions? The quantitative variables that were created were used as descriptive statistics and in the regression model to answer the first research question.

Quantitative Analyses

Descriptive statistics and t-tests were used to answer the second research question: What is the effect of using collaborative tests on students' attitudes towards statistics?

Multiple regression. The multiple regression model was used because the dependent variable was continuous, and more than one independent variable was used. Multiple regression can be used to explore the relationship between one dependent variable and a collection of independent variables. By using multiple regression, the effects the independent variables have on the dependent variable can be investigated. Relationships between two variables can also be analyzed while controlling for other variables in the model (Agresti & Finlay, 1997). Multiple regression uses interval variables; categorical variables can also be used but they require dummy coding if they include more than two categories (Lewis-Beck, 1980).

The multiple regression model: $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$
where k is the number of predictor, $\beta_0 \dots \beta_k$ are parameters, β_0 is the intercept, β_k is the slope. Y is the dependent variables, $X_1 \dots X_k$ are the independent variables and ε_{ij} is the

error term (Lewis-Beck, 1980). In this study, the two multiple regression models will be used:

Model 1: Midterm exam= Section +
CLT1&2Discussion1+CLT1&2Discussion2+ SPCT+Pretest

Model 2: Final exam= Section + CLTDiscussion1+CLTDiscussion2+
SPCT+Midterm exam score

Section: 0= Nonconsensus, 1=Consensus

CLT1&2Discussion1= Quality of discussion on tests #1 and 2, one
dimension (Social, Cognitive or Teaching)

CLT1&2Discussion2= Quality of discussion on tests #1 and 2, one
dimension (Social, Cognitive or Teaching)

CLTDiscussion1= Quality of discussion on all three tests, one dimension
(Social, Cognitive or Teaching)

CLTDiscussion2= Quality of discussion on all three tests, one dimension
(Social, Cognitive or Teaching)

Pretest= Students score on the pretest CAOS

SPCT=Students Perception on Collaborative Tests scale score

The first model used the Midterm scores as the dependent variable; the second model used the Final exam as the dependent variable. Both the Midterm exam and the Pretest variables were seen as measures of students' prior knowledge but at different times, with Pretest measuring students' prior knowledge at the start of the semester and the Midterm exam measuring students' knowledge at week 8 of the semester. The Midterm exam was used for the second model as an independent variable instead of the Pretest variable, because it was believed to be a better measure of prior knowledge for the Final exam.

Because the discussion (CLT1&2Discussion and CLTDiscussion) variables (Social, Teaching and Cognitive) are proportionally related to each other, entering all three of them into the regression model would not make sense, as they could be strongly correlated. Instead, it will be enough to include only two of them in the model. Hierarchical multiple regression and correlation between the discussion variables and the dependent variables will be explored to see which discussion variables will be used for the two models. Predictor variables will be entered into the model in three blocks: first the three covariates, Section, Pretest and SPCT, then the CLT1&2Discussion1 variables and then the second CLT1&2Discussion2. By using hierarchical multiple regression the variance each discussion variables accounts for in the dependent variable was investigated. It provided information on how much new variance in the dependent variable can be explained by adding each CLT1&2Discussion variable to the model (Field, 2005).

Assumptions of multiple regression. In an effort to make accurate inference about the actual population values of the parameters, the regression model needs to meet the following assumptions.

Linearity. The relationship between the dependent variable and the independent variables should be linear because the regression model used here assumes a linear effect. To determine if this assumption is violated a plot of standardized residuals against the standardized predicted values was created (Agresti & Finlay, 1997). The graph should not show any sort of curve, the dots should be randomly dispersed around 0 in order for the linearity assumption to hold (Field, 2005).

Normality. Residuals in the model are assumed random and normally distributed with a mean of 0. To check if this assumption is violated residuals are plotted around their mean value 0. This assumption is met if the histogram shows an approximately bell shape about 0 (Agresti & Finlay, 1997). A normal probability plot can also be used here, if the plot shows a straight line it corresponds to a normal distribution and the assumption is met (Field, 2005).

Independence. This assumption is also called autocorrelation is assumes that the residual for one observation is not correlated with a residual for another observation (Lewis-Beck, 1980). To see if this assumption is met, a Durbin-Watson test for serial correlations between residuals was conducted. This test checks to see if adjacent residuals are correlated. The Durbin-Watson test statistics varies between 0 and 4, the residuals are uncorrelated at a value of 2. A negative correlation is indicated for values higher than 2 and a positive correlation for values lower than 2 (Field, 2005).

Homoscedasticity. The assumption of homoscedasticity is met if the variance of the residual for the predictor variable is constant at each level for the predictor variable. If the variance is very unequal it is called *heteroscedasticity*. To check if the assumption of homoscedasticity is met a plot, just like the one created for linearity assumption of standardized residuals against the standardized predicted values was used. The dots on the graph should be randomly dispersed around 0, and the dots should not funnel out since that would be an indicator for heteroscedasticity (Field, 2005).

Multicollinearity. The predictor variables in the regression model should not be perfectly or highly correlated with each other. To see if this assumption is met, a

correlation matrix for all predictors variables was created and any correlation higher than .8 violates this assumption. In addition, the variance inflation factor (VIF) was also used. The VIF checks if a strong linear relationship is present between any of the predictor variables. An average VIF value greater than 1 violates the assumption of multicollinearity (Field, 2005).

Interaction effect. When the impact of one independent variable depends on the value of another independent variable an interaction effect is present (Lewis-Beck, 1980). To see if an interaction effect between the predictor variables was present, an interaction model was tested. Before that, the predictor variables were mean-centered to protect against multicollinearity between the independent variables and the interaction terms (Howell, 2007). The interaction model for the two regression models included more than one interaction term because more than two independent variables were presented. In this case, the interaction terms were cross-products for each pair of the independent variables in the model. Because hierarchical regression was used, the final regression model might include fewer than four independent variables, which would then make the interaction model smaller. However, if the final model included all four independent variables, this is how the interaction model for the two regression models would look:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_1 X_2 + \beta_6 X_1 X_3 + \beta_7 X_1 X_4 + \beta_8 X_2 X_3 + \beta_9 X_2 X_4 + \beta_{10} X_3 X_4$$

If the interaction terms are not significant, they should be dropped from the model (Agresti & Finlay, 1997). Before the interaction term is created, the dependent variables will be mean-centered to reduce multicollinearity.

This chapter presented the research methods that were used in the study. It described the setting the study took place in and its subjects. Instruments and measurements that were used were described along with their timeline of administration. It covered the data analysis, both the qualitative and the quantitative methods that were used in the study. Next chapter will describe the results of these analyses.

Chapter 4

Results

The previous chapter described the methods used for gathering and analyzing the data used in this study. This chapter presents the results of the study and the analysis of these results. Six major sections are included in this chapter. The first section presents a descriptive analysis of the students who were enrolled in the two sections of EPSY 3264. The second section includes reliability and scale scores for the subscales that were constructed and used to answer the research questions. The next three sections of this chapter are devoted to each of the three research questions. A multiple regression analysis was used to answer the first research question, descriptive analyses consisting of sample means, standard deviations, correlations, scale scores, and t-tests were used to answer the second and third research questions. Finally, additional analysis not addressed with the three research questions is provided.

Examining the Two Online Sections

Students were randomly assigned to different sections of the course. The random assignment was possible because the course was offered online. It would have been difficult to randomly assign students to different sections in a face-to-face course due to time availability for students.

Personal information regarding the students who were enrolled was limited. The available information from students' enrollment was gender, academic level (i.e., freshman, sophomore, junior, senior) and primary program of study. In addition, two questions on the SATS-36 asked students if they had been enrolled in an online course before and how many online courses they had completed.

Chi-square analyses were run to check if students enrolled in the two sections were similar in regard to the distributions of females and males and academic level. Table 10 shows the proportions of academic levels between the two sections. When there was low count in some cells due to the small sample size, a Monte Carlo test was used to simulate a p-value using 10,000 replicates for the analysis (Field, 2005). Monte Carlo tests were used for the analysis of academic level and number of online course students had completed. There was no difference in students' academic levels between the two sections ($\chi^2 = 4.683$, $p = 0.338$). The analysis also showed that the number of female and male students in the two sections was similar ($\chi^2 = 1.04$, $df = 1$, $p = 0.308$). This indicates that the random assignment to the two sections was successful and that students in both sections were similar in terms of the demographic variables.

Table 10

Proportions of Students' Academic Levels between Sections

	<i>Section</i>	
	Consensus (N = 32)	Nonconsensus (N = 27)
Sophomore	.188	.222
Junior	.375	.185
Senior	.251	.481
Nondegree	.125	.074
Graduate Student	.063	.037

The two questions that were added to the pre-SATS-36 regarding students experience taking online courses revealed that more students in the nonconsensus section had been enrolled in online courses before taking this course. Table 11 shows the proportion of students that had been enrolled in an online course prior to taking this course for each section. The difference between the sections was statistically significant ($\chi^2 = 5.562$, $df=1$, $p = 0.018$). To see if there was a difference in the number of online courses students had completed between sections the nonparametric Mann-Whitney U test was used because the sample size were small and the distribution for the number of online courses students had completed was not normally distributed. There was a statistically significant difference between the number of online courses students had completed between two sections ($U = 110$, $P = 0.000$). The mean for the number of

online courses students had completed before this course was higher for students in the nonconsensus section (3.96) compared to 1.08 for students in the consensus section.

These two questions were based on a sample of 47 students, or those students who took both the pre- and postversion of the SATS-36. In the consensus section, 24 out of 32 students completed the survey and for the nonconsensus section, 23 out of 27 students completed it.

Table 11

Proportion and Frequency of Students that Have Been Enrolled in Online Courses Before

<i>Have you been enrolled in an online course before?</i>	<i>Section</i>	
	<i>Consensus</i>	<i>Nonconsensus</i>
Yes	12 (.50)	19 (.826)
No	12 (.50)	4 (.173)

Reliability and Scale Scores

An additional reliability analysis was conducted for the 13 different scale scores that were produced for the Students Perception on Collaborative Tests (SPCT) and on the pre- and postversions of the Survey Of Attitudes Toward Statistics (SATS-36) instruments. These scale scores were used to help answer the three research questions.

Table 12 includes reliability coefficients and a descriptive summary for the average scale scores that were produced.

Table 12.

Reliability and Descriptive Summary for the Average Scale Scores

	<i>Scale score for Scales</i>	<i>Coefficient Alpha</i>	<i>N</i>	<i>Items Used</i>	<i>Average Score</i>	<i>Min Score</i>	<i>Max Score</i>
	SPCT	.867	54	All 18 items, 15*	2.89	1.83	3.89
Pre	Affect	.853	46	3, 4*, 15*, 18*, 19, 28*	3.99	1.83	7.00
	Cognitive	.807	47	5*, 11*, 26*, 31, 32, 35*	4.98	3.5	7.00
	Value	.875	47	7*, 9, 10, 13*, 16*, 17, 21*, 25*, 33*	5.47	3.78	7.00
	Difficulty	.823	45	6, 8*, 22, 24*, 30*, 34*, 36*	3.56	1.57	5.57
	Interest	.885	47	12, 20, 23,29	5.06	2.25	7.00
	Effort	.621	47	1, 2, 14, 27	6.44	4.75	7.00
Post	Affect	.817	45	3, 4*, 15*, 18*, 19, 28*	4.00	1.33	6.00
	Cognitive	.770	47	5*, 11*, 26*, 31, 32, 35*	4.97	2.67	6.83
	Value	.895	47	7*, 9, 10, 13*, 16*, 17, 21*, 25*, 33*	5.13	2.11	6.89
	Difficulty	.781	47	6, 8*, 22, 24*, 30*, 34*, 36*	3.79	1.57	5.57
	Interest	.926	47	12, 20, 23,29	4.33	1.00	6.75
	Effort	.769	47	1, 2, 14, 27	5.90	1.00	7.00

* Indicates an item where the responses have been reversed.

These high coefficient alphas indicate a satisfactory level of internal consistency between items in each subscale except for the Pre–Effort scale that had the lowest coefficient alpha .621. Because the psychometrics of the SATS-36 have been well established in the literature (e.g., Schau, 2005; Tempelaar, 2007), the Effort scale will be used here despite its low coefficient alpha.

Examining the First Research Question: What is the impact of using collaborative tests in an online statistics course on students learning?

Model 1: Midterm Exam as the Dependent Variable

To explore how using collaborative tests in an online statistics course impacts students' learning, a hierarchical multiple regression model was used. Midterm exam score was the dependent variable in all models. The possible predictor variables consisted of the three covariates (Section, Pretest and SPCT) and the three variables that measured the proportion of each type of posting (Teaching, Cognitive or Social) for the first and second group tests, both of which occurred prior to the Midterm exam. These latter three variables were named CLT1&2Discussion-Teaching, CLT1&2Discussion-Cognitive and CLT1&2Discussion-Social, respectively

To determine which discussion variables to use in the full model, the bivariate correlations among the Midterm exam and the three CLT1&2Discussion variables (Cognitive, Teaching and Social) was produced (see Table 13). Based on the bivariate correlations, the Cognitive discussion variable was excluded from the full model because it had a high correlation with the Social discussion variable and it had a low correlation with the dependent variable.

Table 13

Correlation for CLT1&2Discussion (Cognitive, Teaching and Social) and the Midterm Exam Variables

	1	2	3	4
(1) Midterm exam	1.000			
(2) CLT1&2Discussion-Teaching	.156	1.000		
(3) CLT1&2Discussion-Cognitive	.156	-.309*	1.000	
(4) CLT1&2Discussion-Social	-.262*	-.443**	-.716**	1.000

* $p < .05$, ** $p < .001$.

The full model was:

Model 1: Midterm exam = Section+ SPCT + Pretest +
CLT1&2Discussion-Social + CLT1&2Discussion-Teaching

The hierarchical regression model that was used had three steps, the first step included the three covariates Section, Pretest and SPCT, in the second step the CLT1&2Discussion-Social variable was added and in the third step the CLT1&2Discussion-Teaching variable was added. CLT1&2Discussion-Social was added first because it had a higher and a significant correlation with the dependent variable.

Table 14 reports the results of the three steps for the hierarchical regression for model 1. Based on these results, a reduced model with the three independent variables, Section, Pretest and CLT1&2Discussion-Social was used. Table 14 shows that CLT1&2Discussion-Teaching did not add much to the variance that was explained when it was added to the model with a statistically nonsignificant $\Delta R^2 = .010$. Furthermore, when CLT1&2Discussion-Teaching was in the model the relationship between Midterm

exam and CLT1&2Discussion-Social became weaker and statistically nonsignificant. The SPCT variable did not have a significant effect ($p > .05$) on the dependent variable at any stages of the hierarchical regression. It did not add anything to accounting for variance in the dependent variable in the overall model while it used one degree of freedom, and because the sample size was small at $n=54$, the SPCT variable was excluded from the final model.

Table 14

*Hierarchical Multiple Regression Reports for Model 1, Midterm Exam as Dependent**Variable*

	<i>B</i>	<i>SE B</i>	β	<i>t-test</i>	<i>p-value</i>
Step 1					
Constant	14.498	4.336		3.343	.002
Section	1.648	1.270	.161	1.298	.200
Pretest	.487	.140	.432	3.468	.001*
SPCT	.059	.074	.099	.795	.430
Step 2					
Constant	17.082	4.294		3.978	.000
Section	2.223	1.240	.217	1.793	.079
Pretest	.458	.135	.407	3.393	.001**
SPCT	.059	.070	.100	.837	.406
CLT1&2Discussion-Social	-7.674	3.256	-.286	-2.357	.022*
Step 3					
Constant	16.331	4.404		3.708	.001
Section	2.423	1.268	.237	1.911	.062
Pretest	.465	.136	.413	3.426	.001*
SPCT	.040	.074	.068	.542	.590
CLT1&2Discussion-Social	-6.432	3.601	-.240	-1.786	.080
CLT1&2Discussion-Teaching	4.352	5.309	.116	.820	.416

Note. $R^2 = .233^*$ for Step 1; $\Delta R^2 = .078^*$ for Step 2; $\Delta R^2 = .010$ for Step 3. * $p < .05$, ** $p < .001$.

Table 15 shows the results of the regression for the final model with the three independent variables that were used. The R^2 for the model was 0.293. Section, Pretest and CLT1&2Discussion-Social accounted for 29.3% of the variation in Midterm exam. The two independent variables Section and Pretest had a positive relationship with Midterm exam. The partial regression coefficient for Pretest was statistically significant ($p < .05$). The standard deviation for Pretest was 4.495 so by using the standardized coefficient β , as the Pretest variable increased by one standard deviation (4.495), the scores on the Midterm exam increased by .410 standard deviations when the effects of Section and CLT1&2Discussion-Social were held constant. An increase of one point on the Pretest was associated with an increase of .455 points on the Midterm exam while controlling for the effect of Section and CLT1&2Discussion-Social. The relationship between CLT1&2Discussion-Social and Midterm exam was negative and statistically significant ($p < .05$); for students who had a higher frequency of social indicators on the first and second group test, the Midterm exam score tended to be lower. The standard deviation for CLT1&2Discussion-Social was .191, so by using the standardized coefficient β , as the CLT1&2Discussion-Social variable increased by one standard deviation (.191), the scores on the Midterm exam decreased by -.289 standard deviations when the effects of Section and Pretest were held constant. The relationship between Section and Midterm exam was positive and not statistically significant ($p > .05$).

Table 15

Multiple Regression Reports for the Final Model with Midterm Exam as a Dependent

Variable

	<i>B</i>	<i>SE B</i>	β	<i>t-test</i>	<i>p-value</i>
Constant	20.134	2.366		8.511	.000
Section	2.200	1.165	.222	1.888	.064
Pretest	.455	.127	.410	3.570	.001*
CLT1&2Discussion1-Social	-7.505	3.058	-.289	-2.454	.017*

Note. $R^2 = .293$, * $p < .05$, ** $p < .001$.

To see if any accurate inferences could be made about the actual population, the final model with Midterm exam as the dependent variable needed to meet the assumptions outlined in Chapter 3.

Checking the assumptions for Model 1: Midterm exam as the dependent variable. To check the assumption of linearity and homoscedasticity, a plot of standardized residuals against the standardized predicted values was used (see Figure 1). Looking at the scatterplot in Figure 1 it can be seen that the values are randomly and evenly dispersed around a standardized residual value of zero which means that the assumptions of both linearity and homoscedasticity hold.

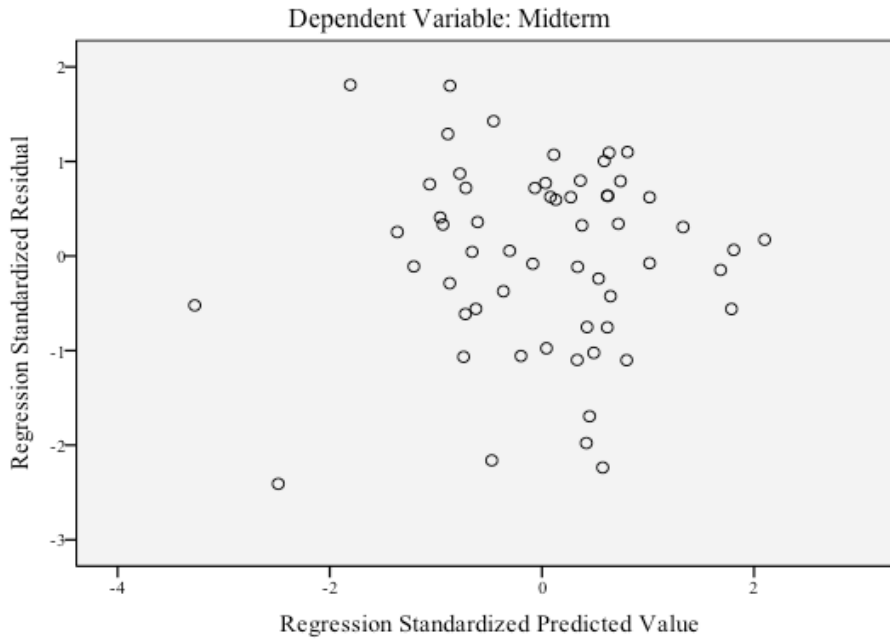


Figure 1. Scatterplot of the standardized residuals against the standardized predicted values for Model 1.

A normal probability plot of Midterm exam scores (see Figure 2) was created to check for the assumption of normality. Because the points follow the expected diagonal line, we can assume that the distribution of the standardized residuals is approximately normal and that the assumption of normality holds.

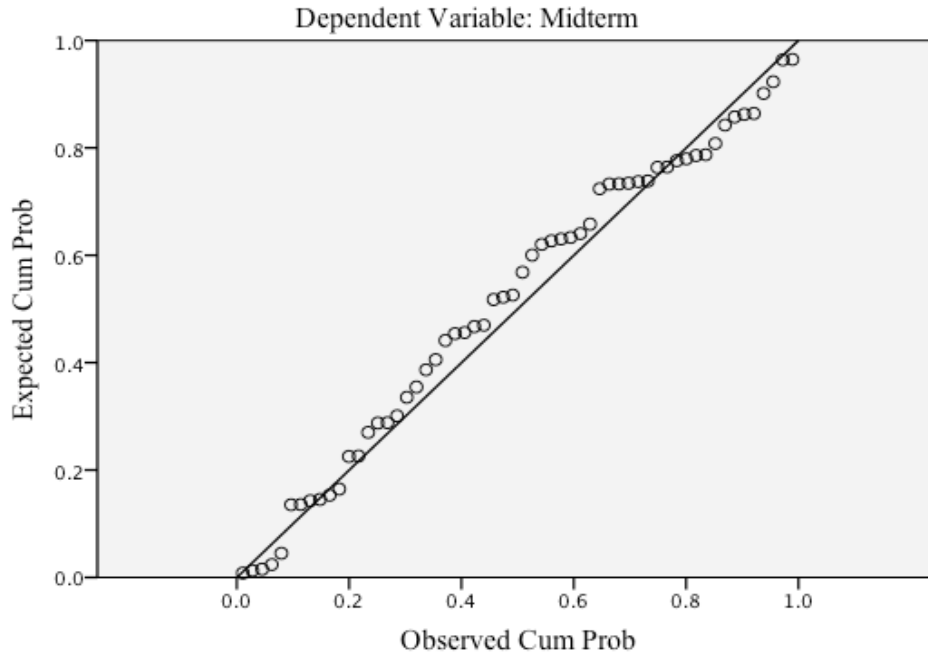


Figure 2. Normal P-P Plot of regression standardized residuals.

The Durbin-Watson test for serial correlation between residuals is used to see if the assumption of independence was met. The Durbin-Watson test was 1.569, which is close to 2, and suggests that no serial correlation was present meaning that the assumption of independence holds (Field, 2005).

To check for multicollinearity, the correlation between the predictor variables and the variance inflation factor (VIF) were explored. The highest correlation was only .221 between Section and Discussion-Social and the average VIF was 1.039, which is acceptable (Field, 2005). We can assume that the assumption of multicollinearity was not violated.

To test for interactions between the independent variables, the variables were mean-centered and a new regression model was built with the interaction terms included.

Variables in the model ending with C are mean-centered, for example SectionC. Because the final model with Midterm exam as the dependent variable only included three independent variables, the interaction model that was used included six parameters:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3$$

Table 16 shows the results of the regression for the final model with the 3 independent variables mean-centered and the three mean-centered interaction terms. Only 1 of the interaction terms SectionCxCLT1&2Discussion-SocialC was significant. The 2 other interaction terms were not significant ($p > .05$).

Table 16

Model 1, the Final Model with the Interaction Terms

	<i>B</i>	<i>SE B</i>	β	<i>t-test</i>	<i>p-value</i>
Step 1					
Constant	26.009	.567		45.868	.000
SectionC	2.200	1.165	.222	1.888	.064
PretestC	.455	.127	.410	3.570	.001*
CLT1&2Discussion-SocialC	-7.505	3.058	-.289	-2.454	.017*
Step 2					
Constant	26.389	.546		48.365	.000
SectionC	1.900	1.090	.192	1.743	.087
PretestC	.500	.121	.451	4.151	.000**
CLT1&2Discussion-SocialC	-4.725	3.013	-.182	-1.568	.123
SectionCxPretestC	-.371	.247	-.167	-1.502	.139
SectionC xCLT1&2Discussion-SocialC	-14.382	6.552	-.270	-2.195	.033*
PretestCxCLT1&2Discussion-SocialC	.670	.683	.117	.981	.331

Note. $R^2 = .293^*$ for Step 1; $\Delta R^2 = .128^*$ for Step 2; * $p < .05$, ** $p < .001$.

Because one of the interaction terms was significant ($p < .05$), an interaction model for the regression equation was created including the interaction term that was significant. The other two interaction terms were excluded from the model since they were not significant. The new regression equation that was used is below:

$$\text{The } E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_3$$

Table 17 includes the results for the interaction model. The R^2 for the model was .375. The three mean-centered independent variables and the interaction term for the mean-centered variables SectionC and CLT1&2Discussion-SocialC accounted for 37.5% of the variation in Midterm exam. This is an increase of 8% over the model without the interaction term. What is interesting here is that once the interaction term was added to the model, the CLT1&2Discussion-SocialC variables was no longer statistically significant ($p > .05$). It appears that the interaction of SectionC and CLT1&2Discussion-SocialC explains more of the variance in the dependent variable Midterm exam than CLT1&2Discussion-SocialC by itself.

Table 17

Multiple Regression Results for the Interaction Model, Midterm Exam as Dependent Variable

	<i>B</i>	<i>SE B</i>	β	<i>t-test</i>	<i>p-value</i>
Constant	26.357	.554		47.558	.000
SectionC	1.929	1.110	.195	1.737	.088
PretestC	.461	.121	.416	3.815	.000**
CLT1&2Discussion-SocialC	-4.884	3.069	-.188	-1.592	.117
SectionCxCLT1&2Discussion-SocialC	-16.124	6.124	-.302	-2.633	.011*

$R^2 = .375^{**}$, * $p < .05$, ** $p < .001$.

Figure 3 was created to visualize the relationship between the interaction term (SectionC and CLT1&2Discussion-SocialC) and Midterm exam. It appears that

CLT1&2Discussion-SocialC acts as a moderator in the relationship between Section and Midterm, which changes as a function of the level of the CLT1&2Discussion-SocialC. For students who were enrolled in the consensus section, their scores on the Midterm exam vary depending on their score on the CLT1&2Discussion-Social when Pretest is controlled for. Students who had a low proportion of Social dimension on the two group tests in the consensus section got, on average, a higher score on the Midterm compared to students who had a mean or a high proportion of social dimension in the same section. This difference is larger in the consensus section compared to the nonconsensus section, as shown in Figure 3.

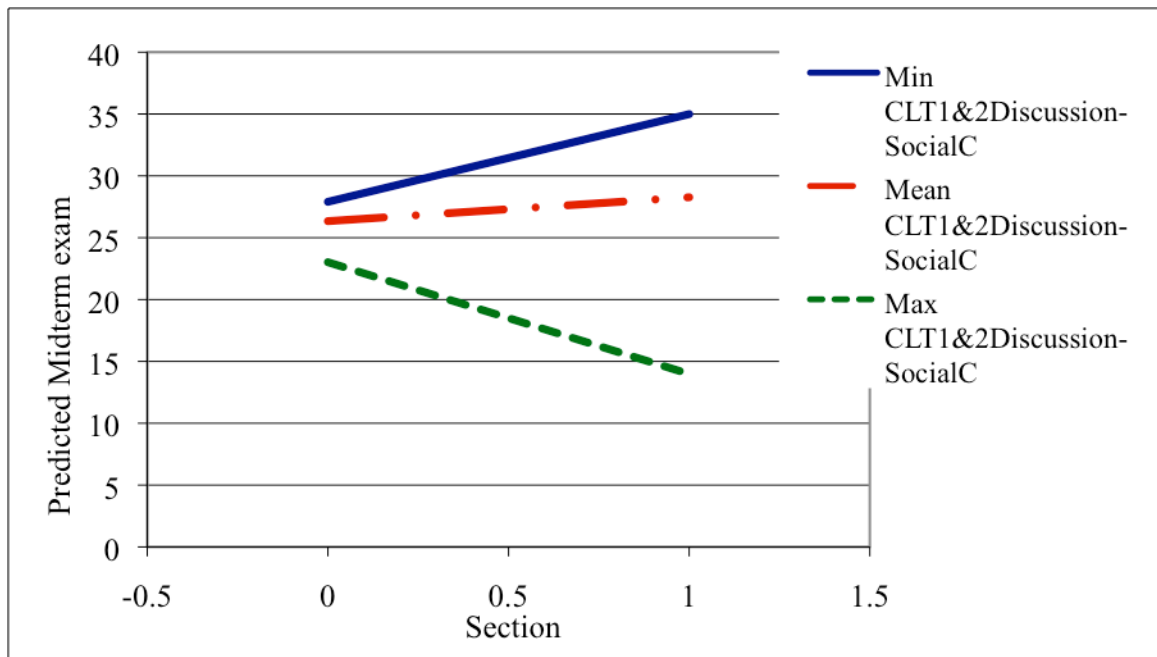


Figure 3. Interaction relationship between Midterm exam score and sections (Nonconsensus=0, Consensus=1), controlling for CLT1&2Discussion-Social.

Model 2: Final Exam as the Dependent Variable

The second hierarchical multiple regression model that was used to explore how using collaborative tests in an online statistics course impacts students' learning included

scores on the Final exam as a dependent variable and the Midterm exam scores as one of the independent variables.

For the first model with the Midterm exam as the dependent variable the discussion variables included proportion of posting (Cognitive, Social and Teaching) from only the first two collaborative tests. In the second model, however, the discussion variables that were used included proportion of posting from all three-collaborative tests. Again, to determine which CLTDiscussion variables would be included in the model, a correlation matrix between the dependent variable, Final Exam, and the three-CLTDiscussion variables was produced (see Table 18). Based on the bivariate correlations, the Cognitive CLTDiscussion variable was excluded from the first model because it had a high correlation with the CLTDiscussion Social variable and it had a low correlation with the dependent variable.

Table 18

Correlation for CLTDiscussion (Teaching, Cognitive and Social) and the Final Exam

Variable

	1	2	3	4
(1) Final Exam	1.000			
(2) CLTDiscussion-Teaching	.177	1.000		
(3) CLTDiscussion-Cognitive	.053	-.480**	1.000	
(4) CLTDiscussion-Social	.109	-.428**	-.588**	1.000

$p < .05$, ** $p < .001$.

The full model is presented below:

Model 2: Final Exam = Section + SPCT + Midterm exam +
CLTDiscussion-Teaching + CLTDiscussion-Social

Again, the hierarchical regression model that was used had three steps, the first step included the three covariates Section, Midterm exam and SPCT, in the second step the CLTDiscussion-Teaching variable was added and in the third step the CLTDiscussion-Social variable was added. CLTDiscussion-Teaching was added first because it had a higher correlation with the dependent variable.

Table 19 reports the results of the three steps for the hierarchical regression for model 2. Based on these results a reduced model with the three independent variables, Section, Midterm exam and CLTDiscussion-Teaching was used. Table 19 shows that CLTDiscussion-Social did not add much to the variance that was explained when it was added to the model with a statistically nonsignificant $\Delta R^2 = .001$. Again, the SPCT

variable did not add anything significant to the model at any stages of the hierarchical regression. Therefore, similar to model 1, the SPCT variable was not used for model 2.

Table 19

Hierarchical Multiple Regression Reports for Model 2, Final Exam as the Dependent Variable

	<i>B</i>	<i>SE B</i>	β	<i>t-test</i>	<i>p-value</i>
Step 1					
Constant	6.506	4.262		1.527	.133
Section	.403	1.157	.041	.348	.729
Midterm	.508	.114	.532	4.454	.000**
SPCT	.062	.066	.110	.934	.355
Step 2					
Constant	7.166	4.081		1.756	.085
Section	-.138	1.128	-.014	-.122	.903
Midterm	.524	.109	.550	4.803	.000**
SPCT	.105	.066	.186	1.590	.118
CLTDiscussion-Teaching	-12.321	5.121	-.285	-2.406	.020*
Step 3					
Constant	7.553	4.456		1.695	.097
Section	-.140	1.139	-.014	-.123	.903
Midterm	.526	.111	.552	4.759	.000**
SPCT	.105	.067	.187	1.580	.121
CLTDiscussion-Teaching	-12.840	5.650	-.297	-2.273	.028*
CLTDiscussion-Social	-1.045	4.584	-.028	-.228	.821

Note. $R^2 = .320^*$ for Step 1; $\Delta R^2 = .072^*$ for Step 2; $\Delta R^2 = .001$ for Step 3. * $p < .05$, ** $p < .001$.

Table 20 shows the results of the regression for the model with the three independent variables that were used and the Final Exam as dependent. The R^2 for the model was .385, which tells us that Section, Midterm exam and CLTDiscussion-Teaching accounted for 38.5% of the variation in Final Exam. The two independent variables Section and Midterm exam had a positive relationship with Final exam scores. Midterm exam was statistically significant ($p < .05$); an increase of one point on the Midterm exam (out of 35 points) was associated with an increase of .565 points on the Final exam while controlling for the effect of Section and CLTDiscussion-Teaching. The relationship between CLTDiscussion-Teaching and the Final exam scores was negative and statistically significant ($p < .05$), meaning the higher the frequency of Teaching indicators on the three group tests, the lower the scores on the Final Exam tended to be. The standard deviation for CLTDiscussion-Teaching was .121 so by using the standardized coefficient β we can say as the CLTDiscussion-Teaching variable increased by one standard deviation (.121), the scores on the Final Exam decreased by .247 standard deviations when controlling for the effects of Section and Midterm.

Table 20

*Multiple Regression Reports for the Final Model with Final Exam as a Dependent**Variable*

	<i>B</i>	<i>SE B</i>	β	<i>t-test</i>	<i>p-value</i>
Constant	10.753	2.767		3.886	.000
Section	.051	1.097	.005	.047	.963
Midterm	.565	.102	.598	5.534	.000**
CLTDiscussion-Teaching	-10.045	4.521	-.247	-2.222	.030*

Note. $R^2 = .385$, * $p < .05$, ** $p < .001$.

As before the assumptions for the regression model needed to be checked to see if accurate inferences could be made about the actual population.

Checking the assumptions for Model 2: final exam as the dependent variable.

The scatterplot of the standardized residuals against the standardized predicted values in Figure 4, shows that the values are randomly and evenly dispersed around a residual value of zero which means that the assumptions of both linearity and homoscedasticity hold.

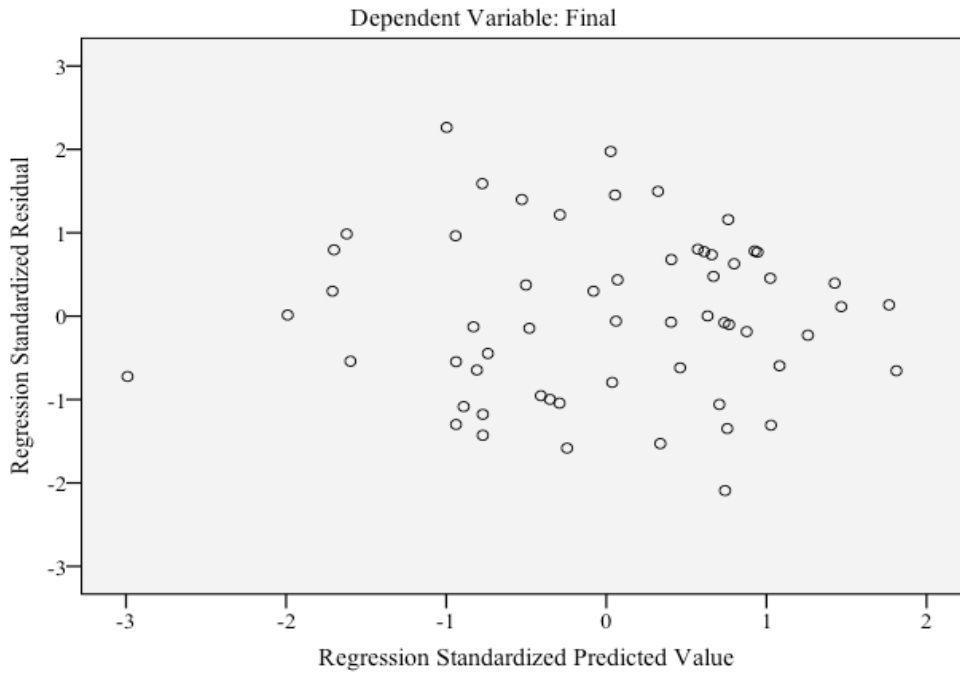


Figure 4. Scatterplot of the standardized residuals against the standardized predicted values for Model 2.

A normal probability plot (Figure 5) plot shows an approximately straight line, we therefore assume that the distribution of the standardized residuals is approximately normal and that the assumption of normality holds.

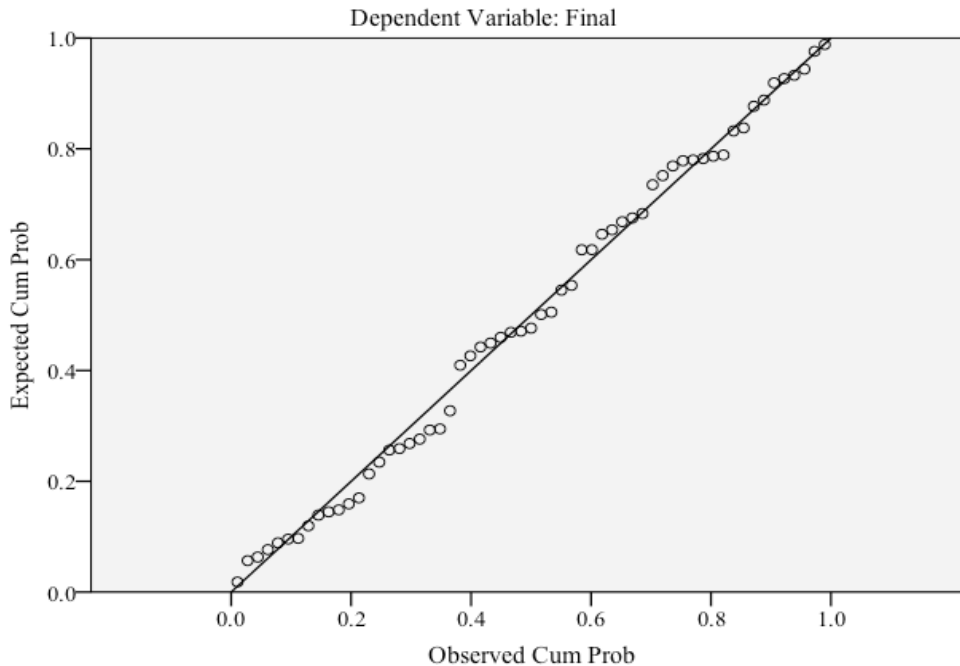


Figure 5. Normal P-P Plot of the regression standardized residuals.

The Durbin-Watson test was 1.634, which is close to 2 this suggests that no serial correlation is present meaning that the assumption of independence holds (Field, 2005). The highest correlation was only $-.273$ between Section and CLTDiscussion-Teaching and the average variance inflation factor (VIF) was 1.088, which is acceptable (Field, 2005). Based on this we can assume that the assumption of multicollinearity is not violated.

To test for interaction we first mean-centered the independent variables and then ran a new regression model with the interaction terms included. Just like model 1 with the Midterm exam as the dependent variable, the model with the Final Exam score as the dependent variable only included three independent variables; the interaction model that was used included six parameters:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3$$

Table 21 shows the results of the regression for the model with Final Exam as a dependent variable and with the three independent variables mean-centered and the three mean-centered interaction terms. None of the three interaction terms was significant ($p > .05$). Therefore, the interaction terms were not included in the final model.

Table 21

Model 2, the Model with the Interaction Terms, Final Exam as a Dependent Variable

	<i>B</i>	<i>SE B</i>	β	<i>t-test</i>	<i>p-value</i>
Step 1					
Constant	23.049	.520		44.296	.000**
SectionC	.051	1.097	.005	.047	.963
MidtermC	.565	.102	.598	5.534	.000**
CLTDiscussion-TeachingC	-10.045	4.521	-.247	-2.222	.030*
Step 2					
Constant	23.291	.592		39.360	.000**
SectionC	-.404	1.163	-.041	-.347	.730
MidtermC	.609	.111	.645	5.475	.000**
CLTDiscussion-TeachingC	-15.175	6.266	-.374	-2.422	.019*
SectionCxMidtermC	-.080	.230	-.042	-.346	.731
SectionC xCLTDiscussion-TeachingC	14.489	12.565	.172	1.153	.254
MidtermCxCLTDiscussion-TeachingC	.552	.974	.067	.566	.574

Note. $R^2 = .385$ ** for Step 1; $\Delta R^2 = .021$ for Step 2; * $p < .05$, ** $p < .001$.

Examining the Second Research Question: What is the effect of using collaborative tests on students' attitudes towards statistics?

In an effort to help answer the second research question about the relationship between students' attitudes towards statistics and working on different collaborative tests, information was gathered using the SATS-36 instrument. The subscales that were produced from the SATS-36 were used to see if there were any differences between sections. Table 22 shows the means and standard deviations and the t-tests for the presubscales. No significant differences ($p > .05$) were found between the two sections on scores on any of the six pre-SATS-36 subscales.

Table 22

Mean Preresponses on the SATS-36 Subscales by Section

	<i>Consensus Section</i>		<i>Nonconsensus Section</i>		<i>t-test</i>	<i>p-value</i>
	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>		
Affect	3.7222	1.08644	4.2391	1.10076	-1.620	.112
Cognitive	4.8889	.70139	5.0870	.96905	-.805	.425
Value	5.3519	.76186	5.5942	.95500	-.964	.340
Difficulty	3.6012	.59721	3.5584	1.14278	.161	.873
Interest	4.7500	1.20009	5.3913	1.09965	-1.908	.063
Effort	6.3854	.56616	6.5109	.56144	-.762	.450

Table 23 shows that there were no significant differences ($p > .05$) on the six post-SATS-36 subscales between the two sections.

Table 23

Mean Postresponses on the SATS-36 Subscales by Section

	<i>Consensus Section</i>		<i>Nonconsensus Section</i>		<i>t-test</i>	<i>p-value</i>
	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>		
Affect	4.1667	1.03353	3.7754	1.34803	1.105	.275
Cognitive	5.0694	.81043	4.8043	1.13223	.926	.359
Value	5.1667	.89850	5.1111	1.20558	.180	.858
Difficulty	4.0238	.73771	3.5590	.97740	1.845	.072
Interest	4.2292	1.36317	4.4457	1.64448	-.492	.625
Effort	5.8958	.75512	5.9239	1.33875	-.089	.929

Another way to detect a possible difference on the subscales was to explore the mean difference, or change score between pre- and post-SATS-36 for the six subscales. Table 24 shows the mean change score for each subscale. A positive score reflects improvement from the pre- to post-SATS-36 instruments, which would mean an increase in attitude towards statistics. A negative score reflects a decline in attitude towards statistics as measured by the SATS-36 instruments. As shown in Table 24, there were no significant differences ($p > .05$) found between the two sections on their mean difference scores on the six SATS-36 subscales.

Table 24

Mean Difference Scores on the SATS-36 Subscales by Section

	<i>Consensus Section</i>		<i>Nonconsensus Section</i>		<i>t-test</i>	<i>p-value</i>
	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>		
Affect	.1111	1.26326	-.1667	1.09954	-.792	.432
Cognitive	.9097	1.07224	1.1970	1.05865	.913	.366
Value	-.1289	1.06414	-.2803	1.20252	-.458	.649
Difficulty	1.5467	1.00821	1.8189	1.33507	-.234	.431
Interest	1.3423	1.47817	1.5714	1.49123	.523	.604
Effort	-1.5300	1.65724	-1.6250	1.00519	.795	.816

To explore if there was any difference between the pre- and post-SATS-36 subscales within each section, a series of one-sample t-tests were conducted to see if the mean difference scores were significantly different from zero. The results of these tests are displayed below in Table 25. For both sections, there was a significant increase ($p < .05$) (from pretest to posttest) on four subscales: Cognitive, Difficulty, Interest and Effort. Students in both sections had a significant decrease on the effort subscale, which means that the amount of effort they put in learning statistics at the end of the course was less than what they expected to put in at the start of the course.

Table 25

Test of Mean Difference Scores on SATS-36 Subscales within Section

	<i>Consensus Section</i>		<i>Nonconsensus Section</i>	
	<i>t-test</i>	<i>p-value</i>	<i>t-test</i>	<i>p-value</i>
Affect	.431	.671	-.711	.485
Cognitive	4.156	.000**	5.303	.000**
Value	-.606	.550	-1.093	.287
Difficulty	7.670	.000**	6.390	.000**
Interest	4.449	.000**	4.943	.000**
Effort	-4.616	.000**	-7.583	.000**

* $p < .05$, ** $p < .001$.

Examining the Third Research Question: How does using a required consensus on collaborative tests vs. a nonconsensus approach effect group discussions?

The quantitative variables that were constructed from discussion posts during the collaborative test using the Pozzi et al. (2007) framework were explored in depth to help answer the third research question.

Figure 6 shows the mean proportion for the three dimensions--Social, Cognitive, and Teaching—for both sections across the three group tests. It can be seen that the Cognitive dimension had the highest mean for all three tests. The mean proportion for the Teaching dimension decreased from the first to the last test, while the means increased slightly from the first to the last test for the two other dimensions (Cognitive and Social).

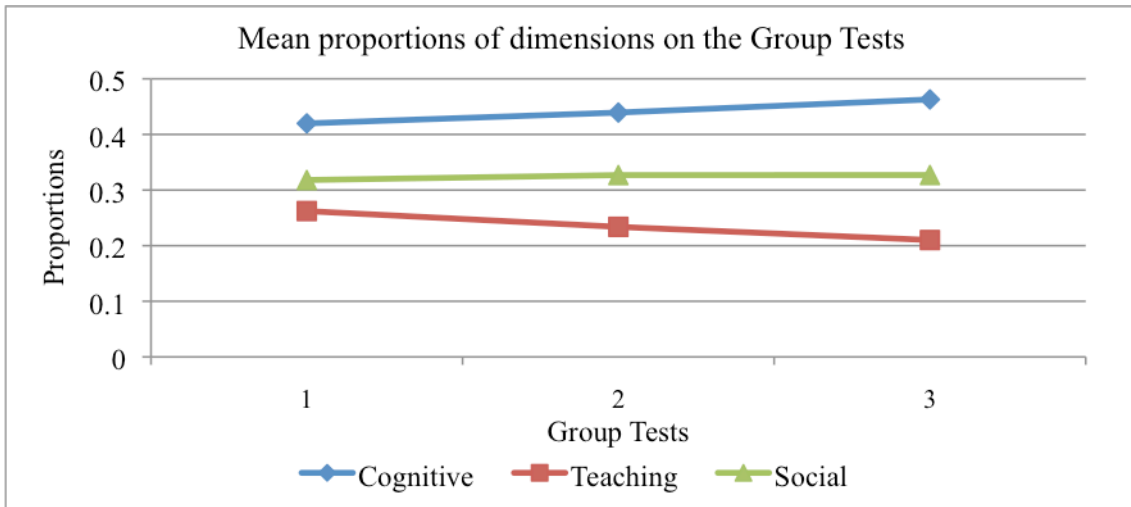


Figure 6. Mean proportions of three dimensions on the three group tests.

Between the two sections, the largest and only statistically significant difference ($p < .05$) was on the Teaching dimension. In the consensus section, the mean proportion for that dimension was 20.3% compare to 27.0% in the nonconsensus section (see Table 26). For the other two dimensions, Cognitive and Social, the difference between the sections was only 2 to 3.85% with a higher proportion in the consensus section. It appears that there was a higher proportion of the Teaching dimension that took place in the nonconsensus section compared to the consensus section.

Table 26

Tests of Mean Proportions of Different Dimensions between the Two Sections

	<i>Consensus Section</i>		<i>Nonconsensus Section</i>		<i>t-test</i>	<i>p-value</i>
	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>		
Social	.3385	.14249	.3001	.11787	-1.114	.270
Cognitive	.4581	.14664	.4303	.12345	-.779	.439
Teaching	.2034	.14632	.2696	.07134	2.143	.036

The difference between proportions of the three dimensions between the two sections can be seen in Figure 7. The difference between the two sections on the Teaching dimensions is clear from that graph.

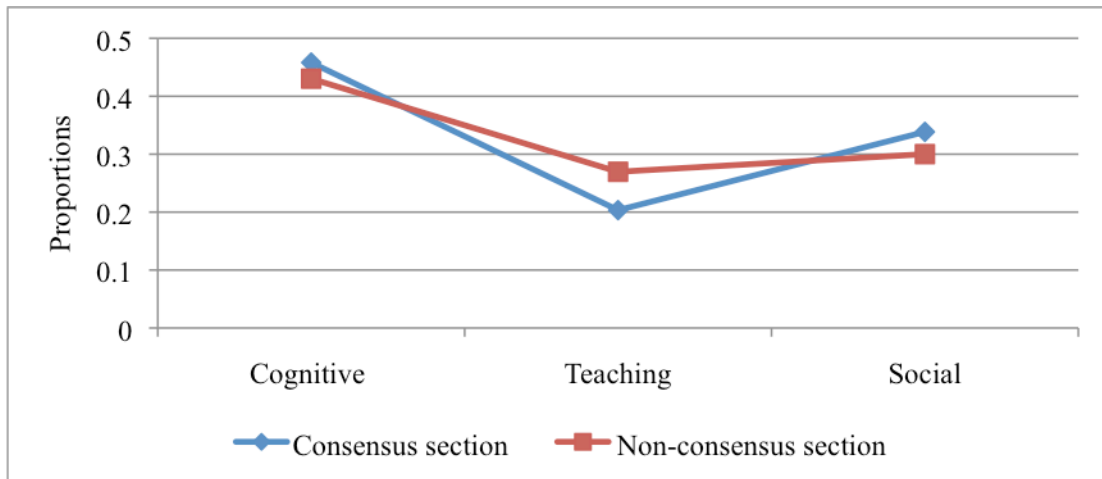


Figure 7. Mean proportion of the three dimensions between sections.

Table 27 shows the mean proportions of the 10 indicators for the three dimensions between sections. The largest difference is for the Cognitive dimension indicator *Exploration* (which is about expressing agreement or disagreement, sharing ideas and information, brainstorming and negotiating). In the consensus section, the mean proportion was 25.58% compared to 20.1% in the nonconsensus section. The mean proportion for the *Affection* indicator (which is about expression of emotions, intimacy, and personal anecdotes) for the Social dimension was also higher for the consensus section (6.5%) compared to 3.64% in the nonconsensus section.

However, the only statistically significant difference ($p < .05$) for the two sections was between the three Teaching dimensions indicators. The consensus section had a higher mean proportion for the *Organisational matters* indicator (which is about introducing topics, providing explanations for methods and letting students know of deadlines) while the nonconsensus section had a higher mean for the *Direct instruction* indicator (which is about recommending activities, pointing out misconceptions, providing feedback and assessment that confirm understanding) and the *Facilitation discourse* indicator (which is about identifying areas of agreements/ disagreement to achieve consensus, encouraging, acknowledging or reinforcing participants contribution, setting the climate for learning).

Table 27

Tests of Mean Proportions for the 10 Indicators between Sections

	<i>Consensus Section</i>		<i>Nonconsensus Section</i>		<i>t-test</i>	<i>p-value</i>
	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>		
Social presence						
Affection	.0650	.07262	.0364	.05214	-1.710	.093
Cohesion	.2734	.12126	.2637	.10826	-.323	.748
Cognitive presence						
Revelation	.0994	.06785	.0911	.08464	-.418	.678
Exploration	.2558	.13906	.2010	.09062	-1.756	.084
Integration	.0573	.06288	.0659	.05716	.540	.591
Resolution	.0252	.03874	.0354	.04159	.975	.334
Metareflection	.0204	.03328	.0369	.08629	1.003	.320
Teaching presence						
Direct instruction	.0501	.05742	.0834	.05730	2.219	.030*
Facilitation	.1254	.10693	.1745	.07004	2.045	.046*
Organisational matters	.0279	.03445	.0117	.02232	-2.101	.040*

* $p < .05$, ** $p < .001$.

There was no statistically significant difference ($p > .05$) between the sections on the Participative dimension on the three group tests. In both sections, students' mean proportion for participation was highest on group test 2.

Table 28

Tests of Mean Proportions for the Participative Dimension between Sections

	<i>Consensus Section</i>		<i>Nonconsensus Section</i>		<i>t-test</i>	<i>p-value</i>
	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>		
Group test #1	.273	.127	.295	.116	.675	.502
Group test #2	.367	.122	.358	.105	-.307	.760
Group test #3	.358	.168	.346	.124	-.319	.751

Cognitive dimension. The proportion of indicators within the Cognitive dimension for both sections can be seen in Figure 8. The *Exploration* indicator was most common followed by the *Revelation* indicator (which is about recognizing a problem, demonstrating a sense of puzzlement and explaining a point of view) in both sections. The difference between the two sections on the proportions of indicators for the Cognitive dimension is minimal as can be seen on Figure 8.

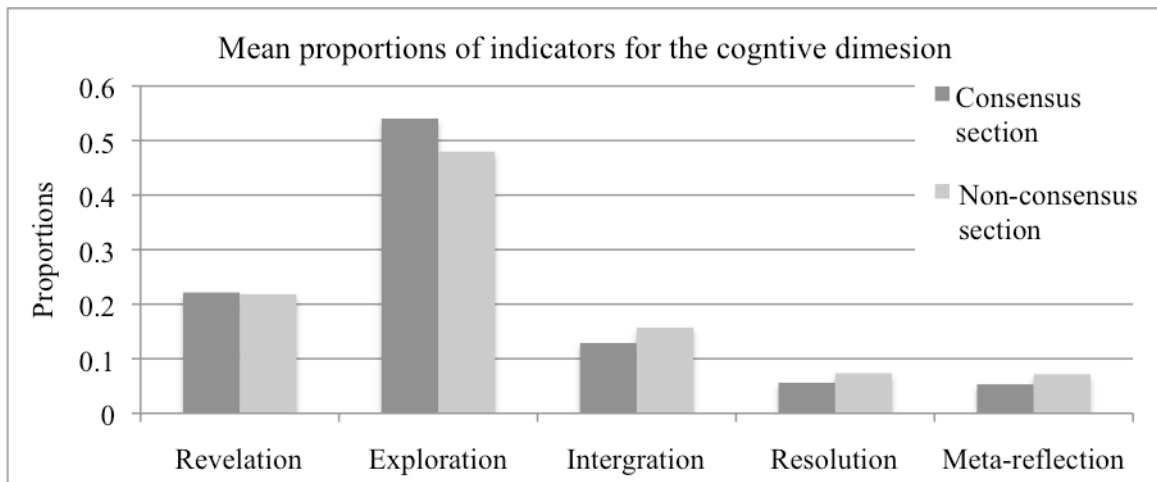


Figure 8. Mean proportions of indicators on the cognitive dimension between the two sections.

None of the mean proportions for Cognitive dimension indicators were significantly different ($p < .05$) between the two sections (see Table 29).

Table 29

Test of Mean Proportions for the Cognitive Dimension Indicators between Sections

	<i>Consensus Section</i>		<i>Nonconsensus Section</i>		<i>t-test</i>	<i>p-value</i>
	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>		
Revelation	.2215	.15201	.2184	.21837	-.064	.949
Exploration	.5403	.16441	.4796	.21975	-1.212	.231
Integration	.1289	.13649	.1571	.13986	.780	.439
Resolution	.0560	.07817	.0734	.08287	.826	.412
Metareflection	.0532	.08534	.0715	.15766	.566	.573

Social dimension. Figure 9 shows the proportion of the two indicators on the Social dimension between the sections. Both sections had high proportions of the *Cohesion* indicator (which is about vocatives, references to the group, and salutations); in the nonconsensus section it was 88% compare to 81% in the consensus section.

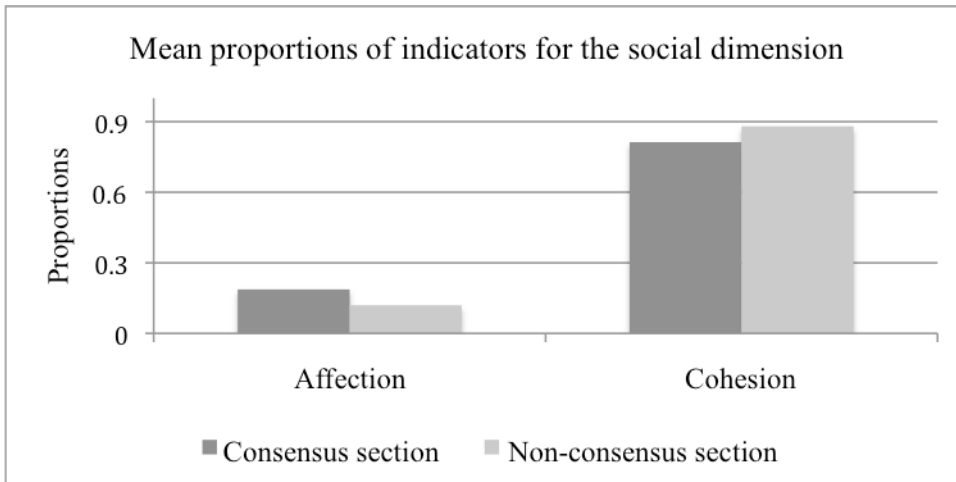


Figure 9. Mean proportions of indicators on the social dimension between sections.

There was not a significant difference ($p > .05$) in mean proportions for these two indicators between the sections, see Table 30.

Table 30

Test of Mean Proportions for the Social Dimension Indicators between Sections

	<i>Consensus Section</i>		<i>Nonconsensus Section</i>		<i>t-test</i>	<i>p-value</i>
	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>		
Affection	.1801	.18387	.1197	.16490	-1.309	.196
Cohesion	.8199	.18387	.8803	.16490	1.309	.196

Teaching dimension. The mean proportion for the three indicators on the Teaching dimension between the sections can be seen on Figure 10. The *Facilitation* indicator had the highest proportion in both sections; it was higher in the nonconsensus section 64.6% compared to 57.7% in the consensus section. The largest difference

between the sections was for the *Organizational matters* indicator. In the consensus section the mean proportion was 12.47% compared to 5.11% in the nonconsensus section.

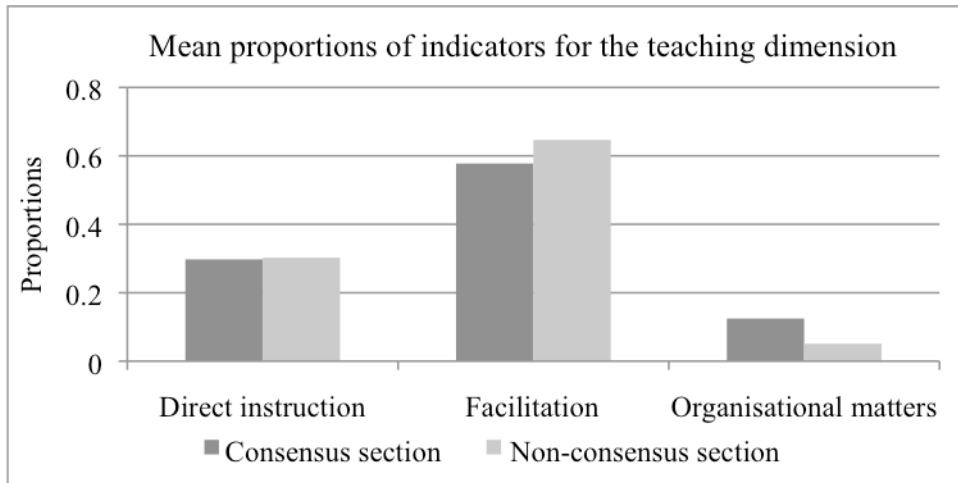


Figure 10. Mean proportions of indicators on the teaching dimension between sections.

The mean proportion for the *Organizational matters* indicator was significantly different between the sections ($p < .05$), with a higher mean in the consensus section (see Table 31).

Table 31

Test of Mean Proportions for the Teaching Dimension Indicators between Sections

	<i>Consensus Section</i>		<i>Nonconsensus Section</i>		<i>t-test</i>	<i>p-value</i>
	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>		
Direct instruction	.2857	.26995	.3024	.20090	.255	.800
Facilitation	.5946	.24663	.6465	.19835	.839	.405
Organisational matters	.1197	.11380	.0511	.09865	-2.328	.024*

* $p < .05$, ** $p < .001$.

Correlations. Table 32 shows the bivariate correlations among the 10 indicators. The strongest significant relationship was $r = .857$ between the two Teaching dimension indicators *Direct instruction* and *Organizational matters*. The strongest correlation between indicators from different dimensions was $r = -.452$ between the Cognitive dimension indicator *Exploration* and the Teaching dimension indicator *Direct instruction*. For the Social dimension, the strongest correlation was $r = .387$ for the *Affection* indicator and the Teaching dimension indicator *Direct instruction*.

Table 32

Correlations between the 10 Indicators

<i>Indicators</i>	<i>Correlations between Indicators</i>									
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	<i>9</i>	<i>10</i>
(1) Revelation	1.000	.028	-.348**	-.016	-.258*	-.130	-.238	-.037	.059	-.071
(2) Exploration	.028	1.000	-.076	-.335**	-.302*	-.452**	-.392**	-.284*	-.164	-.225
(3) Integration	-.348**	-.076	1.000	.268*	-.035	-.042	.311*	-.202	-.149	-.097
(4) Resolution	-.016	-.335**	.268*	1.000	.403**	.047	.215	-.051	-.072	-.080
(5) Metareflection	-.258*	-.302*	-.035	.403**	1.000	.001	-.106	.066	.009	.127
(6) Direct instruction	-.130	-.452**	-.042	.047	.001	1.000	.491**	.857**	.387**	-.262*
(7) Facilitation	-.238	-.392**	.311*	.215	-.106	.491**	1.000	.041	-.111	.102
(8) Organisational matters	-.037	-.284*	-.202	-.051	.066	.857**	.041	1.000	.246	-.358**
(9) Affection	.059	-.164	-.149	-.072	.009	.387**	-.111	.246	1.000	-.139
(10) Cohesion	-.071	-.225	-.097	-.080	.127	-.262*	.102	-.358**	-.139	1.000

Additional Analysis

Students' responses on the Students Perception on Collaborative Tests (SPCT) instrument are explored in this section since this information were not used to answer the three research questions. Of the 59 students who were enrolled in the two sections, 54 completed the SPCT instrument (29 students from the consensus section and 25 students the nonconsensus section) and no statistically significant difference ($p > .05$) was found in responses on the instrument between the two sections. Overall responses were positive towards the collaborative tests: the average score for the SCPT scale score was 2.89 with a minimum score of 1.83 and a maximum score of 3.89 out of 4 (see Table 12).

When asked if *Instructions for doing group tests resulted in everyone in the group contributing equally*, 59.3% of respondents agreed or strongly agreed. Only 9.3% of the students strongly disagreed with the statement *In general, I was an active participant in all three of the group tests*. Other students either agreed or strongly agreed with this statement. For the statement, *Working together on group tests helped me remember information that I had forgotten more than if I had taken the test on my own*, 75.9% of students agreed or strongly agreed. About 74% agreed or strongly agreed with the following statement: *Working together on group tests often helped me revise my initial answers on the tests* and 77.8% of the students who responded agreed or strongly agreed that *Participation in group tests was an important aspect of learning statistics in this course*. In the consensus section, 41.3% of those that responded agreed or strongly agreed with the statement *I would have preferred to take individual tests*, while in the nonconsensus section the proportion that agreed or strongly agreed to the similar statement, *I would have preferred to take only individual tests* was only 24%. For the

consensus section 68.9% of students would have preferred to discuss their answers with their group but to submit them individually while 40% of the students in the nonconsensus group preferred to discuss and submit their answers with their group.

The survey included two open-ended questions that asked respondents to list what they liked and did not like most about group tests. For these two open-ended questions, 46 students responded to what they liked most about the group tests and 42 students replied to what they did not like about the group tests. The next section highlights these responses.

What students liked most about the group tests. Students' responses to what they liked most about the group tests were classified into three major themes: comparing and understanding, group work, and confidence in statistics. The themes overlapped in some cases. Comparing and understanding included responses that mentioned the option of comparing each other's answers and helping with understanding, either by asking questions or teaching to another group member. This was the most common of all the responses. An example of this type of response is below:

Being able to discuss individual questions with other people was helpful. Sometimes I'd remember certain things by talking in the discussions; sometimes I'd realize that my first thoughts were inaccurate. Helping others understand was helpful too- I'd try to teach someone why I'd gotten a particular answer, and would understand it better when I explained it to them.

Group work included responses that highlighted the work done by the group, such as timely posting and interactions, as these two responses illustrate:

Being in an online course, the group tests were one way to help me interact with my classmates.

Our group was awesome at explaining answers and helping others see why they answered the way they did. We were all very active with the tests.

Not everybody liked their group however as this response shows:

Not much, I had a very lazy group that waited until the 11th hour to discuss the answers. I actually did worse on the tests than if I had submitted them myself.

Some responses included references to students' own mathematics or statistics ability or confidence. For example:

I don't have a lot of confidence when it comes to statistics. Group tests reinforced when I was on the right page. It also helped in my understanding of concepts I didn't quite understand. I typically loathe group work, but it was actually one of the most beneficial aspects of this course.

Responses were similar between the consensus and nonconsensus sections and a majority of the comments were related to the comparing and understanding theme.

What students did not like about the group tests. Students' responses to what they did not like about the group tests were classified into two major themes: grading and participation. Again, in some cases, the themes overlapped. Here there was a difference in the responses between the sections. For the consensus section, the grading was mostly about disliking that students had to agree about answers and also that one person was responsible for turning in those answers, as this response shows:

Accountability. Not in the participation sense but in more in a, I'd like to see what the group leader is going to submit before they submit it. Like a preview page where all members view and sign off on it. Otherwise they were great!

Participation was mostly about the discussion part of the tests, and how active the groups were. For example:

Lack of participation in group. Very hard to collaborate with peers with varying background in math/stats, meaning that some truly wanted to learn material while others just wanted to "get it over with"...

Some students did not feel that mandatory discussion was beneficial for them:

I didn't like that you had to post if you didn't have questions after your initial post, I didn't think it was helpful unless you were confused in which case you would have posted anyway.

For the nonconsensus section, responses were mostly about participation in addition to some students who complained about having to take the individual part of the test in that section.

Summary

This chapter presented the results from the study. It described the two online sections, in terms of students' gender, academic level and primary program. It outlined the reliability and scale scores that were produced for the subscales that were used to help answer the research questions. The three research questions were examined using hierarchical multiple regression, descriptive statistics, correlations and t-tests. The open-ended responses from the SPCT instrument that were used were also examined. The next chapter will summarize the findings from this study, discuss its value and limitations, and discuss what meaning this might have for the future use of collaborative tests in online introductory statistics courses.

Chapter 5

Discussion

This chapter summarizes the main findings of the study. It will explore how the study addressed each of the research questions. It discusses some of the limitations of the research and summarizes implications for future research.

The study explored the effects of using collaborative tests in an online introductory statistics course on students' learning. Three collaborative tests were implemented in two online sections of the EPSY-3264 Basic & Applied Statistics course in Fall 2011. Two different treatments were used:

- The Consensus section: students worked on collaborative tests together, reached a consensus on answers and turned answers in as a group.
- The Nonconsensus section: students worked on three collaborative tests together but turned them in individually.

In both treatments, the collaborative tests were graded in terms of correctness and students' participation on discussion for each test. Full participation was defined as one initial comment with answers to the test, followed by two questions or comments regarding the content of the test. The number of students who took part in the study was 59; in the Consensus section there were 32 students and in the Nonconsensus section there were 27 students. Students were randomly assigned to treatments to protect against confounding. The study sought to answer these three research questions:

1. What is the impact of using collaborative tests in an online statistics course on students' learning?

2. What is the effect of using collaborative tests on students' attitudes towards statistics?
3. How does using a required Consensus on collaborative tests vs. a Nonconsensus approach affect group discussions?

The following types of instruments were used to collect quantitative data for the study:

- The Comprehensive Assessment of Important Outcomes in Statistics (CAOS) test was used to see if there was a change in students' learning between the two treatments. CAOS was administered both as a pretest at the beginning of the semester and as a final exam at the end of the semester. Only 33 items from the 40 items on the CAOS were used.
- A Midterm exam was made up of items from the CAOS test (items that were not used on the pretests and final exam), from the ARTIST online database, and items used previously in the course. The Midterm was used to measure students' learning at the middle of the semester.
- The Survey of Attitudes Toward Statistics (SATS-36) was used to explore the effects of using different formats of collaborative tests on students' attitudes towards statistics. A Pre- and Postmeasure of the SATS-36 were used.
- The Students Perception on Collaborative Tests (SPCT) survey was used to measure students' perceptions towards taking the collaborative tests.

Qualitative data were also gathered using a framework (Pozzi et al., 2007) that evaluates and monitors computer-supported collaborative learning processes. In this

study, the focus was on the three dimensions that take place in a learning community: Cognitive, Social and Teaching. The Cognitive dimension refers to how much learners are “able to construct and confirm meaning through sustained reflection and discourse in a critical community of inquiry” (Garrison et al., 2001 in Persico et al., 2010, p. 9). The Social dimension refers to the social presence of students in the course while the Teaching dimension is the “ the design, facilitation, and direction of cognitive and social processes for the purpose of realizing personally meaningful and educationally worthwhile learning outcomes” (Anderson et al., 2001 in Pozzi et al., 2007, p. 174). Indicators that express each of these three dimensions were identified from students’ discussion comments during the collaborative tests. Number of indicators varied for each dimension. Five indicators express the Cognitive dimension, two express the Social dimension, and three express the Teaching dimension. The unit of analysis was each post or discussion comment during the collaborative test. A maximum of 3 indicators were identified for each discussion comment. Indicators could be from the same or different dimensions. Proportions out of total frequency of indicators for the three dimensions were computed and used as quantitative variables in the data analysis.

The data analysis included a hierarchical multiple regression model that was used to explore the effects of using collaborative tests on students’ learning. In addition, descriptive statistics and t-tests were used to examine changes in students’ attitudes towards statistics and the effects of using different formats of collaborative tests on group discussion.

Research Question 1. What is the impact of using collaborative tests in an online statistics course on students' learning?

Students' scores increased from the pretest to the final exam. On all three collaborative tests, discussion comments classified as Cognitive were most common; these comments did not however explain a significant amount of variability in scores on either the Midterm exam or the final exam. However, students who displayed more discussion comments classified as Social or Teaching on the collaborative tests had lower scores on the Midterm exam and the final exams. In this study the impact of the collaborative tests on students' learning was negative for students who displayed more discussion comments that were classified as Social or Teaching dimensions during the collaborative tests.

For the first two collaborative tests, discussion comments classified as Social had a significant negative effect on students' scores on the Midterm exam. This means that students who displayed more of a Social dimension on their discussion comments when working on the first two collaborative tests got a lower score on the Midterm exam. For the two treatments, there was a larger effect for students in the Consensus section between their Midterm exam score and discussion comments classified as Social compared to students in the Nonconsensus section. This might indicate that students in the Consensus section that had more discussion comments classified as *Social* on the two collaborative tests had less understanding of the material compared to students in the Nonconsensus section with similar amount of discussion comments classified as *Social*. It could be that students who were not fully grasping the material in the Consensus section might have displayed more of a Social dimension in their discussion comments on the

collaborative tests in order to fulfill the participation requirement. These students might still have received a good grade on the collaborative tests due to the consensus part. Nonetheless, the grading on the collaborative tests, where students needed to participate to earn a credit, might have influenced the discussion in a way that it was more superficial regarding the content instead of trying to understand or discover misconceptions they might have had regarding the material. While their peers in the Nonconsensus section had more discussion comments classified as Teaching, their discussion revolved more about pointing out misconceptions and asking for clarifications regarding the material, possibly because there was more at stake for them due to the individual grading. However, it can also be argued that there was much at stake for students' in the Consensus section because they needed to reach agreement regarding the answers. There was not a significant difference between the two treatments and discussion comments classified as Social, the only difference were that the effects for these discussion comments and Midterm exam scores were larger in the Consensus section. Because of this, it is hard to explain the interaction between discussion comments classified as Social on the first two collaborative tests and the two treatments.

For all three collaborative tests, discussion comments classified as Teaching had negative effects on scores on the final exam. In other words, students who displayed more discussion comments classified as Teaching got a lower score on the final exam. This might indicate that students who had more discussion comments classified as Teaching might have been trying to teach each other but might not have been teaching the right things. These students might have been struggling more with the material without realizing it since their discussion on the three collaborative tests was more geared to the

three indicators *Direct instruction* (recommending activities, pointing out misconceptions, providing feedback and assessment that confirm understanding), *Facilitating discourse* (identifying areas of agreements/ disagreement to achieve consensus, encouraging, acknowledging or reinforcing participants contribution, setting the climate for learning) and *Organisational matter* (introducing topics, providing explanations for methods and letting students know of deadlines) within the Teaching dimension.

Research Question 2. What is the effect of using collaborative tests on students' attitudes towards statistics?

Students' attitudes in both treatments increased in terms of their intellectual knowledge, skills, and interest towards statistics. However, at the end of the semester, students' perceived statistics to be more difficult and they put less effort into learning statistics.

An increase in students' attitudes towards learning has been reported both with using collaborative learning (Potthast, 1999) and with using collaborative tests (e.g., Giraud & Enders, 2000; Ioannou & Artion, 2010). However, because this study did not include a group of students who took the same test individually without the collaborative part, we cannot say that the noticeable increase in students' attitudes in terms of their intellectual knowledge, skills, and interest towards statistics was due to the three collaborative tests, other factors in both treatments might have contributed to this.

From the beginning to the end of the semester, students' attitudes towards their own intellectual knowledge and skills in regards to statistics increased. Their interest in statistics as a subject also increased. Interestingly, at the end of the course students'

attitude towards the difficulty of statistics was greater than before, which means that students perceived statistics to be more difficult at the end of the course compared to when they started it. This is connected to students' attitudes towards effort, because students also felt that they put less effort in learning statistics at the end of the course. Students' decrease in the effort they put into learning statistics at the end of the course might also very well influence why they perceived the subject more difficult at the end of the course.

Between the two treatments, no significant difference was found in changes in students' attitudes. In this study, the benefits of using collaborative tests on students' attitudes towards statistics are therefore not related to the specific format of the collaborative tests. However, in both treatments there was a noticeable increase in students' attitudes in terms of their intellectual knowledge, skills, and interest towards statistics. It is unclear what caused this increase and because a third treatment without collaborative tests was not included we cannot determine that the increase was due to the three collaborative tests.

Research Question 3. How does using a required consensus on collaborative tests vs. a nonconsensus approach affect group discussions?

When students' discussions on the three collaborative tests were explored in terms of the proportion of discussion comments that were classified as Cognitive, Social and Teaching, the only significant difference found between the two treatments was that students in the Nonconsensus section had more discussion comments classified as Teaching. This was surprising because of the way the collaborative tests were set up in the Consensus section, when students had to reach a consensus on the tests and turn them

in as a group. That format was expected to affect the discussion on the tests in a way that students might have discussed more practical matters such as when and how to review and turn the test in. Logistical matters like these are classified as *Teaching* dimension according to the Pozzi et al., (2007) framework. This was still the case though because, when indicators for the three dimensions were explored, the only significant difference found between the two treatments was for the Teaching dimensions indicator *Organisational matters*, which was more present in the Consensus section. The *Organisational matters* indicator is about introducing topics, providing explanations for methods and letting students know of deadlines, this difference does not come as a surprise because students in the Consensus section had to compile and turn the collaborative tests in together. There was more at stake in terms of logistics in turning the test in for students in the Consensus section compared to their peers in the Nonconsensus section who turned their tests in individually. The *Organisational matters* indicator was the least common indicator in the Nonconsensus section, but the other two indicators for the Teaching dimension *Direct instruction* and *Facilitation* were more common in the Nonconsensus section compared to the Consensus section.

Without the Teaching dimension, the effects of using two different formats of collaborative tests on group discussions seem to be similar because no other significant differences were found in group discussions for students using a required Consensus compared to students using a Nonconsensus approach on the collaborative tests. In both treatments, discussion comments classified as Cognitive were the most common for the three collaborative tests. This does not come as a surprise because students were working on an assessment together that was graded among other things on the number of correct

items. It makes sense that in both treatments most discussion comments were focused towards discussing the tests and their contents. Discussion comments classified as Social were the second most common after the Cognitive dimension.

Limitations of the Study

The study had limitations that affect the conclusions drawn from the results. One limitation relates to the CAOS test, which was used to measure important student learning outcomes. CAOS contains 40 items to assess learning outcomes for all students completing a basic introductory statistics course. In this study, only 33 items out of the 40 items on CAOS were used in the pre and posttests. Not using all the items on the CAOS might have affected the reliability and validity of the measurements obtained using this instrument.

Generalization of the results of the study to other introductory statistics courses may be limited to courses that also use collaborative learning as a regular method of teaching. Implementing only collaborative tests without including other collaborative activities might produce different results. Furthermore, adding one treatment group that would not receive collaborative tests might have provided useful information on the effect of using collaborative tests. Conclusions based on this study would be stronger if the experiment had been repeated over another semester to see if the same results emerged.

Students' familiarity with the online environment might have influenced the results. Because even though students were randomly assigned between treatments, students in the Nonconsensus section had more experience in taking online course before taking this course compared to students in the Consensus section.

Implications for Teaching Online Statistics

While this study did not show a significant difference between the two collaborative test formats, students reported a positive perspective towards the collaborative tests. A majority of students in both treatments preferred to take collaborative tests. Considering what has been pointed out in the literature (Garfield & Ben-Zvi, 2008) about the negative views students hold about statistics courses being both difficult and an unpleasant experiences, any effort that challenges this long standing idea should be carefully constructed. Instructors of statistics should explore the use of collaborative tests in their online introductory courses. Based on the results here the effects of the different formats of collaborative tests remain unclear. The increase in students' attitudes in terms of their intellectual knowledge, skills, and interest towards statistics might have been because of students' experience working on collaborative tests or because students worked in collaborative groups throughout the semester; it is hard to know without including a third treatment where there would be no collaboration on tests.

Instructors who want to use collaborative tests in online statistics courses should do so, and if they can, they should include a control group, where there is no collaboration. They should try different collaborative formats that fit the course, for example by offering students the choice between two or more different formats. Or to offer different variations on the collaborative test to cater to different students' learning needs. Instructors might keep in mind that in this study, discussion comments classified as Teaching were more common in the Nonconsensus section, and the negative effects of discussion comments classified as Social on the Midterm exam score were larger in the Consensus section.

Instructors might also want to keep in mind fairness in grading and how much of the overall grade the collaborative tests would account for. In this study, in order to ensure that every student would participate in the collaborative test, the grading was based on correctness and participation on the test. The collaborative tests accounted for 20% of the final grade, while individual assessment and assignments accounted for 58.6%.

This study suggests that the use of collaborative tests in online introductory statistics courses has a positive impact on students' attitudes toward statistics. These results support the use of collaborative activities and assessments in online introductory statistics courses as well as face to face classes.

Implications for Future Research

With the expected increase in enrollment in online course in the coming years (Allen & Seaman, 2010), the need to conduct research on the online introductory statistics course is warranted. More and more courses, including statistics, will be offered online and the need for effective quality teaching methods and assessment in these courses will increase as well. Many questions remain unanswered when it comes to teaching statistics online and using collaborative tests in online courses in general. Those include the effects of using different test formats, the appropriate group size, type of questions used, etc.

The online environment offers new ways in exploring both how statistics is taught online and how collaborative tests are used. Applying some of the principles of experimental design like randomizing to groups is more applicable to do in online courses compared to face-to-face courses due to schedule conflict. For example, in this

study, the effects of confounding variables were minimized, like time of class and instructor by having the same instructor teach both courses online.

Some research has been done concerning the teaching of introductory statistics courses in face-to-face settings, one of the products of this work are the Guidelines for Assessment and Instruction on Statistics Education (GAISE). The GAISE recommendations among other things, suggest the use of active learning and having students work in groups however it is still unclear if these recommendations also apply to the online environment. Very little research exists on the online introductory statistics courses and the assessment and instruction of these courses remain uncertain. Even though this study did not establish much significant difference between the two collaborative formats that were used it does contribute significantly to the scarce literature on online introductory statistics courses. The results of the study suggest that there are benefits of using collaborative learning and testing in online introductory statistics courses, students had positive perspective towards taking collaborative tests and majority of students preferred to take collaborative tests. There was also an increase in students' attitudes in terms of their intellectual knowledge, skills, and interest towards statistics but it is unclear if that was due to the collaborative tests like prior research has shown. Because of this, more research in regards to the online introductory statistics courses is needed to support these findings and to help us better understand how we can structure and teach high quality online introductory statistics courses.

References

- Agresti, A., and Finlay, B. (1997), *Statistical Methods for the Social Sciences*, 3rd ed., Upper Saddle River, NJ: Prentice Hall.
- AIMS Project. (NA). AIMS Project Adapting and Implementing Innovative Material in Statistics. Retrieved May 15, 2011 from <http://www.tc.umn.edu/~aims/>
- Allen, I. E., & Seaman, J. (2010). *Learning on demand: Online education in the United States, 2009*. Retrieved from Sloan Consortium, Babson Survey Research Group website:
<http://sloanconsortium.org/publications/survey/pdf/learningondemand.pdf>
- Bakker, J. R. (2009). *Web-based vs. classroom instruction of statistics*. (Unpublished PhD). The Ohio State University, 2009.
- Ben-Zvi, D. (2007). Using wiki to promote collaborative learning in statistics education. *Technology Innovations in Statistics Education*, 1(1).
- Breedlove, W., Burkett, T., & Winfield, I. (2004). *Collaborative testing and test performance*. Retrieved from
<http://find.galegroup.com.floyd.lib.umn.edu/gtx/infomark.do?&contentSet=IAC-Documents&type=retrieve&tabID=T002&prodId=PROF&docId=A126683337&source=gale&srcprod=PROF&userGroupName=mnaumntwin&version=1.0>
- Cobb, P., & McClain, K. (2004). Propose design principles for the teaching and learning of elementary statistics. In D. Ben-Zvi, & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 375-396). Dordrecht, Netherlands: Kluwer.

- Curtis, D. D., & Lawson, M. J. (2001). Exploring collaborative online learning. *Journal of Asynchronous Learning Networks*, 5(1).
- Dabbagh, N., & Bannan-Ritland, B. (2005). *Online learning: Concepts, strategies, and application*. Columbus, OH: Pearson Merrill Prentice Hall.
- delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, 6(2). Retrieved from [http://www.stat.auckland.ac.nz/~iase/serj/SERJ6\(2\)_delMas.pdf](http://www.stat.auckland.ac.nz/~iase/serj/SERJ6(2)_delMas.pdf)
- Delucchi, M. (2006). The efficacy of collaborative learning groups in an undergraduate statistics course. *College Teaching*, 54(2), 244-248.
- DeVaney, T. A. (2010). Anxiety and attitude of graduate students in on-campus vs. online statistics courses. *Journal of Statistics Education*, 18(1).
- Dillenbourg, P. (1999). What do you mean by collaborative learning? In P. Dillenbourg (Ed.), *Collaborative-learning: Cognitive and computational Approaches*. (pp. 1-19). Oxford, UK: Elsevier.
- Dutton, J., & Dutton, M. (2005). Characteristics and performance of students in an online section of business statistics. *Journal of Statistics Education*, 13(3).
- Eggen, P., & Kauchak, D. (2006). *Educational psychology: Windows on classrooms* (7th ed.). New Jersey, NJ: Pearson Merrill Prentice Hall.
- Everson, M. (2006). Group discussion in online statistics courses. *ELearn Magazine*, 2006(4).

- Everson, M. G., & Garfield, J. (2008). An innovative approach to teaching online statistics courses. *Technology Innovations in Statistics Education*, 2(1).
- Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). London, UK: Sage.
- GAISE. (2005). *GAISE college report*. Retrieved from <http://www.amstat.org/education/gaise/GAISECollege.htm>
- Garfield, J. & Franklin, C. (2011). Assessment of learning, for learning, and as learning in statistics education. In C. Batanero, G. Burrill, C. Reading and A. Rossman (eds.), *Teaching statistics in school mathematics—Challenges for teaching and teacher education: A joint ICMI/IASE study* (pp. 133–145). New York, NY: Springer.
- Garfield, J., Zieffler, A., Kaplan, D., Cobb, G., Chance, B., & Holcomb, J. (2011). Rethinking assessment of student learning in statistics courses. *The American Statistician*, 65(1), 1.
- Garfield, J., delMas, R., & Zieffler, A. (2008, June). *AIMS: Adapting and implementing innovative materials*. A CAUSEway Workshop presented at the University of Minnesota, Minneapolis, MN.
- Garfield, J. & Ben-Zvi, D. (2008). *Developing Students' Statistical Reasoning: Connecting Research and Teaching Practice*. Dordrecht, the Netherlands: Springer.
- Garfield, J. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review*, 75(3), 372.

- Garfield, J., Hogg, B., Schau, C., & Whittinghill, D. (2002). First courses in statistical science: The status of educational reform efforts. *Journal of Statistics Education*, 10(2), 456.
- Garfield, J. (1995). How students learn statistics. *International Statistical Review*, 63(1), 25.
- Garfield, J. (1993). Teaching statistics using small-group cooperative learning. *Journal of Statistics Education*, 1(1)
- Garrison, D. R., & Anderson, T. (2003). *E-learning in the 21st century a framework for research and practice*. London, UK: RoutledgeFalmer.
- Giraud, G. (1997). Cooperative learning and statistics instruction. *Journal of Statistics Education*, 5(3), 1.
- Giraud, G., & Enders, C. (2000). *The effects of repeated cooperative testing in an introductory statistics course*. Retrieved from <http://search.ebscohost.com.floyd.lib.umn.edu/login.aspx?direct=true&db=eric&AN=ED445103&site=ehost-live>
- Giuliodori, M., Lujan, H. L., & DiCarlo, S. E. (2009). Student interaction characteristics during collaborative group testing. *Advances in Physiology Education*, 33(1), 24-29.
- 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.
- Gunnarsson, C. L. (2001). *Student attitude and achievement in an online graduate statistics course*. (Unpublished EdD). University of Cincinnati, 2001,

- Haberyan, A., & Barnett, J. (2010). Collaborative testing and achievement: Are two heads really better than one? *Journal of Instructional Psychology*, 37(1), 32-41.
- Harasim, L. (2000). Shift happens: Online education as a new paradigm in learning. *The Internet and Higher Education*, 3(1-2), 41.
- Helmericks, S. G. (1993). Collaborative testing in social statistics: Toward gemeinstat. *Teaching Sociology*, 21, 287-297.
- Henri, F. (1992). Computer conferencing and content analysis. In A. R. Kaye (Ed.), *In collaborative learning through computer conferencing: The najaden papers* (pp. 115-136). New York, NY: Springer.
- Hicks, J. (2007). Students' views of cooperative learning and group testing. *Radiologic Technology*, 78(4), 275.
- Hong, K., Lai, K., & Holton, D. (2003). Students' satisfaction and perceived learning with a web-based course. *Educational Technology & Society*, 6(1).
- Howell, D.C. (2007). *Statistical Methods for Psychology* (6th ed). Belmont, CA: Duxbury Press.
- Ioannou, A., & Artino, A. R. J. (2010). Learn more, stress less: Exploring the benefits of collaborative assessment. *College Student Journal*, 44(1).
- Jianxia, D., Durrington, V. A., & Mathews, J. G. (2007). Online collaborative discussion: Myth or valuable learning tool. *Journal of Online Learning and Teaching*, 3(2), 94.
- Johnson, D. W., Johnson, R. T., & Holubec, E. J. (2008). *Cooperation in the classroom* (Eighth edition ed.). Edina, MN: Interaction Book.

- Johnson, D. W., Johnson, R. T., & Stanne, M. (2000). *Cooperative learning methods: A meta-analysis*. Retrieved December 1, 2010, from <http://www.clcrc.com/pages/cl-methods.html>
- Johnson, D. W., & Johnson, R. T. (1996). Cooperation and the use of technology. In D. Jonassen (Ed.), *Handbook of research for educational communications and technology* (pp. 785). London, UK: MacMillan.
- Johnson, D. W., Johnson, R. T., & Smith, K. A. (1991). *Active learning: Cooperation in the college classroom*. Edina, MN: Interaction Book.
- Kapitanoff, S. H. (2009). Collaborative testing: Cognitive and interpersonal processes related to enhanced test performance. *Active Learning in Higher Education*, 10(1), 56-70. doi:10.1177/1469787408100195
- Keller, C. M., & Steinhorst, R. K. (1995). Using small groups to promote active learning in the introductory statistics course: A report from the field. *Journal of Statistics Education*, 3(2).
- Kieser, A. L., & Golden, F. O. (2009). *Using online office applications: Collaboration tools for learning*. Retrieved from <http://find.galegroup.com.floyd.lib.umn.edu/gtx/infomark.do?&contentSet=IAC-Documents&type=retrieve&tabID=T002&prodId=PROF&docId=A234310653&source=gale&srcprod=PROF&userGroupName=mnaumntwin&version=1.0>
- Lewis-Beck, M. 1980. *Applied Regression: An introduction*. Sage Series on Quantitative Applications in the Social Sciences Nr. 22
- Lusk, M., & Conklin, L. (2003). Collaborative testing to promote learning. *Journal of Nursing Education*, 42(3), 121-124.

- Magel, R. C. (1998). Using cooperative learning in a large introductory statistics class. *Journal of Statistics Education*, 6(3).
- Mahle, M. (2007). Interactivity in distance education. *Distance Learning*, 4(1), 47.
- Manca, S., Persico, D., Pozzi, F., & Sarti, L. (NA). *A model to monitor and evaluate online collaborative learning processes*. Retrieved December 10, 2010, from <http://spaziofir.itd.cnr.it/CSCL/modello.htm>
- Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review*, 65, 123–137.
- Palloff, R., & Pratt, K. (2004). *Collaborating online: Learning together in community*. San Francisco, CA: Jossey-Bass.
- Palloff, R., & Pratt, K. (2007). *Building online learning communities: Effective strategies for the virtual classroom* (2nd ed.). San Francisco, CA: Jossey-Bass.
- Perkins, D. V., & Saris, R. N. (2001). A "jigsaw classroom" technique for undergraduate statistics courses. *Teaching of Psychology*, 28(2).
- Persico, D., Pozzi, F., & Sarti, L. (2010). Monitoring collaborative activities in computer supported collaborative learning. *Distance Education*, 31(1), 5-22
- Potthast, M. J. (1999). Outcomes of using small-group cooperative learning experiences in introductory statistics courses. *College Student Journal*, 33(1)
- Pozzi, F. (2010). Using jigsaw and case study for supporting online collaborative learning. *Computers & Education*, 55(1), 67-75.
- Pozzi, F., Manca, S., Persico, D., & Sarti, L. (2007). A general framework for tracking and analysing learning processes in computer-supported collaborative learning

- environments. *Innovations in Education & Teaching International*, 44(2), 169-179.
- Rao, S. P., Collins, H. L., & DiCarlo, S. E. (2002). Collaborative testing enhances student learning. *Advances in Physiology Education*, 26(1), 37-41.
- Resta, P., & Laferrière, T. (2007). Technology in support of collaborative learning. *Educational Psychology Review*, 19(1), 65.
- Roberts, T. S. (Ed.). (2004). *Online collaborative learning: Theory and practice*. Hershey, PA: Information Science.
- Roseth, C. J., Johnson, D. W., & Johnson, R. T. (2008). Promoting early adolescents' achievement and peer relationships: The effects of cooperative, competitive, and individualistic goal structures. *Psychological Bulletin*, 134(2), 223.
- Roseth, C. J., Garfield, J. B., & Ben-Zvi, D. (2008). Collaboration in learning and teaching statistics. *Journal of Statistics Education*, 16(1). Retrieved from <http://www.amstat.org/publications/jse/v16n1/roseth.pdf>
- Sandahl, S. S. (2009). Collaborative testing as a learning strategy in nursing education: A review of the literature. *Nursing Education Perspectives*, 30(3), 171-175.
- Schou, S. B. (2007). A study of student attitudes and performance in an online introductory business statistics class. *Electronic Journal for the Integration of Technology in Education*, 6, 71-78.
- Schau, C. (2005). CS Consultants, LLC website. Retrieved May 1, 2011 from <http://www.evaluationandstatistics.com>

- Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 957-1009). Reston, VA: National Council of Teachers of Mathematics.
- Shen, J., Hiltz, S. R., & Bieber, M. (2006). Collaborative online examinations: Impacts on interaction, learning, and student satisfaction. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 36(6), 1045-1053.
- Shen, J., Hiltz, S. R., & Bieber, M. (2008). Learning strategies in online collaborative examinations. *Professional Communication, IEEE Transactions on*, 51(1), 63-78.
- Shindler, J. V. (2004). Greater than the sum of the parts: Examining the soundness of collaborative exams in teachers education courses. *Innovative Higher Education*, 22(4), 273-283.
- Simkin, M. G. (2005). An experimental study of the effectiveness of collaborative testing in an entry-level computer programming class. *Journal of Information Systems Education*, 16(3), 273.
- Simonson, M. R., Smaldino, S., Albright, M., & Zvacek, S. (2000). *Teaching and learning at a distance* (Fourth ed.). Boston, MA: Allyn & Bacon.
- Slavin, R. E. (1991). Synthesis of research on cooperative learning. *Educational Leadership*, 48(5), 71-81. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=9108121689&site=ehost-live>
- Stahl, G., Koschmann, T., & Suthers, D. (2006). Computer-supported collaborative learning: An historical perspective. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences*. Cambridge, UK: Cambridge University Press.

- Swan, K., Shen, J., & Hiltz, S. R. (2006). Assessment and collaboration in online learning. *Journal of Asynchronous Learning Networks, 10*(1).
- Tallent-Runnels, M. K., Thomas, J. A., Lan, W. Y., Cooper, S., Ahern, T. C., Shaw, S. M., & Liu, X. (2006). Teaching courses online: A review of the research. *Review of Educational Research, 76*(1), 93-135.
- Tempelaar, D., Van Der Loeff, S., and Gijsselaers, W. (2007). A structural equation model analyzing the relationship of students' attitudes toward statistics, prior reasoning abilities and course performance. *Statistics Education Research Journal, 6*(2).
- Tudor, G. E. (2006). Teaching introductory statistics online- satisfying the students. *Journal of Statistics Education, 14*(3).
- Utts, J., Sommer, B., Acredolo, C., Maher, M. W., & Matthews, H. R. (2003). A study comparing traditional and hybrid internet-based instruction in introductory statistics classes. *Journal of Statistics Education, 11*(3).
- Ward, B. (2004). The best of both worlds: A hybrid statistics course. *Journal of Statistics Education, 12*(3).
- Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education, 46*(1), 71-95.
- Wenger, E. (2006). Communities of practice: A brief introduction. Retrieved May 20, 2011 from <http://www.ewenger.com/theory/index.htm>
- Wisnbaker, J. (2003). Extending the journey toward a virtual introductory statistics course. *Paper Presented at the Meeting of the International Association for Statistical Education Conference "Statistics and the Internet"*, Berlin, Germany.

Zhang, J. (2002). Teaching statistics on-line: Our experience and thoughts. *Paper Presented at the Annual Meeting of the International Conference on Teaching Statistics*, Cape Town, South Africa.

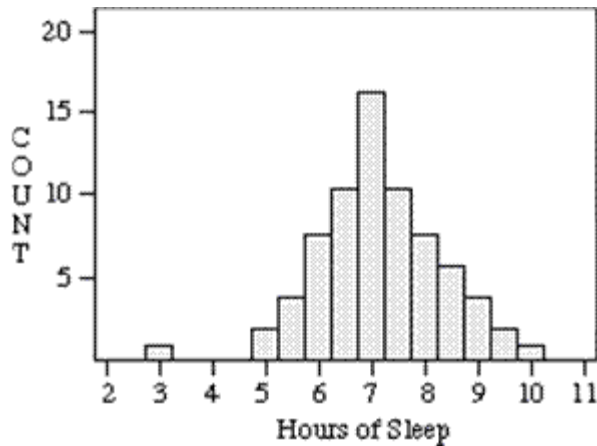
Zimbardo, P. G., Butler, L. D., & Wolfe, V. A. (2003). Cooperative college examinations: More gain, less pain when students share information and grades. *Journal of Experimental Education*, 71(2), 101.

Appendix A

Instruments

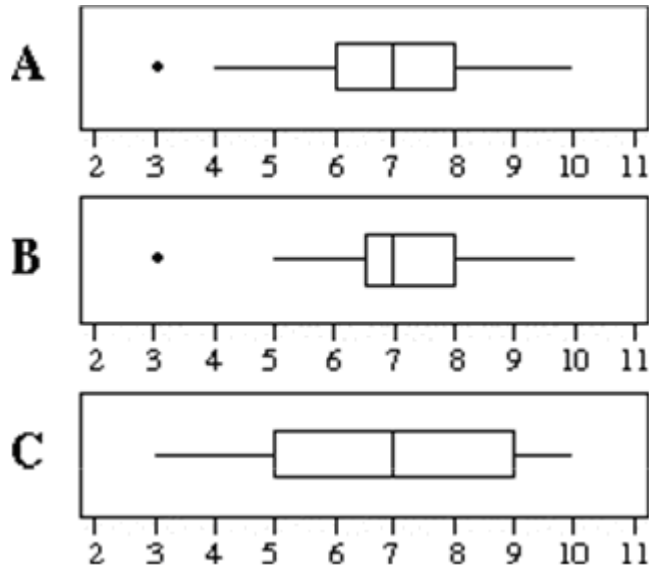
A-1 Comprehensive Assessment of Important Outcomes in Statistics (CAOS)

The following graph shows a distribution of hours slept last night by a group of college students.



1. Select the statement below that gives the most complete description of the graph in a way that demonstrates an understanding of how to statistically describe and interpret the distribution of a variable.
 - a. The bars go from 3 to 10, increasing in height to 7, then decreasing to 10. The tallest bar is at 7. There is a gap between three and five.
 - b. The distribution is normal, with a mean of about 7 and a standard deviation of about 1.
 - c. Most students seem to be getting enough sleep at night, but some students slept more and some slept less. However, one student must have stayed up very late and got very few hours of sleep.
 - d. The distribution of hours of sleep is somewhat symmetric and bell-shaped, with an outlier at 3. The typical amount of sleep is about 7 hours and overall range is 7 hours.

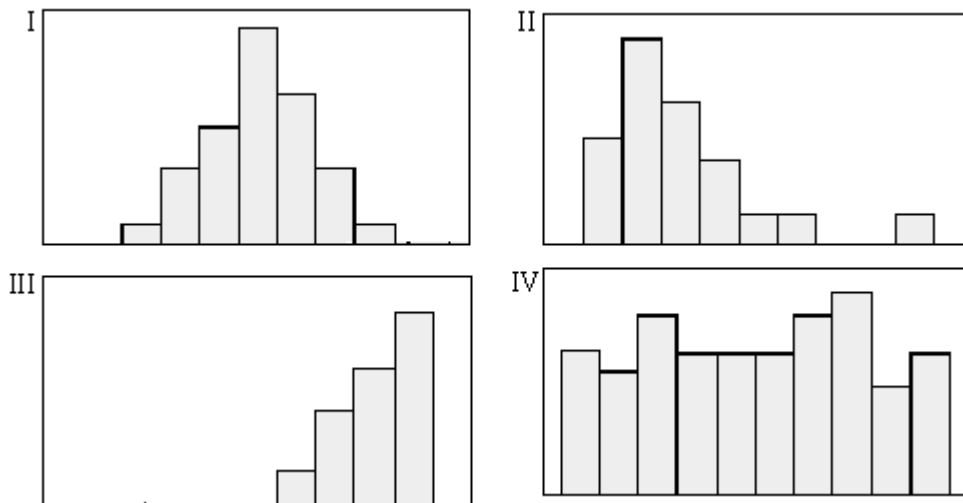
2. Which box plot seems to be graphing the same data as the histogram in question 1?



- a. Boxplot A.
- b. Boxplot B.
- c. Boxplot C.

Items 3 to 5 refer to the following situation:

Four histograms are displayed below. For each item, match the description to the appropriate histogram.



3. A distribution for a set of quiz scores where the quiz was very easy is represented by:
 - a. Histogram I.
 - b. Histogram II.
 - c. Histogram III.
 - d. Histogram IV.

4. A distribution for a set of wrist circumferences (measured in centimeters) taken from the right wrist of a random sample of newborn female infants is represented by:
 - a. Histogram I.
 - b. Histogram II.
 - c. Histogram III.
 - d. Histogram IV.

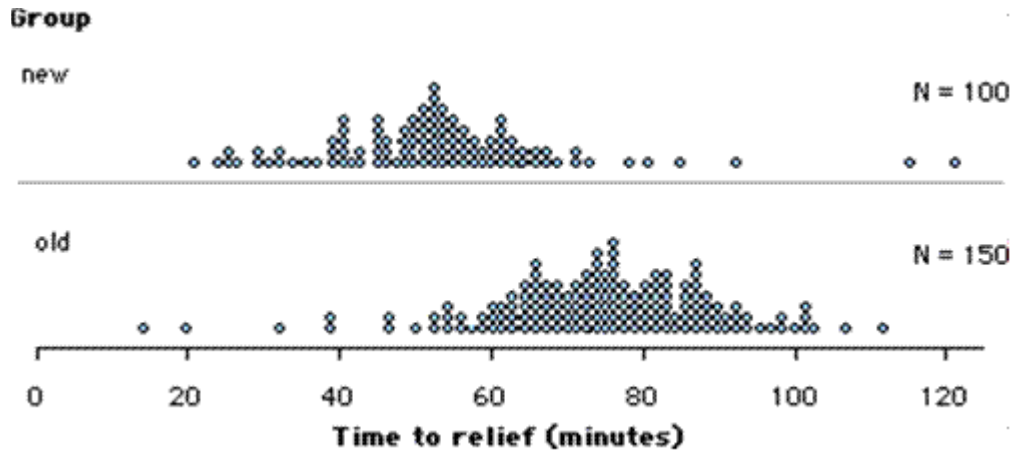
5. A distribution for the last digit of phone numbers sampled from a phone book (i.e., for the phone number 968-9667, the last digit, 7, would be selected) is represented by:
 - a. Histogram I.
 - b. Histogram II.
 - c. Histogram III.
 - d. Histogram IV.

6. A recent research study randomly divided participants into groups who were given different levels of Vitamin E to take daily. One group received only a placebo pill. The research study followed the participants for eight years to see how many developed a particular type of cancer during that time period. Which of the following responses gives the best explanation as to the purpose of randomization in this study?
 - a. To increase the accuracy of the research results.
 - b. To ensure that all potential cancer patients had an equal chance of being selected for the study.
 - c. To reduce the amount of sampling error.
 - d. To produce treatment groups with similar characteristics.
 - e. To prevent skewness in the results.

Items 7 to 9 refer to the following situation:

A drug company developed a new formula for their headache medication. To test the effectiveness of this new formula, 250 people were randomly selected from a larger population of patients with headaches. 100 of these people were randomly assigned to receive the new formula medication when they had a headache, and the other 150 people

received the old formula medication. The time it took, in minutes, for each patient to no longer have a headache was recorded. The results from both of these clinical trials are shown below. Items 11, 12, and 13 present statements made by three different statistics students. For each statement, indicate whether you think the student's conclusion is valid.



7. The old formula works better. Two people who took the old formula felt relief in less than 20 minutes, compared to none who took the new formula. Also, the worst result - near 120 minutes - was with the new formula.
 - a. Valid.
 - b. Not valid.

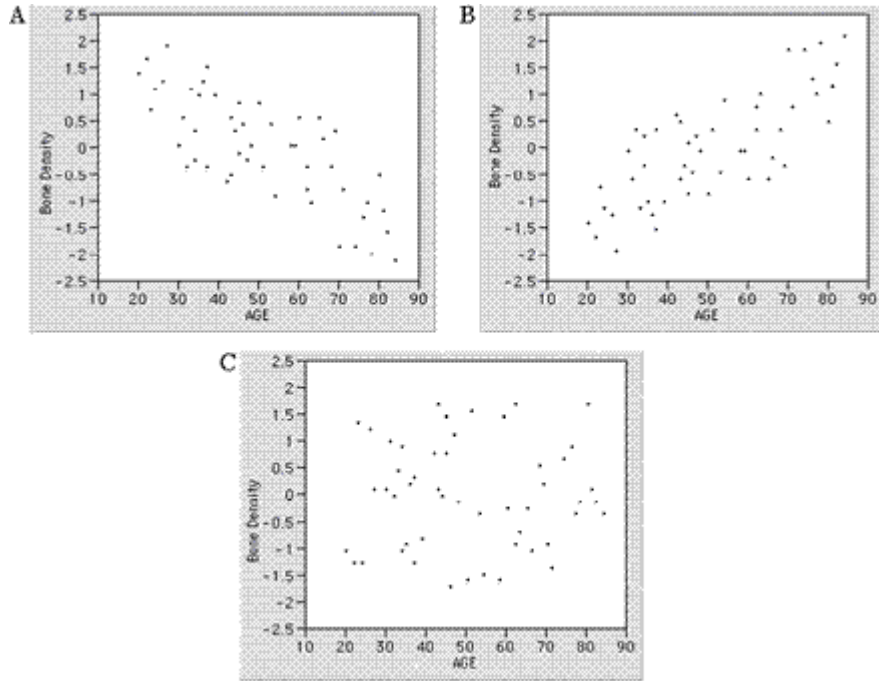
8. The average time for the new formula to relieve a headache is lower than the average time for the old formula. I would conclude that people taking the new formula will tend to feel relief about 20 minutes sooner than those taking the old formula.
 - a. Valid.
 - b. Not valid.

9. I would not conclude anything from these data. The number of patients in the two groups is not the same so there is no fair way to compare the two formulas.
 - a. Valid.
 - b. Not valid.

10. A certain manufacturer claims that they produce 50% brown candies. Sam plans to buy a large family size bag of these candies and Kerry plans to buy a small fun size bag. Which bag is more likely to have more than 70% brown candies?
 - a. Sam, because there are more candies, so his bag can have more brown candies.

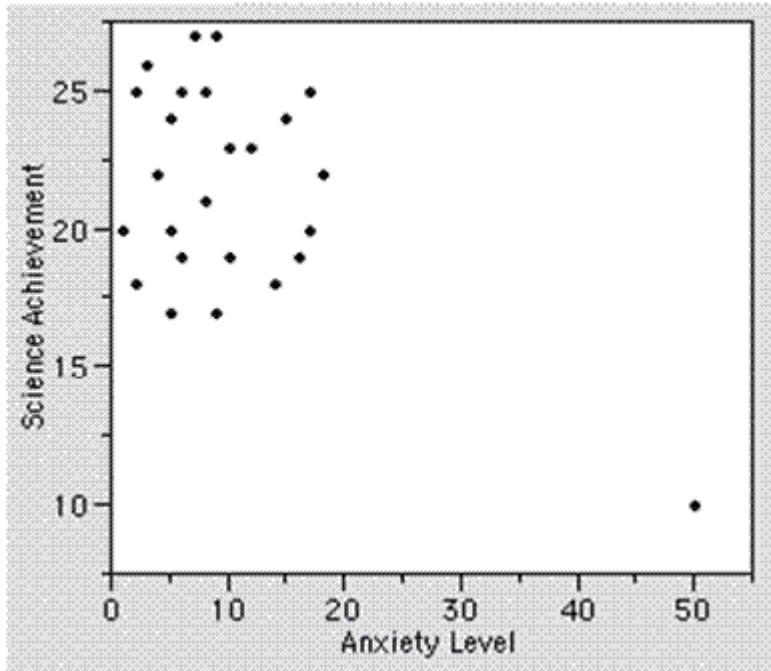
- b. Sam, because there is more variability in the proportion of browns among larger samples.
 - c. Kerry, because there is more variability in the proportion of browns among smaller samples.
 - d. Kerry, because most small bags will have more than 50% brown candies.
 - e. Both have the same chance because they are both random samples.
11. Imagine you have a barrel that contains thousands of candies with several different colors. We know that the manufacturer produces 35% yellow candies. Five students each take a random sample of 20 candies, one at a time, and record the percentage of yellow candies in their sample. Which sequence below is the most plausible for the percent of yellow candies obtained in these five samples?
- a. 30%, 35%, 15%, 40%, 50%.
 - b. 35%, 35%, 35%, 35%, 35%.
 - c. 5%, 60%, 10%, 50%, 95%.
 - d. Any of the above.
12. A graduate student is designing a research study. She is hoping to show that the results of an experiment are statistically significant. What type of p -value would she want to obtain?
- a. A large p -value.
 - b. A small p -value.
 - c. The magnitude of a p -value has no impact on statistical significance.

13. Bone density is typically measured as a standardized score with a mean of 0 and a standard deviation of 1. Lower scores correspond to lower bone density. Which of the following graphs shows that as women grow older they tend to have lower bone density?



- a. Graph A.
- b. Graph B.
- c. Graph C.

14. The following scatterplot shows the relationship between scores on an anxiety scale and an achievement test for science. Choose the best interpretation of the relationship between anxiety level and science achievement based on the scatterplot.



- a. This graph shows a strong negative linear relationship between anxiety and achievement in science.
- b. This graph shows a moderate linear relationship between anxiety and achievement in science.
- c. This graph shows very little, if any, linear relationship between anxiety and achievement in science.
15. Researchers surveyed 1,000 randomly selected adults in the U.S. A statistically significant, strong positive correlation was found between income level and the number of containers of recycling they typically collect in a week. Please select the best interpretation of this result.
- a. We can not conclude whether earning more money causes more recycling among U.S. adults because this type of design does not allow us to infer causation.
- b. This sample is too small to draw any conclusions about the relationship between income level and amount of recycling for adults in the U.S.
- c. This result indicates that earning more money influences people to recycle more than people who earn less money.

Items 16 and 17 refer to the following situation:

A researcher in environmental science is conducting a study to investigate the impact of a particular herbicide on fish. He has 60 healthy fish and randomly assigns each fish to either a treatment or a control group. The fish in the treatment group showed higher levels of the indicator enzyme.

16. Suppose a test of significance was correctly conducted and showed no statistically significant difference in average enzyme level between the fish that were exposed to the herbicide and those that were not. What conclusion can the graduate student draw from these results?
 - a. The researcher must not be interpreting the results correctly; there should be a significant difference.
 - b. The sample size may be too small to detect a statistically significant difference.
 - c. It must be true that the herbicide does not cause higher levels of the enzyme.

17. Suppose a test of significance was correctly conducted and showed a statistically significant difference in average enzyme level between the fish that were exposed to the herbicide and those that were not. What conclusion can the graduate student draw from these results?
 - a. There is evidence of association, but no causal effect of herbicide on enzyme levels.
 - b. The sample size is too small to draw a valid conclusion.
 - c. He has proven that the herbicide causes higher levels of the enzyme.
 - d. There is evidence that the herbicide causes higher levels of the enzyme for these fish.

Items 18 to 20 refer to the following situation:

A research article reports the results of a new drug test. The drug is to be used to decrease vision loss in people with Macular Degeneration. The article gives a p -value of .04 in the analysis section. Items 25, 26, and 27 present three different interpretations of this p -value. Indicate if each interpretation is valid or invalid.

18. The probability of getting results as extreme as or more extreme than the ones in this study if the drug is actually not effective.
 - a. Valid.
 - b. Invalid.
19. The probability that the drug is not effective.
 - a. Valid.
 - b. Invalid.
20. The probability that the drug is effective.
 - a. Valid.
 - b. Invalid.

Items 21 to 24 refer to the following situation:

A high school statistics class wants to estimate the average number of chocolate chips in a generic brand of chocolate chip cookies. They collect a random sample of cookies, count the chips in each cookie, and calculate a 95% confidence interval for the average number of chips per cookie (18.6 to 21.3). Items 28, 29, and 30 present four different interpretations of these results. Indicate if each interpretation is valid or invalid.

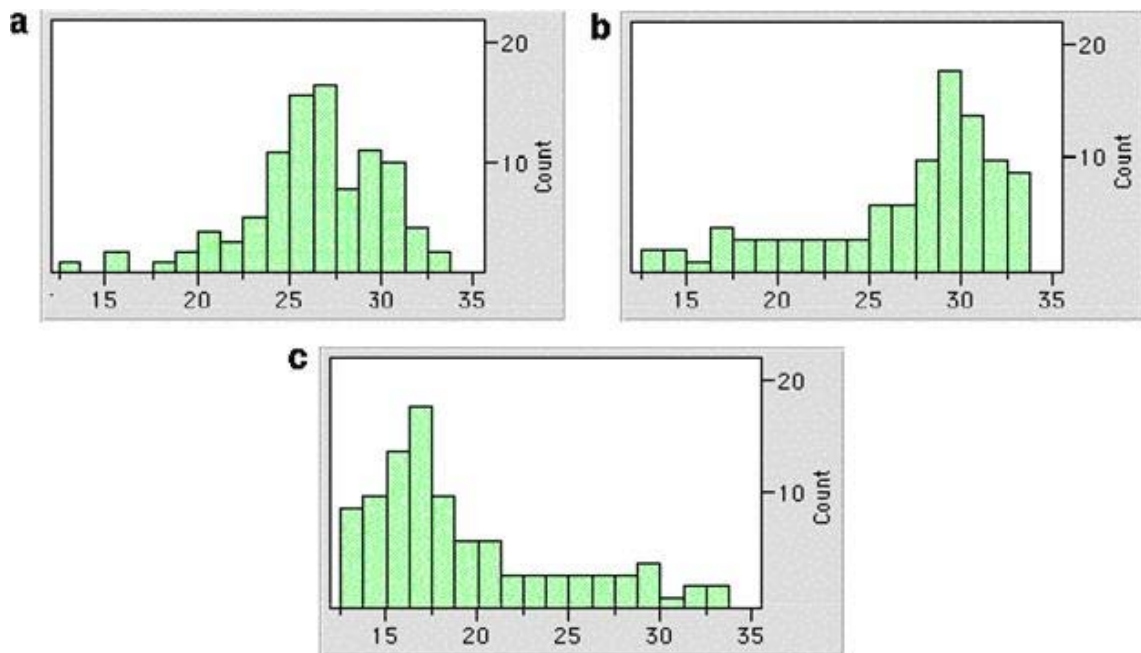
21. We are 95% certain that each cookie for this brand has approximately 18.6 to 21.3 chocolate chips.
 - a. Valid.
 - b. Invalid.
22. We expect 95% of the cookies to have between 18.6 and 21.3 chocolate chips.
 - a. Valid.
 - b. Invalid.
23. We would expect about 95% of all possible sample means from this population to be between 18.6 and 21.3 chocolate chips.
 - a. Valid.
 - b. Invalid.

24. We are 95% certain that the confidence interval of 18.6 to 21.3 includes the true average number of chocolate chips per cookie.
- Valid.
 - Invalid.
25. It has been established that under normal environmental conditions, adult largemouth bass in Silver Lake have an average length of 12.3 inches with a standard deviation of 3 inches. People who have been fishing Silver Lake for some time claim that this year they are catching smaller than usual largemouth bass. A research group from the Department of Natural Resources took a random sample of 100 adult largemouth bass from Silver Lake and found the mean of this sample to be 11.2 inches. Which of the following is the most appropriate statistical conclusion?
- The researchers cannot conclude that the fish are smaller than what is normal because 11.2 inches is less than one standard deviation from the established mean (12.3 inches) for this species.
 - The researchers can conclude that the fish are smaller than what is normal because the sample mean should be almost identical to the population mean with a large sample of 100 fish.
 - The researchers can conclude that the fish are smaller than what is normal because the difference between 12.3 inches and 11.2 inches is much larger than the expected sampling error.

A study examined the length of a certain species of fish from one lake. The plan was to take a random sample of 100 fish and examine the results. Numerical summaries on lengths of the fish measured in this study are given.

Mean	26.8mm
Median	29.4mm
Standard Deviation	5.0 mm
Minimum	12.mm
Maximum	33.4mm

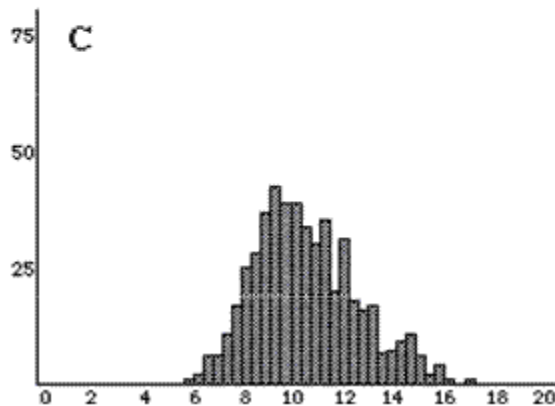
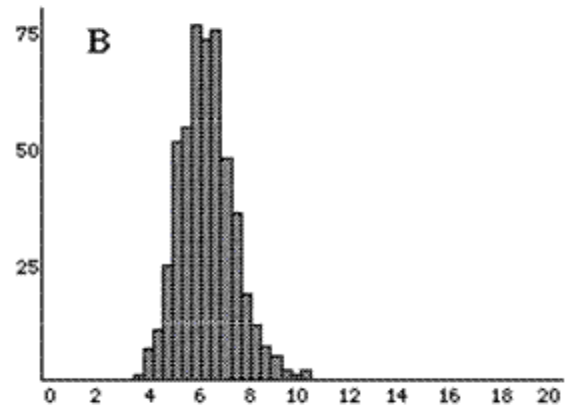
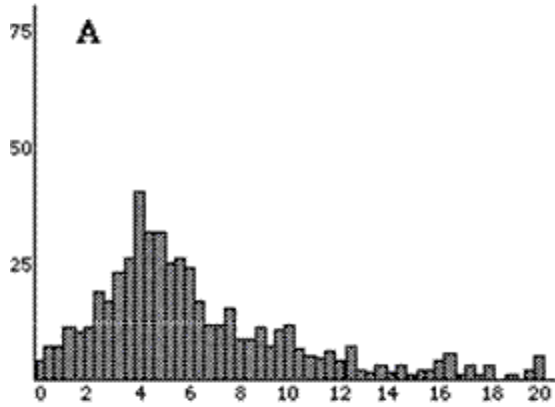
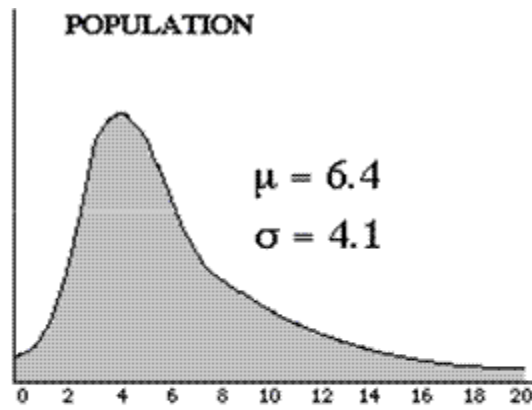
26. Which of the following histograms is most likely to be the one for these data?



- a. Histogram a.
- b. Histogram b.
- c. Histogram c.

Items 27 and 28 refer to the following situation:

Four graphs are presented below. The graph at the top is a distribution for a population of test scores. The mean score is 6.4 and the standard deviation is 4.1.



27. Which graph (A, B, or C) do you think represents a single random sample of 500 values from this population?
- Graph A
 - Graph B
 - Graph C
28. Which graph (A, B, or C) do you think represents a distribution of 500 sample means from random samples each of size 9?
- Graph A
 - Graph B
 - Graph C
29. This table is based on records of accidents compiled by a State Highway Safety and Motor Vehicles Office. The Office wants to decide if people are less likely to have a fatal accident if they are wearing a seatbelt. Which of the following comparisons is most appropriate for supporting this conclusion?

Safety Equipment in Use	Injury		ROW TOTAL
	Nonfatal	Fatal	
Seat Belt	412,368	510	412,878
No Seat Belt	162,527	1,601	164,128
COLUMN TOTAL	574,895	2,111	577,006

- Compare the ratios $510/412,878$ and $1,601/164,128$
- Compare the ratios $510/577,006$ and $1,601/577,006$
- Compare the numbers 510 and 1,601

30. A student participates in a Coke versus Pepsi taste test. She correctly identifies which soda is which four times out of six tries. She claims that this proves that she can reliably tell the difference between the two soft drinks. You have studied statistics and you want to determine the probability of anyone getting at least four right out of six tries just by chance alone. Which of the following would provide an accurate estimate of that probability?
- Have the student repeat this experiment many times and calculate the percentage time she correctly distinguishes between the brands.
 - Simulate this on the computer with a 50% chance of guessing the correct soft drink on each try, and calculate the percent of times there are four or more correct guesses out of six trials.
 - Repeat this experiment with a very large sample of people and calculate the percentage of people who make four correct guesses out of six tries.
 - All of the methods listed above would provide an accurate estimate of the probability.
31. A college official conducted a survey to estimate the proportion of students currently living in dormitories about their preference for single rooms, double rooms, or multiple (more than two people) rooms in the dormitories on campus. Which of the following does NOT affect the college official's ability to generalize the survey results to all dormitory students?
- Five thousand students live in dormitories on campus. A random sample of only 500 were sent the survey.
 - The survey was sent to only first-year students.
 - Of the 500 students who were sent the survey, only 160 responded.
 - All of the above present a problem for generalizing the results.
32. The number of people living on American farms has declined steadily during the last century. Data gathered on the U.S. farm population (millions of people) from 1910 to 2000 were used to generate the following regression equation: Predicted Farm Population = $1167 - .59(\text{YEAR})$. Which method is best to use to predict the number of people living on farms in 2050?
- Substitute the value of 2050 for YEAR in the regression equation, and compute the predicted farm population.
 - Plot the regression line on a scatterplot, locate 2050 on the horizontal axis, and read off the corresponding value of population on the vertical axis.

- c. Neither method is appropriate for making a prediction for the year 2050 based on these data.
 - d. Both methods are appropriate for making a prediction for the year 2050 based on these data.
33. The following situation models the logic of a hypothesis test. An electrician uses an instrument to test whether or not an electrical circuit is defective. The instrument sometimes fails to detect that a circuit is good and working. The null hypothesis is that the circuit is good (not defective). The alternative hypothesis is that the circuit is not good (defective). If the electrician rejects the null hypothesis, which of the following statements is true?
- a. The circuit is definitely not good and needs to be repaired.
 - b. The electrician decides that the circuit is defective, but it could be good.
 - c. The circuit is definitely good and does not need to be repaired.
 - d. The circuit is most likely good, but it could be defective.

A-2 Midterm

Items 1 and 2 refer to the following situation: A college statistics class conducted a survey of how students spend their money. They gathered data from a large random sample of college students who estimated how much money they typically spent each week in different categories (e.g., food, entertainment, etc.). The following statistics were calculated for money spent weekly on food: mean = \$31.52; median = \$30.00; interquartile range = \$34.00; standard deviation = \$21.60; range = \$132.50.

1. A student states that the median food cost tells you that a majority of students in this sample spend about \$30 each week on food. How do you respond?
 - a. Agree, the median is an average and that is what an average tells you.
 - b. Agree, \$30 is representative of the data.
 - c. Disagree, a majority of students spend more than \$30.
 - d. Disagree, the median tells you only that 50% of the sample spent less than \$30 and 50% of the sample spent more.

2. The class determined that a mistake had been made and a value entered as 138 should have been entered as 38. They recalculate all of the statistics. Which of the following would be true?
 - a. The value of the median decreases, the value of the mean stays the same.
 - b. The values of the median and mean both decrease.
 - c. The value of the median stays the same, the value of the mean decreases.

3. What is the main purpose of **random assignment** in experiments? Please explain. (Points: 2)

Items 4 and 5 refer to the following situation:

4. Suppose two researchers wanted to determine if aspirin reduces the chance of a heart attack.

Researcher 1 studied the medical records of 500 randomly selected patients. For each patient, he recorded whether the person took aspirin every day and if the person had ever had a heart attack. Then he reported the percentage of heart attacks for the patients who took aspirin every day and for those who did not take aspirin every day. What type of study did Researcher 1 conduct?

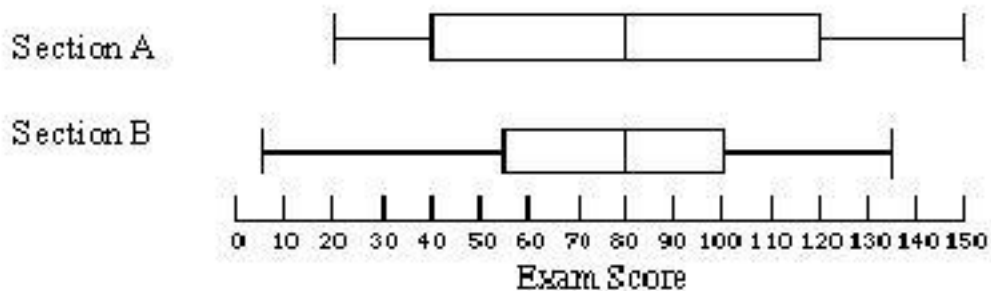
- a. Observational
- b. Experimental
- c. Survey
- d. None of the above

5. Researcher 2 also studied 500 patients that visited a regional hospital in the last year. He randomly assigned half (250) of the patients to take aspirin every day and the other half to take a placebo every day. Then after a certain length of time he reported the percentage of heart attacks for the patients who took aspirin every day and for those who did not take aspirin every day. What type of study did Researcher 2 conduct?
 - a. Observational
 - b. Experimental
 - c. Survey
 - d. None of the above

6. A medical doctor estimates the prevalence of broken arms among high school athletes by studying the records of her own patients. What type of sample is the doctor using in her study?

7. Return again to Question 6. Using her chosen sample, can the doctor generalize her results to a population of all high school athletes? Explain.

Items 8 to 10 refer to the following situation: The two boxplots below display final exam scores for all students in two different sections of the same course.



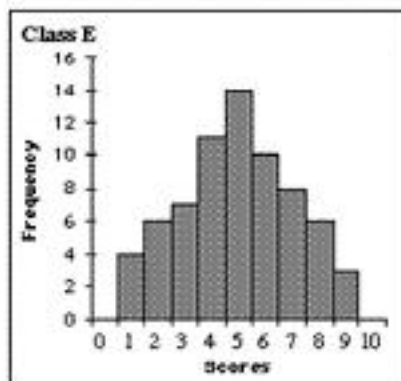
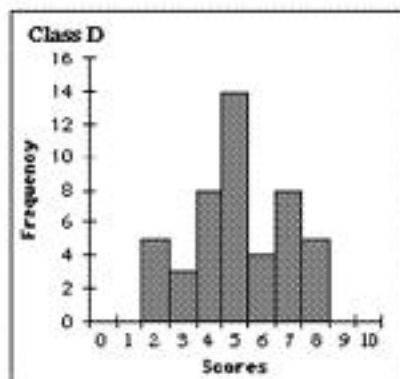
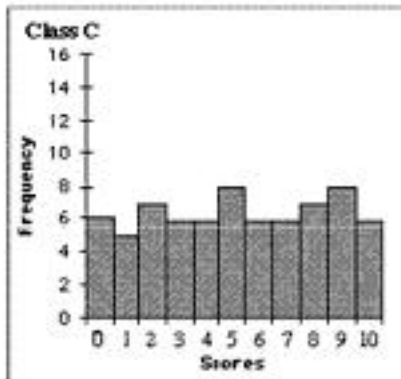
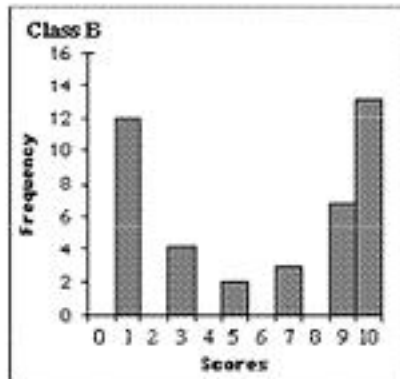
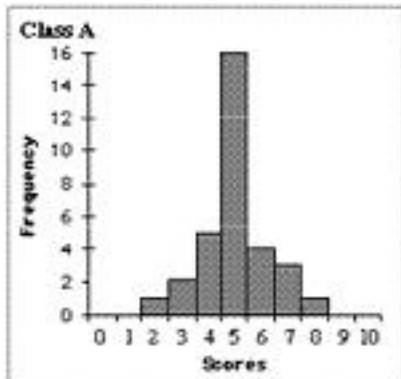
8. Which section would you expect to have a greater standard deviation in exam scores?
 - a. Section A
 - b. Section B.
 - c. Both sections are about equal.
 - d. It is impossible to tell.

9. Which data set has a greater percentage of students with scores at or below 30?
 - a. Section A
 - b. Section B.
 - c. Both sections are about equal.
 - d. It is impossible to tell.

10. Which section has a greater percentage of students with scores at or above 80?
- Section A
 - Section B.
 - Both sections are about equal.

Items 11 and 12 refer to the following situation:

Five histograms are presented below. Each histogram displays test scores on a scale of 0 to 10 for one of five different statistics classes.



11. Which of the classes would you expect to have the lowest standard deviation, and why?
- Class A, because it has the most values close to the mean.
 - Class B, because it has the smallest number of distinct scores.
 - Class C, because there is no change in scores.
 - Class A and Class D, because they both have the smallest range.
 - Class E, because it looks the most normal.
12. Which of the classes would you expect to have the highest standard deviation, and why?
- Class A, because it has the largest difference between the heights of the bars.
 - Class B, because more of its scores are far from the mean.
 - Class C, because it has the largest number of different scores.
 - Class D, because the distribution is very bumpy and irregular.
 - Class E, because it has a large range and looks normal.
13. A university administrator wanted to know the average amount of time students spend studying for their classes each week. She surveyed a random sample of 100 university students. It was found that these 100 students reported spending an average of 11.6 hours studying, with a standard deviation of 1.2 hours. After reading the results of the study, a university statistics professor suggested that there might have been some response bias. Explain what the statistics professor meant by response bias in the context of this particular study.
14. Jean lives about 10 miles from the college where she plans to attend a 10-week summer class. There are two main routes she can take to the school, one through the city and one through the countryside. The city route is shorter in miles, but has more stoplights. The country route is longer in miles, but has only a few stop signs and stoplights. Jean sets up a randomized experiment where each day she tosses a coin to decide which route to take that day. She records the following data for 5 days of travel on each route.

Country Route - 17, 15, 17, 16, 18

City Route - 18, 13, 20, 10, 16

It is important to Jean to arrive on time for her classes, but she does not want to arrive too early because that would increase her parking fees. Based on the data gathered, which route would you advise her to choose?

- The Country Route, because the times are consistently between 15 and 18 minutes.
- The City Route, because she can get there in 10 minutes on a good day and the average time is less than for the Country Route.

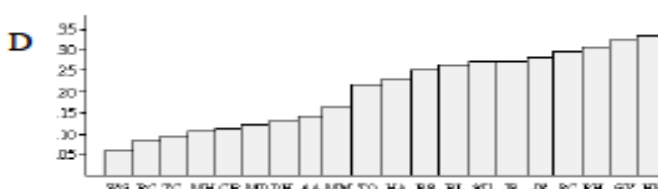
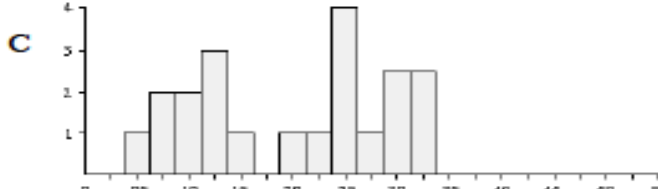
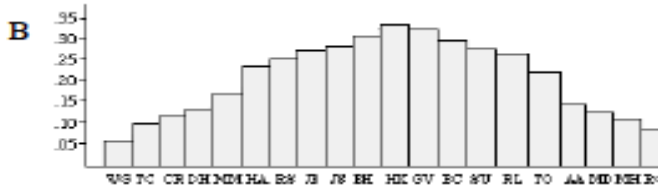
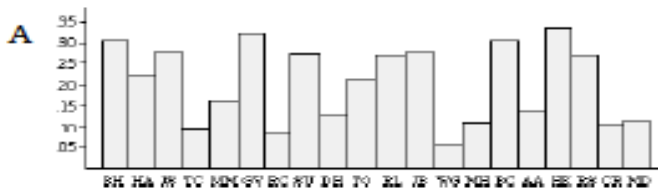
c. Because the times on the two routes have so much overlap, neither route is better than the other. She might as well flip a coin.

15. A baseball fan likes to keep track of statistics for the local high school baseball team. One of the statistics she recorded is the proportion of hits obtained by each player based on the number of times at bat as shown in the table below. Which of the following graphs gives the best display of the distribution of proportion of hits in that it allows the baseball fan to describe the shape, center and spread of the variable, proportion of hits?

Player	Proportion of hits
BH	0.305
HA	0.229
JS	0.281
TC	0.097
MM	0.167
GV	0.333
RC	0.085

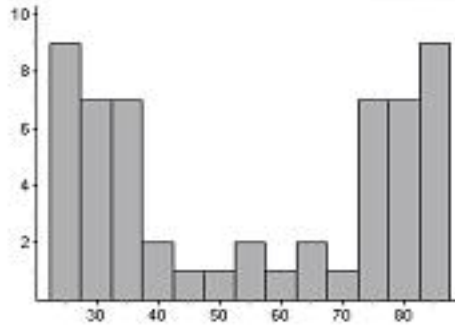
Player	Proportion of hits
SU	0.270
DH	0.136
TO	0.218
RL	0.267
JB	0.270
WG	0.054
MH	0.108

Player	Proportion of hits
BC	0.301
AA	0.143
HK	0.341
RS	0.261
CR	0.115
MD	0.125



Screen Shot :

16. Consider the accompanying histogram, and the list of variables (A, B, C, and D) below it. Which ONE of the variables (A, B, C, or D) shows the correct summary statistics for the histogram?

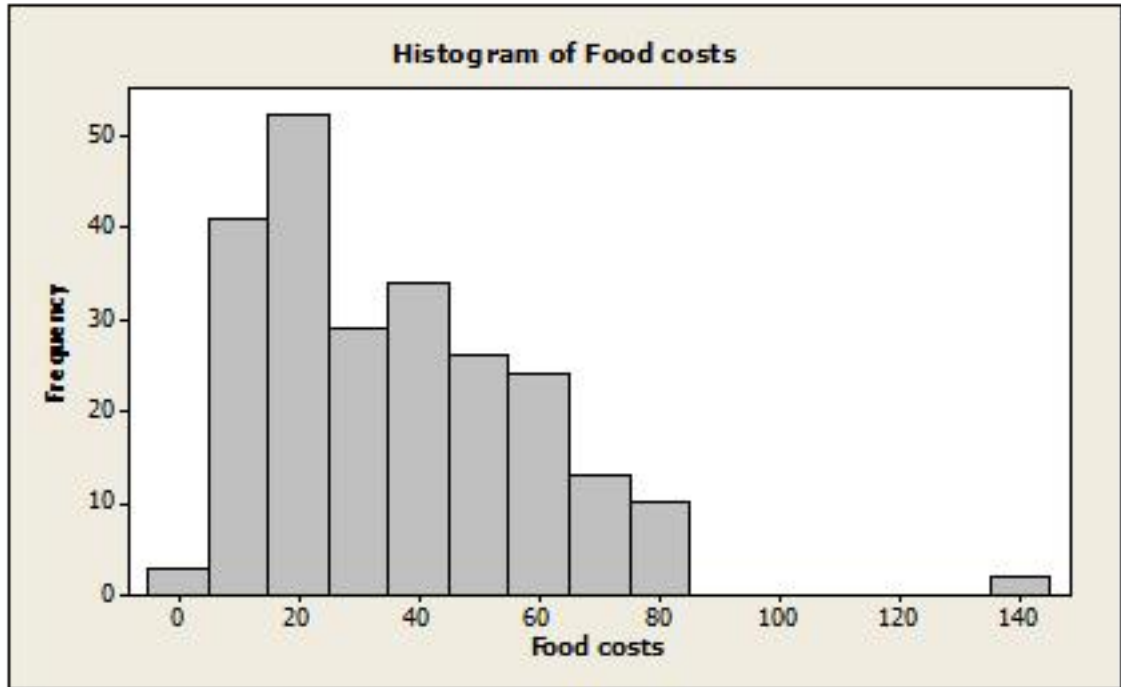


Variable	Mean	Median	Standard Deviation
A	55	55	15
B	55	55	25
C	60	50	25
D	60	60	15

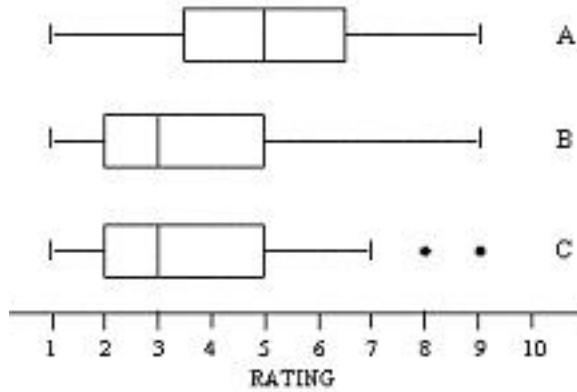
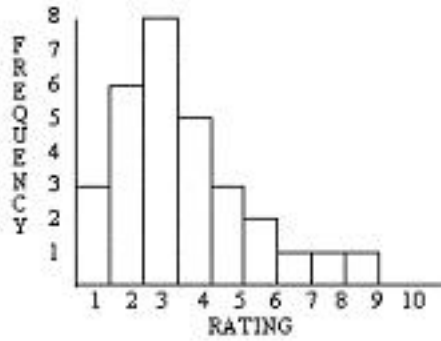
mic

1. A
2. B
3. C
4. D

17. Researchers conducted a survey of how students spend their money, and the accompanying histogram shows the reported weekly food costs (in dollars). What would you say to someone who examined the histogram and said, “There are about fifty students who spend \$20 a week on food” Explain.

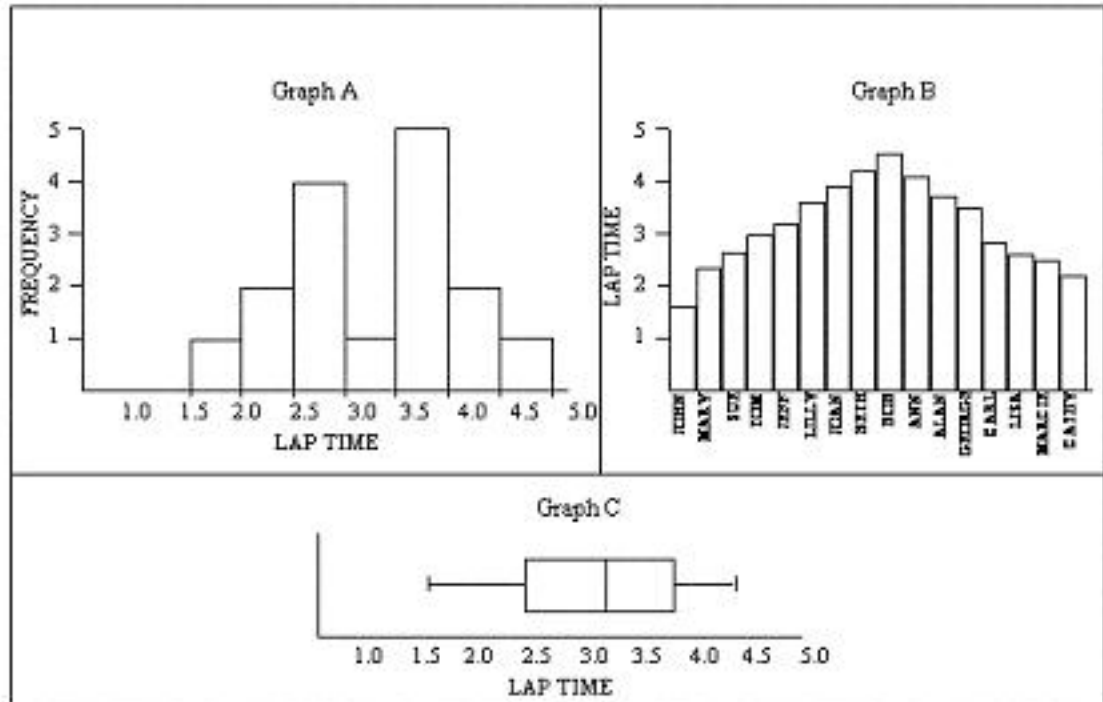


18. One of the items on the student survey for an introductory statistics course was "Rate your aptitude to succeed in this class on a scale of 1 to 10" where 1 = Lowest Aptitude and 10 = Highest Aptitude. Look carefully at the accompanying histogram of the data, and the three box plots below the histogram. Which ONE of the three box plots (A, B, or C) represents the same data set that is shown in the histogram?



1. A
2. B
3. C

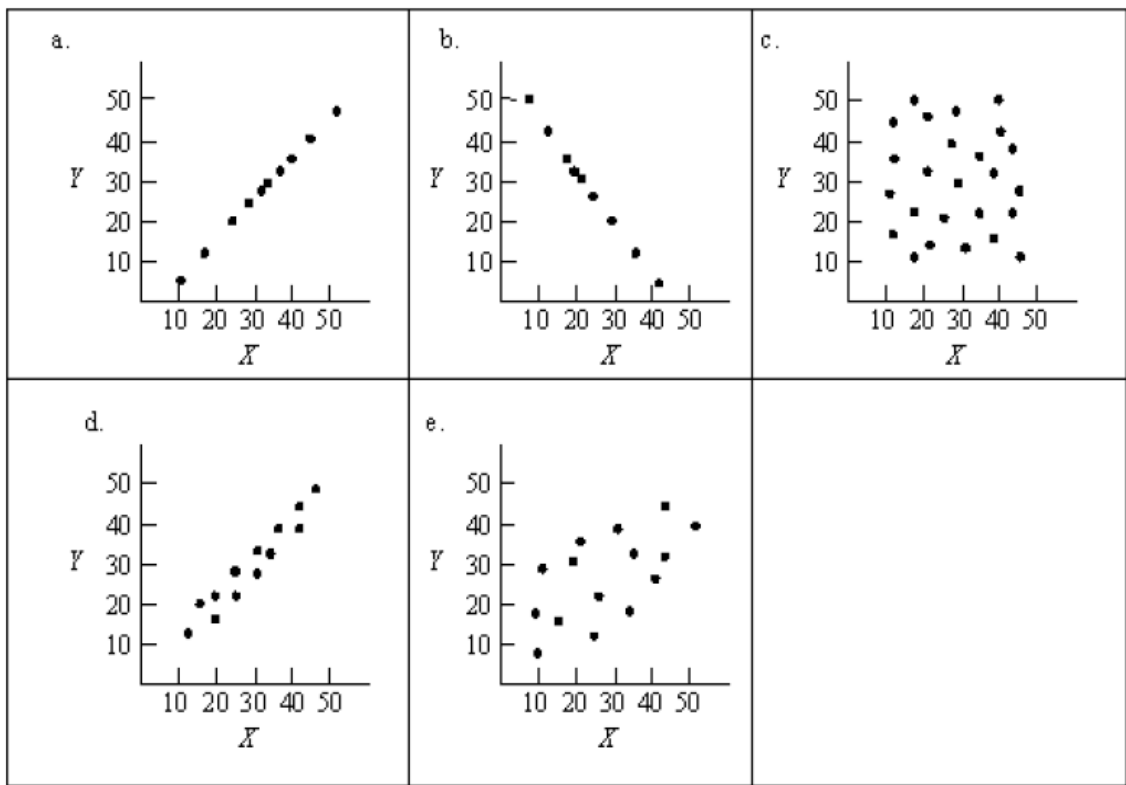
19. A local running club has its own track and keeps accurate records of each member's individual best lap time (in minutes) around the track. Several graphs of this data are shown here. Which of the graphs (A, B, or C) allows you to most easily see the shape of the distribution of running times? Note it may be hard to read the x-axis in Graph B, but the names of the different members are displayed along that axis.



- a. A
 b. B
 c. C
20. A class of 30 introductory statistics students took a 15 item quiz, with each item worth 1 point. The standard deviation for the resulting score distribution is 0. You know that:
- a. about half of the scores were above the mean.
 b. an arithmetic error must have been made.
 c. everyone correctly answered the same number of items.
 d. the mean, median, and mode must all be 0.

21. A teacher gives a 15 item science test. For each item, a student receives one point for a correct answer; 0 points for no answer; and loses one point for an incorrect answer. Total test scores could range from +15 points to -15 points. The teacher computes the standard deviation of the test scores for the class to be -2.30. What do we know?
- The standard deviation was calculated incorrectly.
 - Most students received negative scores.
 - Most students scored below the mean.
 - None of the above.

Items 22 to 24 refer to the following situation: Consider the five scatterplots that are shown below:



22. Select the scatterplot that shows a correlation of zero?
- -
 -
 -
 -

23. Select the scatterplot that shows a correlation of about .60?
- a.
 - b.
 - c.
 - d.
 - e.
24. Select the scatterplot that shows the strongest relationship between the X and Y variables?
- a. a
 - b. b
 - c. a and b
 - d. a and d
 - e. a, b, and d

For questions 25–29 use the following:

The salaries of the CEOs and the stock prices of 24 companies are shown in this scatterplot.



25. In this situation, what do the cases (dots) in the scatterplot represent?
26. Describe the shape, trend, and strength of the relationship (in the scatterplot in question 25) between CEO salary and stock price. (Be sure to describe all three)
27. Write 1-2 sentences explaining the relationship between CEO salary and stock price (see scatterplot Question 25). Write this so someone who doesn't understand statistics can understand (i.e. nonstatistical language)
28. A CEO uses this relation to argue there is a positive correlation between CEO salary and the Stock Price, and therefore, for the good sake of the company's stock

value, his salary should be increased. Do you agree with his argument? Explain why or why not.

29. Could the researcher generalize to all companies from this study (referring back to question 25)? Why or why not?

A-3 Directions on Collaborative Test

Consensus Section

Group Test EPSY 3264

This assignment is a group test that should be completed within your discussion room. The test will become available on Monday at noon and it should be completed by midnight Saturday. You should only work on the test with your group members.

The test consists of 15 questions worth 20 points in total (the group member that submits the test gets 1 extra credit). In order to receive full credit for this test, you have to contribute to the assignment within your discussion group in a specific way (as outlined below).

Grading is based on participation and correctness on the test.

1. Members have to post their initial post before midnight on Wednesday. The initial post should be student's individual answers to the whole test.
2. In addition to the first post, each member has to provide at least two meaningful answers or comments in his or her discussion group. Each post is worth 33.33% of the grade on the test. For example if the score on the test ends up being 15 points and you only contributed two times, you will receive a score of 10 points (which is 66.67% of 15). The three posts (the initial post +two comments/questions/answers) are the minimum and the initial post with individual answers is required to receive a full grade.
3. The group has to come to a consensus regarding answers. For each question, you need to come up with ONE final answer. If two different answers are provided to a question, no credit will be given for that question.
4. One group member should submit one copy of the test through the assignment tool in WebVista once the test is completed. Please include names of those group members who participated.

Make your answers clear and remember to explain when asked in order to receive full credit.

Nonconsensus

Group test EPSY 3264

This assignment is a group test that should take place within your discussion group. You should only work on the test with your group members however the final version of the test should be submitted individually by each group member through assessment tools in WebVista. The test will become available on Monday at noon and it should be completed by midnight Saturday.

The test consists of 15 questions worth 20 points in total. In order to receive full credit for this test, you have to contribute to the assignment within your discussion group in a specific way (as outlined below).

Grading is based on participation and correctness on the test.

1. Members have to post their initial post in their discussion room before midnight on Wednesday. The initial post should be student's individual answers to the whole test.
2. In addition to the first post, each member has to provide at least two meaningful answers or comments in his or her discussion group. Each post is worth 33.33% of the individual grade on the test. For example if the score on the individual test ends up being 15 points and you only contributed two times to the discussion, you will receive a score of 10 points (which is 66.67% of 15). The three posts (the initial post +two comments/questions/answers) are the minimum and the initial post with individual answers is required to receive a full grade.
3. The group does not have to reach a consensus regarding answers, since each group member submits their own final version of the test.
4. Each group member should submit the test individually through the assessment tool in WebVista before the due date on Saturday.

Make your answers clear and remember to explain when asked in order to receive full credit.

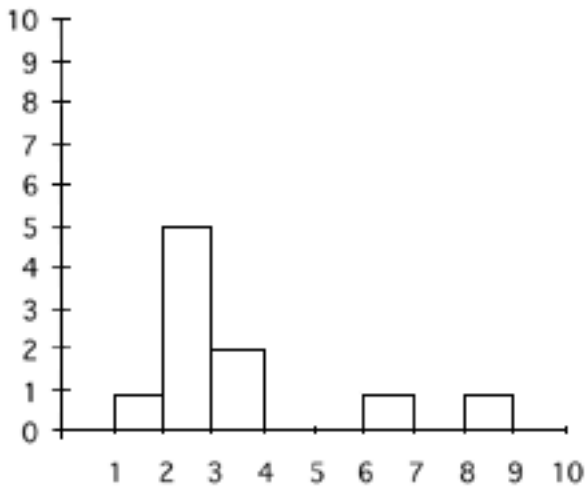
A-4 Collaborative Tests

**Group Test #1
EPSY-3264**

1. Suppose you wanted to conduct a study to compare two medical treatments on people with a specific medical condition. Previous research suggests that a person's current medical condition might be related to their age, sex, previous medical history, body mass index (a measure combining weight and height), and smoking behavior. You are told by your statistical consultant to randomly assign the subjects in your study to two groups, so that one treatment may be randomly assigned to each group. A friend asks why you wouldn't be better off forcing the groups to be balanced with respect to the characteristics listed above, instead of leaving it to chance to make the groups comparable. How would you answer your friend? **(2 points)**
2. What is the difference between a parameter and a statistic? Provide an example using the variable "height" that illustrates this difference. **(1 point)**

For items 3-5 use the following:

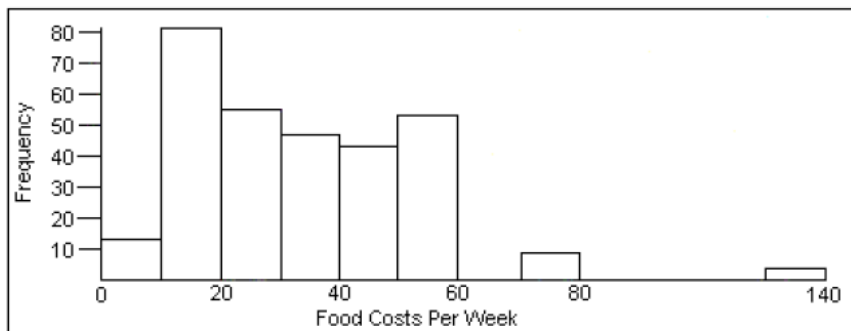
Here is a histogram for a set of test scores from a 10-item makeup quiz given to a group of students who were absent on the day the quiz was given.



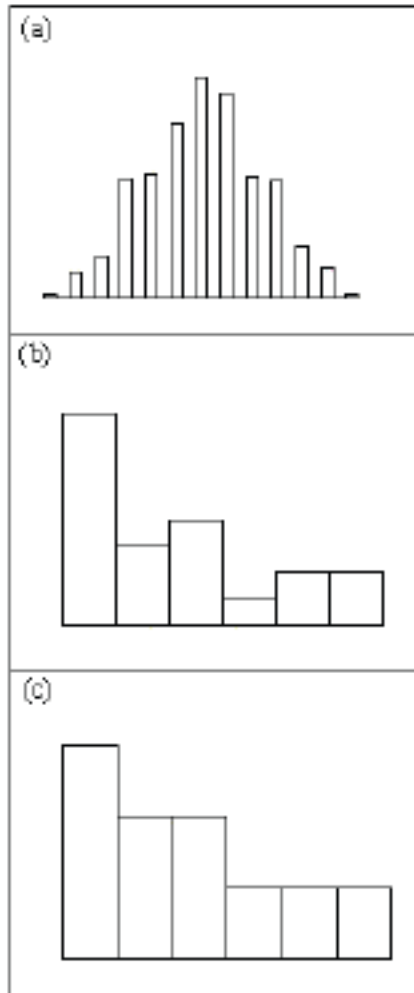
3. What do the numbers on the vertical axis represent? **(1 point)**
 - a. Independent variable
 - b. Scores on the test
 - c. Dependent variable
 - d. Number of Students

4. How many people received scores higher than 4? **(1 point)**
- a. 1
 - b. 2
 - c. 3
 - d. 4
5. How many people took the test and have scores represented in the graph? **(1 point)**
- a. 5
 - b. 10
 - c. 20
6. A researcher is studying the relationship between a vitamin supplement and cholesterol level. What type of study needs to be done in order to establish that the amount of vitamin supplement causes a change in cholesterol level? **(1 point)**
- a. Correlational study
 - b. Time Series study
 - c. Randomized experiment
 - d. Survey
7. The dean of a college would like to determine the feelings of students concerning a new registration fee that would be used to upgrade the recreational facilities on campus. All registered students would pay the fee each term. Which of the following data collection plans would provide the best representation of students' opinions at the school? **(1 point)**
- a. Survey every 10th student who enters the current recreational facilities between the hours of 1:00 and 5:00 pm until 100 students have been asked.
 - b. Randomly sample fifty student ID numbers and send a survey to all students in the sample.
 - c. Place an ad in the campus newspaper inviting students to complete an online survey. Collect the responses of the first 200 students who respond.
 - d. All of the above would be equally effective.
8. A team in the Department of Institutional Review at a large university wanted to study the relationship between completing an internship during college and students' future earning potential. From the same graduating class, they selected a random sample of 80 students who completed an internship and 100 students who did not complete an internship and examined their salaries 5 years past graduation. They found that there was a statistically higher mean salary for the internship group than for the non-internship group. Which of the following interpretations do you think is the most appropriate? **(1 point)**

- a. More students should take internships because having an internship produces a higher salary.
 - b. There could be a confounding variable, such as student major, that explains the difference in mean salary between the internship and no internship groups.
 - c. You cannot draw any valid conclusions because the samples are not the same size.
9. A college statistics class conducted a survey. They gathered data from a large random sample of students who estimated how much money they typically spent each week in different categories (e.g., food, entertainment, etc.). A distribution of the survey results is presented below. One student claims the distribution of food costs basically looks bell-shaped, with one outlier. How would you respond? (1 point)



- a. Agree, it looks pretty symmetric if you ignore the outlier.
 - b. Agree, most distributions are bell-shaped.
 - c. Disagree, it looks more skewed to the left.
 - d. Disagree, it looks more skewed to the right.
 - e. Disagree, it looks more bimodal.
10. M&M Candies reports that the plain M&Ms are manufactured with 30% brown, 20% red, 20% yellow, 10% orange, 10% blue, and 10% green candies. Below are 3 graphs. One represents the colors of individual M&Ms for the population of all M&M Candies, one represents the colors in a sample of 20 candies, and one represents the sampling distribution of the proportion of brown candies in 500 samples of size 50. You should not assume the vertical or horizontal scales are the same. Identify which graph is which and explain your choices. (2 points)



11. Four students at a local high school conducted surveys. Shannon got the names of all 800 children in the high school and put them in a hat, and then pulled out 60 of them. Jake asked 10 students at an after-school meeting of the computer games club. Adam asked all of the 200 children in Grade 10. Claire set up a booth outside of the school. Anyone who wanted to stop and fill out a survey could. She stopped collecting surveys when she got 60 students to complete them. Who do you think has the best sampling method? Why? **(2 points)**

For items 12-15 use the following:

CNN conducted a quick vote poll on September 19, 1999 to determine "What proportion of Americans think that the Miss America pageant is still relevant today?" The poll was conducted on the internet. Here are the results of the poll: Is the Miss America pageant still relevant today? Yes: 1192 votes; No: 4389 votes; Total: 5581 votes.

12. What is the parameter of interest? Please describe. **(1 point)**

13. Based on these results, can reliable conclusions be made about how all Americans feel. Why not? **(1 point)**

14. Do you have recommendations for CNN to improve their poll? Explain briefly. **(2 points)**

15. People who eat lots of fruits and vegetables have lower rates of colon cancer than those who eat little of these foods. Fruits and vegetables are rich in "antioxidants" such as vitamins A, C, and E. Will taking antioxidants help prevent colon cancer? A clinical trial studied this question with 864 people who were at risk of colon cancer.

The subjects were divided into four groups that each received a different type of daily supplement: (1) daily beta carotene, (2) daily vitamins C and E, (3) all three vitamins every day, and (4) daily placebo. After four years, the researchers were surprised to find no significant difference in colon cancer among the groups.

What are the explanatory and response variables in this experiment? **(2 points)**

Group Test 2
EPSY-3264

1. The following counts of raisins were obtained in a class activity. Summarize and describe this data set (**2 points**).

30	40	37	35	42	28	29	24	25	26	23	19	18
----	----	----	----	----	----	----	----	----	----	----	----	----

2. A 30 item math test was graded using the following procedure: a correct response was scored as +1, a blank response was scored 0, and an incorrect response was scored -1. The maximum possible test score was 30; the lowest score possible was -30. The standard deviation of the test scores for the class was reported to be -2.13. Therefore, (**1 point**)
- some students received negative scores
 - the test was too hard for this class
 - the class performed poorly on this test
 - the standard deviation was calculated incorrectly
 - most students received positive scores

3. A college statistics class conducted a survey of how students spend their money. They gathered data from a large random sample of college students who estimated how much money they typically spent each week in different categories (e.g., food, entertainment, etc.).

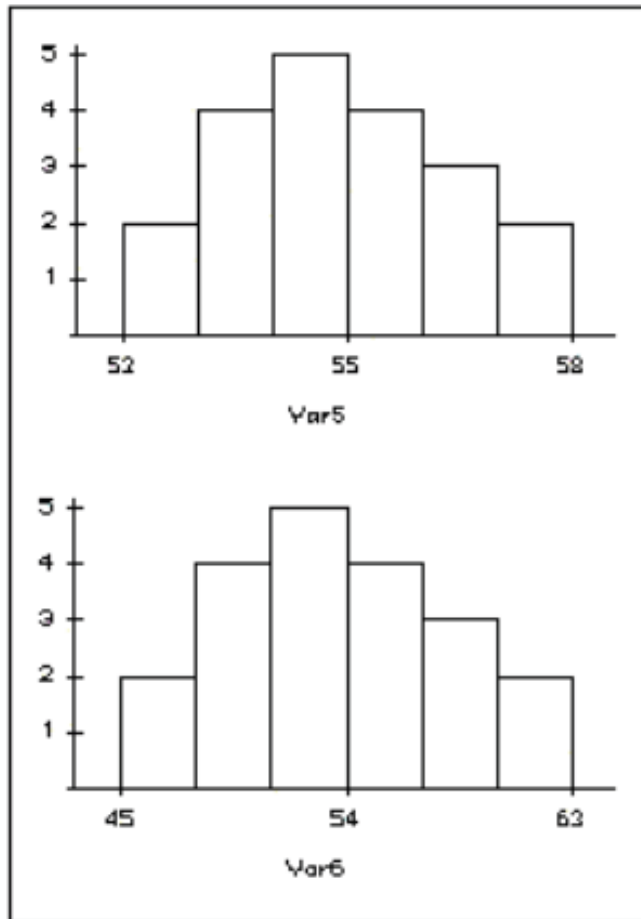
The following statistics were calculated for money spent weekly on food: mean = \$31.52; median = \$30.00; interquartile range = \$34.00; standard deviation = \$21.60; range = \$132.50.

A student states that the median food cost tells you that a majority of students in this sample spend about \$30 each week on food. How do you respond? (**1 point**)

- Agree, the median is an average and that is what an average tells you.
 - Agree, \$30 is representative of the data.
 - Disagree, a majority of students spend more than \$30.
 - Disagree, the median tells you only that 50% of the sample spent less than \$30 and 50% of the sample spent more.
4. As part of its twenty-fifth reunion celebration, the Class of 1980 of State University mailed a questionnaire to its members. One of the questions asked the respondent to give his or her total income last year. Of the 820 members of the class of 1980, the university alumni office had addresses for 583. Of these, 175 returned the questionnaire. The reunion committee computed the mean income given in the responses and announced, "The members of the class of 1980 have enjoyed resounding success. The average income of class members is \$120,000!"

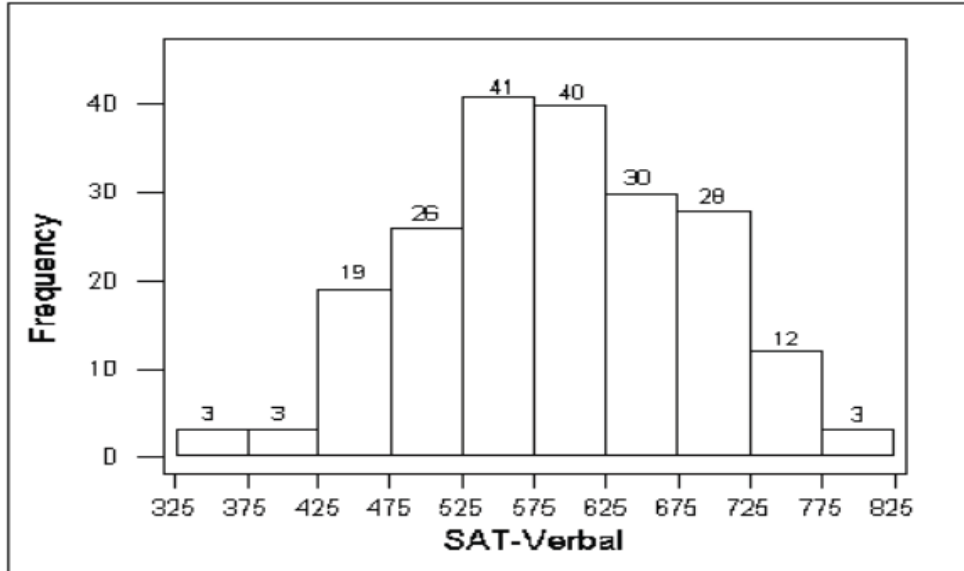
Identify two distinct sources of bias or misleading information in this result, being explicit about the direction of bias you expect. Explain how you might fix each of these problems (2 points).

5. When calculating a standard deviation by hand, what should the sum of the deviations from the mean always equal? Explain why. (1 point)
6. Consider the two histograms displayed below. The histogram labeled "Var5" has a mean of 54 and the histogram labeled "Var6" has a mean of 53. Please indicate which one has a larger standard deviation and WHY that histogram has the larger standard deviation (2 points).



7. Suppose two distributions have exactly the same mean and standard deviation. Someone says that the two distributions have to look exactly alike. Is this True or False? Explain your reasoning (2 points).

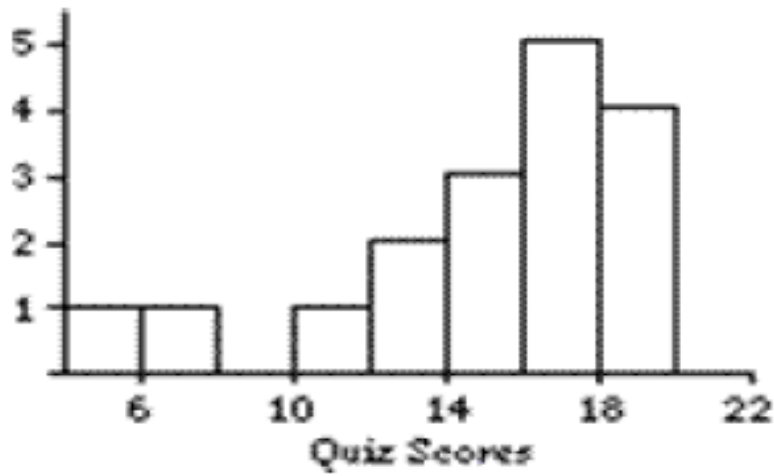
8. Which of the following is most sensitive to outliers? (1 point).
- interquartile range
 - standard deviation
 - median
 - mode
9. The following histogram shows the Verbal SAT scores for 205 students entering a local college in the fall of 2002 (2 points).



Would the five-number summary or the mean and standard deviation be a better summary for this distribution? Explain your choice.

10. The distribution of the top 1% of individual incomes in the US is strongly skewed to the right. In 1997, the two measures of center for the top 1% of individual incomes were \$330,000 and \$675,000. Which number represents the mean income of the top 1% and which number represents the median income of the top 1%? Choose the best answer. (1 point)
- mean = \$330,000 and median = \$675,000
 - median = \$330,000 and mean = \$675,000
 - Not enough information to tell which is which.

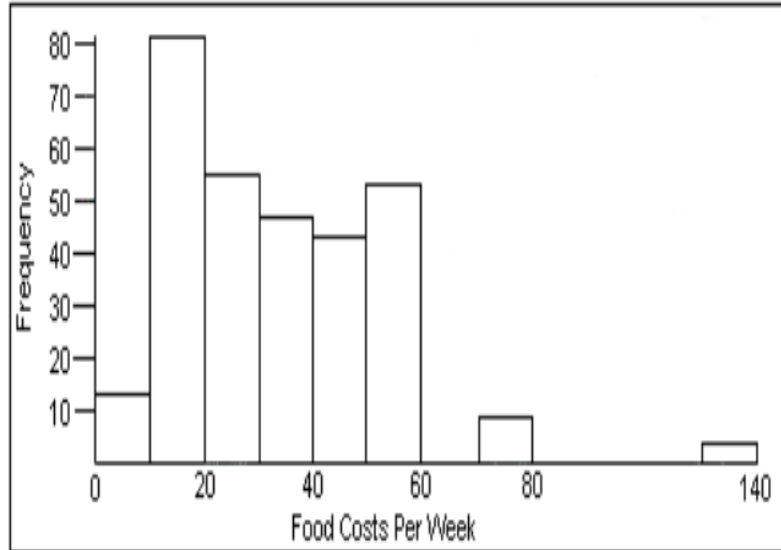
11. For this graphical display of Quiz Scores, which estimates of the mean and median are most plausible? (1 point).



- a. median = 13.0 and mean = 12.0
b. median = 14.0 and mean = 15.0
c. median = 16.0 and mean = 14.3
d. median = 16.5 and mean = 16.2
12. Consider two populations in the same state. Both populations are the same size (22,000). Population 1 consists of all students at the State University. Population 2 consists of all residents in a small town. Consider the variable Age. Which population would most likely have the higher standard deviation for Age? (1 point)
- a. Population 1 would more likely have a higher standard deviation (SD) for age than Population 2.
b. Population 2 would more likely have a higher standard deviation (SD) for age than Population 1.
c. They would likely have the same standard deviation (SD) for age because they have the same population size.
d. There is not enough information to tell.

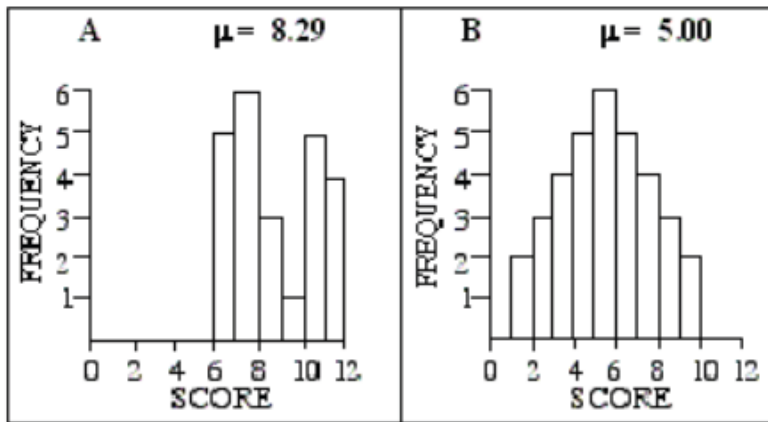
Items 13 and 14 refer to the following situation:

This is a distribution of how much money was spent per week for a random sample of college students.



13. The range for this distribution is \$132.50. Indicate your agreement or disagreement with the following statement: The range is not a useful summary of the variability of this data set. (1 point)
- Agree, because the range is not an accurate statistic.
 - Agree, because the range is too easily influenced by outliers.
 - Disagree, because the range uses all of the information in the data set.
 - Disagree, because students tend to spend any amount of money between \$0 and \$132.50.
14. What is the best measure to use to summarize the variability of this data set? (1 point)
- Range, because it tells you the overall spread of the data.
 - Standard deviation, because it is based on all the information in the data set.
 - Standard deviation, because it is the most commonly used measure of variability.
 - Interquartile range, because it is resistant to outliers.

15. For each pair of graphs, determine which graph has the higher standard deviation (it is not necessary to do any calculations to answer this question). (1 point)



- A has a larger standard deviation than B
- B has a larger standard deviation than A
- Both graphs have the same standard deviation

Group Test 3
EPSY-3264

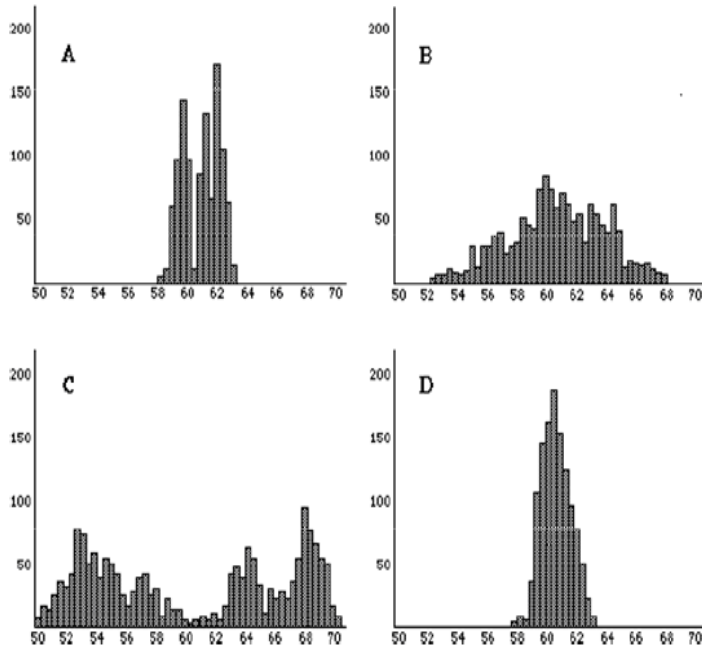
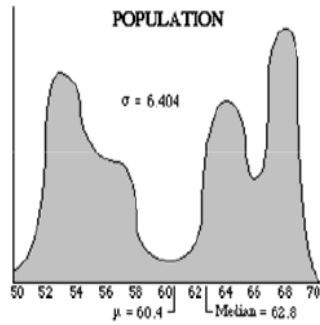
1. A health insurance company is interested in the cholesterol levels for individuals' ages 40 or older. A random sample of 100 individuals was chosen from the target population. The following information was obtained from the sample: average = 158 mg, median = 159 mg, s.d. = 20 mg. One individual has a cholesterol level at 175 mg. Based on only the summary statistics, is this an unusually high level of Cholesterol for someone from this population? Why or why not? **(1 point)**
 - a. Yes, because 175 is 17 mg higher than the mean cholesterol level.
 - b. Yes, because it is better to have a low cholesterol level.
 - c. No, because 175 is less than one standard deviation above the mean.
 - d. No, because it is not above 200 mg, the recommended maximum adult Cholesterol level.

2. A doctor collects a large set of heart rate measurements that approximately follow a normal distribution. He only reports 3 statistics, the mean = 110 beats per minute, the minimum = 65 beats per minute, and the maximum = 155 beats per minute. Which of the following is most likely to be the standard deviation of the distribution? **(1 point)**
 - a. 5
 - b. 15
 - c. 35
 - d. 90

3. A student was studying the relationship between how much money students spend on food and on entertainment per week. Based on a sample size of 270, he calculated a correlation coefficient (r) of .013 for these two variables. Which of the following is an appropriate interpretation? **(1 point)**
 - a. This low correlation of .013 indicates there is no relationship.
 - b. There is no linear relationship but there may be a nonlinear relationship.
 - c. This correlation indicates there is some type of linear relationship.

Items 4 to 8 refer to the following situation:

A hypothetical distribution for a population of test scores is displayed below. The population has a mean of 60.4, a median of 62.8, and a standard deviation of 6.404. Each of the other four graphs labeled A to D represent possible distributions of sample means for random samples drawn from the population.



4. Which graph best represents a distribution of sample means for 1000 samples of size 4? (1 point)

- a. A
- b. B
- c. C
- d. D

5. What do you expect for the variability (spread) of the sampling distribution of the mean from samples of size 4? **(1 point)**
- Same as the population.
 - Less variability than the population (a narrower distribution).
 - More variability than the population (a wider distribution).
6. Which graph best represents a distribution of sample means for 1000 samples of size 50? **(1 point)**
- A
 - B
 - C
 - D
7. What do you expect for the shape of the sampling distribution (the distribution of sample means for all possible samples of size $n = 50$)? **(1 point)**
- Shaped more like a normal distribution.
 - Shaped more like the population.
 - Shaped like neither the population nor the normal distribution.
8. A recent article in an educational research journal reports a correlation of $+0.8$ between math achievement and overall math aptitude for a large sample of students. It also reports a correlation of -0.8 between math achievement and a math anxiety test for the same group of students. Only students with scores on all three measures were included in the study. Which of the following interpretations is the most correct? **(1 point)**
- The correlation of $+0.8$ indicates a stronger relationship than the correlation of -0.8
 - The correlation of $+0.8$ is just as strong as the correlation of -0.8
 - It is impossible to tell which correlation is stronger
9. For a sample of 9 men, the mean weight is 175 lb. with a standard deviation of 15 lb. What is the standard error of the mean (standard deviation of the sampling distribution)? What information does this standard error give you? **(2 points)**

10. The distributions of SAT and LSAT scores are both approximately normal and symmetric. Veronica took both tests (at different times) and would like to know on which test her performance was better. Use the data given on each test to decide which score was better, relative to other people who took each test. Explain your answer. (2 points)

Test	Veronica's score	Mean score	std. deviation
SAT	875	998	203
LSAT	145	150	9

11. Shelly is going to flip a coin 50 times and record the percentage of heads she gets. Her friend Diane is going to flip a coin 10 times and record the percentage of heads she gets. Which person is more likely to get 20% or fewer heads? (2 points)
- Diane because the more you flip the closer you get to 50% and she did only 10 flips.
 - Shelly because the greater the sample size, the greater the variability in the results.
 - Neither because each coin flip is a separate event and the probability of heads is not affected by the number of times flipped.
12. Identify a pair of variables for which you would expect to see a strong correlation but not a cause-and-effect relationship. Suggest an explanation for the association. (1 point)

Items 13 to 14 refer to the following situation:

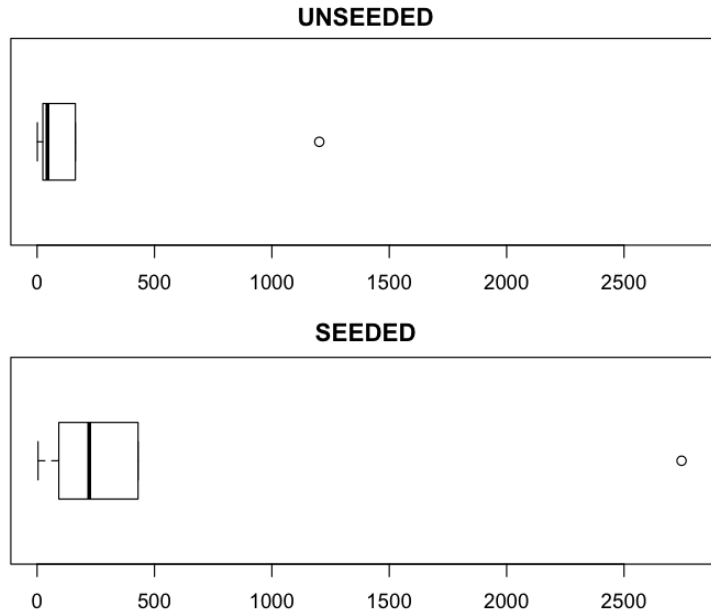
The following data are from a cloud seeding experiment, to determine whether seeding clouds increase rainfall. In this experiment, clouds were randomly assigned to be seeded or not, and the amount of rain generated from each cloud was then measured. The following statistics were obtained. The variable measured is acre-feet of rainfall.

	Count	Mean	Median	StdDev	Min	Max	Range	IntQRange	Lower quartile	Upper quartile
Unseeded	26	164	44	278	1	1202	1202	138.6	24.4	163
Seeded	26	445	222	650.8	4	2745	2741	337.6	92.4	430

13. Just by looking at the statistics, which procedure seems to produce more variability in amounts of rainfall? Why? (1 point)

14. Looking at the statistics, do you see evidence that seeding clouds increases rainfall? Why or why not? (2 points)

Boxplots based on min, Q1, median, Q3 and max



15. A study was planned to examine the length of a certain species of fish on Gull Lake. The initial plan was to take a random sample of 100 fish from this lake using a special net, and examine the results. Numerical summaries on lengths of the fish measured in this study are given. Use them to answer the question. (2 points)

Mean	25.018 in
Median	25.295
Standard Deviation	4.1831
Range	20.73
Min	12.67
Max	33.4
N	78

Notice there were only 78 fish that were actually sampled. This may result in some bias in the fish lengths. Which direction, if either, do you think the bias may be, and why?

A-5 Students Perception on Collaborative Tests (SPCT)

Please rate the extent to which you agree or disagree with each of the following statements as they reflect your experience of taking this online introductory statistics course, especially the three group tests.

1. Instructions for group tests were clear.
Strongly disagree, disagree, agree, strongly agree
2. Instructions for doing group tests resulted in everyone in the group contributing equally.
Strongly disagree, disagree, agree, strongly agree
3. In general, I was an active participant in all three of the group tests.
Strongly disagree, disagree, agree, strongly agree
4. Overall, group members contributed fairly equally to the group tests.
Strongly disagree, disagree, agree, strongly agree
5. Group tests were less stressful than if I had to take the test on my own.
Strongly disagree, disagree, agree, strongly agree
6. Overall, the grading of the group tests were fair.
Strongly disagree, disagree, agree, strongly agree
7. I feel that group members that contribute less to group tests should not receive the same grade as others.
Strongly disagree, disagree, agree, strongly agree
8. I prepared as much for group tests as I would have done for individual tests.
Strongly disagree, disagree, agree, strongly agree
9. Working together on group tests helped me better understand concepts of statistics than if I had taken the test on my own.
Strongly disagree, disagree, agree, strongly agree

10. Participation in group tests increased my confidence in statistics more than if I had taken the test on my own.
Strongly disagree, disagree, agree, strongly agree
11. Working together on group tests helped me remember information that I had forgotten more than if I had taken the test on my own.
Strongly disagree, disagree, agree, strongly agree
12. Working together on group tests often helped me revise my initial answers on the tests.
Strongly disagree, disagree, agree, strongly agree
13. Participation in group tests was an important aspect of learning statistics in this course.
Strongly disagree, disagree, agree, strongly agree
14. Group tests were an effective way to assess my learning in this course.
Strongly disagree, disagree, agree, strongly agree
15. I would have preferred to take individual tests.
Strongly disagree, disagree, agree, strongly agree
- 15.b I would have preferred to take only individual tests.
Strongly disagree, disagree, agree, strongly agree
16. I would have preferred being able to discuss group tests with my group but submit my own individual answers.
Strongly disagree, disagree, agree, strongly agree
- 16.b I would have preferred being able to discuss group tests with my group and submit one set of answers as a group.
Strongly disagree, disagree, agree, strongly agree
17. Participation in group tests helped me earn a higher grade on the group tests than if I had taken the test on my own.
Strongly disagree, disagree, agree, strongly agree

18. Considering the total points for a grade (labs, discussion assignments, article critiques, graph critique, the midterm, group tests and the final exam), the amount of points earned in group tests in this course was a fair way of assessing my learning.

Strongly disagree, disagree, agree, strongly agree

19. Please state what you liked most about group tests.

20. Please state what you did not like about group tests.

A-6 The Survey Of Attitudes Toward Statistics (SATS-36)

Pre survey

DIRECTIONS: The statements below are designed to identify your attitudes about statistics. Each item has 7 possible responses. The responses range from 1 (strongly disagree) through 4 (neither disagree nor agree) to 7 (strongly agree). If you have no opinion, choose response 4. Please read each statement. Mark the one response that most clearly represents your degree of agreement or disagreement with that statement. Try not to think too deeply about each response. Record your answer and move quickly to the next item. Please respond to all of the statements.

	Strongly disagree		Neither disagree nor agree			Strongly agree	
I plan to complete all of my statistics assignments.	1	2	3	4	5	6	7
I plan to work hard in my statistics course.	1	2	3	4	5	6	7
I will like statistics.	1	2	3	4	5	6	7
I will feel insecure when I have to do statistics problems.	1	2	3	4	5	6	7
I will have trouble understanding statistics because of how I think.	1	2	3	4	5	6	7
Statistics formulas are easy to understand.	1	2	3	4	5	6	7
Statistics is worthless.	1	2	3	4	5	6	7
Statistics is a complicated subject.	1	2	3	4	5	6	7
Statistics should be a required part of my professional training.	1	2	3	4	5	6	7
Statistical skills will make me more employable.	1	2	3	4	5	6	7
I will have no idea of what's going on in this statistics course.	1	2	3	4	5	6	7
Statistics is not useful to the typical professional.	1	2	3	4	5	6	7
I plan to study hard for every statistics test.	1	2	3	4	5	6	7
I will get frustrated going over statistics tests in this course.	1	2	3	4	5	6	7

	Strongly disagree		Neither disagree nor agree			Strongly agree	
Statistical thinking is not applicable in my life outside my job.	1	2	3	4	5	6	7
I use statistics in my everyday life	1	2	3	4	5	6	7
I will be under stress during statistics class.	1	2	3	4	5	6	7
I will enjoy taking statistics courses.	1	2	3	4	5	6	7
I am interested in using statistics.	1	2	3	4	5	6	7
Statistics conclusions are rarely presented in everyday life.	1	2	3	4	5	6	7
Statistics is a subject quickly learned by most people.	1	2	3	4	5	6	7
I am interested in understanding statistical information.	1	2	3	4	5	6	7
Learning statistics requires a great deal of discipline.	1	2	3	4	5	6	7
I will have no application for statistics in my profession.	1	2	3	4	5	6	7
I will make a lot of math errors in statistics.	1	2	3	4	5	6	7
I plan to log into the course website two times a week	1	2	3	4	5	6	7
I am scared by statistics.	1	2	3	4	5	6	7
I am interested in learning statistics.	1	2	3	4	5	6	7
Statistics involves massive computations.	1	2	3	4	5	6	7
I can learn statistics.	1	2	3	4	5	6	7
I will understand statistics equations.	1	2	3	4	5	6	7
Statistics is irrelevant in my life.	1	2	3	4	5	6	7
Statistics is highly technical.	1	2	3	4	5	6	7
I will find it difficult to understand statistical concepts.	1	2	3	4	5	6	7

	Strongly disagree		Neither disagree nor agree			Strongly agree	
Most people have to learn a new way of thinking to do statistics.	1	2	3	4	5	6	7

Please notice that the labels for each scale on the rest of this page change from item to item.

	Very poorly			Very well			
How well did you do in mathematics courses you have taken in the past?	1	2	3	4	5	6	7

	Very poor			Very good			
How good at mathematics are you?	1	2	3	4	5	6	7

	Not at all			Great deal			
In the field in which you hope to be employed when you finish school, how much will you use statistics?	1	2	3	4	5	6	7

	Not at all confident			Very confident			
How confident are you that you can master introductory statistics material?	1	2	3	4	5	6	7

	Yes		No		Don't know		
Are you required to take this statistics course (or one like it) to complete your degree program?	1	2	3	4	5	6	7

	Not at all likely			Very likely			
If the choice had been yours, how likely is it that you would have chosen to take any course in statistics?	1	2	3	4	5	6	7

DIRECTIONS: For each of the following statements mark the one best response. Notice that the response scale changes on each item. What is your major? If you have a double major, pick the one that best represents your interests.

- | | | |
|--------------------|---------------------------|---------------------------|
| 1. Arts/Humanities | 6. Education | 11. Sociology/Social Work |
| 2. Biology | 7. Engineering | 12. Other |
| 3. Business | 8. Mathematics/Statistics | |
| 4. Chemistry | 9. Nursing | |
| 5. Economics | 10. Psychology | |

Current grade point average (please estimate if you don't know; give only one single numeric response: e.g., 3.52). If you do not yet have a grade point average, please enter 99: _____

For each of the following three items, give one single numeric response (e.g., 26). Please estimate if you don't know exactly.

Number of credit hours earned toward the degree you are currently seeking (don't count this semester): _____

Number of high school mathematics and/or statistics courses completed: _____

Number of college mathematics and/or statistics courses completed (don't count this semester): _____

Number of online courses completed:

Have you been enrolled in an online course before? 1. Yes 2. No

Degree you are currently seeking: 1. Bachelors 2. Masters 3. Doctorate 4. Other

What grade do you expect to receive in this course?

- | | | | |
|-------|-------|--------|-------|
| 1. A+ | 5. B | 9. C- | 13. F |
| 2. A | 6. B- | 10. D+ | |
| 3. A- | 7. C+ | 11. D | |
| 4. B+ | 8. C | 12. D- | |

In order to describe the characteristics of your class as a whole, we need your responses to the following items.

Your sex: 1. Male 2. Female

Your citizenship: 1. US citizen 2. Foreign student 3. Other

Your age (in years): _____

THANKS FOR YOUR HELP!

Post Survey

DIRECTIONS: The statements below are designed to identify your attitudes about statistics. Each item has 7 possible responses. The responses range from 1 (strongly disagree) through 4 (neither disagree nor agree) to 7 (strongly agree). If you have no opinion, choose response 4. Please read each statement. Mark the one response that most clearly represents your degree of agreement or disagreement with that statement. Try not to think too deeply about each response. Record your answer and move quickly to the next item. Please respond to all of the statements.

	Strongly disagree		Neither disagree nor agree			Strongly agree	
I tried to complete all of my statistics assignments.	1	2	3	4	5	6	7
I worked hard in my statistics course.	1	2	3	4	5	6	7
I like statistics.	1	2	3	4	5	6	7
I feel insecure when I have to do statistics problems.	1	2	3	4	5	6	7
I have trouble understanding statistics because of how I think.	1	2	3	4	5	6	7
Statistics formulas are easy to understand.	1	2	3	4	5	6	7
Statistics is worthless.	1	2	3	4	5	6	7
Statistics is a complicated subject.	1	2	3	4	5	6	7
Statistics should be a required part of my professional training.	1	2	3	4	5	6	7
Statistical skills will make me more employable.	1	2	3	4	5	6	7
I have no idea of what's going on in this statistics course.	1	2	3	4	5	6	7
I am interested in being able to communicate statistical information to others.	1	2	3	4	5	6	7
Statistics is not useful to the typical professional.	1	2	3	4	5	6	7
I tried to study hard for every statistics test.	1	2	3	4	5	6	7
I get frustrated going over statistics tests in this course.	1	2	3	4	5	6	7
Statistical thinking is not applicable in my life outside my job.	1	2	3	4	5	6	7
I use statistics in my everyday life	1	2	3	4	5	6	7
I am under stress when I am logged into the course	1	2	3	4	5	6	7
I enjoy taking statistics courses.	1	2	3	4	5	6	7
I am interested in using statistics.	1	2	3	4	5	6	7

	Strongly disagree		Neither disagree nor agree			Strongly agree	
Statistics conclusions are rarely presented in everyday life.	1	2	3	4	5	6	7
Statistics is a subject quickly learned by most people.	1	2	3	4	5	6	7
I am interested in understanding statistical information.	1	2	3	4	5	6	7
Learning statistics requires a great deal of discipline.	1	2	3	4	5	6	7
I will have no application for statistics in my profession.	1	2	3	4	5	6	7
I make a lot of math errors in statistics.	1	2	3	4	5	6	7
I tried to log into the course website two times a week	1	2	3	4	5	6	7
I am scared by statistics.	1	2	3	4	5	6	7
I am interested in learning statistics.	1	2	3	4	5	6	7
Statistics involves massive computations.	1	2	3	4	5	6	7
I can learn statistics.	1	2	3	4	5	6	7
I understand statistics equations.	1	2	3	4	5	6	7
Statistics is irrelevant in my life.	1	2	3	4	5	6	7
Statistics is highly technical.	1	2	3	4	5	6	7
I find it difficult to understand statistical concepts.	1	2	3	4	5	6	7
Most people have to learn a new way of thinking to do statistics.	1	2	3	4	5	6	7

NOTICE that the labels for the scale on each of the following items differ from those used above.

	Very poor				Very good		
How good at mathematics are you?	1	2	3	4	5	6	7
	Not at all				Great deal		
In the field in which you hope to be employed when you finish school, how much will you use statistics?	1	2	3	4	5	6	7

	<u>Not at all confident</u>				<u>Very confident</u>		
How confident are you that you have mastered introductory statistics material?	1	2	3	4	5	6	7

	<u>Not at all</u>				<u>Great deal</u>		
As you complete the remainder of your degree program, how much will you use statistics?	1	2	3	4	5	6	7

	<u>Not at all likely</u>				<u>Very likely</u>		
If you could, how likely is it that you would choose to take another course in statistics?	1	2	3	4	5	6	7

	<u>Very easy</u>				<u>Very difficult</u>		
How difficult for you is the material currently being covered in this course?	1	2	3	4	5	6	7

DIRECTIONS: For each of the following statements mark the one best response. Notice that the response scale changes on each item.

Do you know definitely what grade you will receive in this course?

1. Yes 2. No

What grade do you expect to receive in this course?

1. A+ 5. B 9. C- 13. F
 2. A 6. B- 10. D+
 3. A- 7. C+ 11. D
 4. B+ 8. C 12. D-

In a usual week, how many hours did you spend outside of class studying statistics? Give only one single numeric response that is a whole number _____

	<u>Very low</u>				<u>Very high</u>		
In the past week, how would you describe your overall stress level?	1	2	3	4	5	6	7

THANKS FOR YOUR HELP!

Appendix B

Syllabus

Instructor

Audbjorg Bjornsdottir M.A.
Department of Educational Psychology
Office: Educational Sciences Building 192
Office Hours: Online hours on Mondays from 11 am to 12 pm

Teaching Assistants

Cengiz Zopluoglu
Office: 190 Education science building
Email: zoplu001@umn.edu

Chu-Ting Chung
Office: 140 Education Sciences Building
Email: chung162@umn.edu

TA office hours: The TAs will hold Face-to-Face office hours every Friday in 325 Peik Hall computer lab. Exact times will be announce later

IMPORTANT NOTE: The Educational Psychology computer lab in 325 Peik Hall will be open on Fridays for general student use. This is one of the few labs on campus (in addition to 355 Peik Hall) where you can access the *Fathom* software for EPSY 3264. Exact times will be announced later

COURSE DESCRIPTION

This course is designed to provide an overview of introductory statistics. The topics to be covered in this course include sampling methods, experimental design, data exploration (e.g., using graphical and numerical summaries), data modeling and simulation, normal distributions, sampling distributions, methods of statistical inference (estimation and testing), and correlation. Upon completion of this introductory course, students should be able to: (1) think critically about statistics used in magazines, newspapers, and journal articles, (2) reason about data and (3) apply the knowledge gained in the course to begin to answer simple research questions using empirical data.

This course is intended for undergraduate students who have completed a high school algebra course, but *not previously* studied statistics. Students should also have familiarity with computers and technology (e.g., internet browsing, Microsoft Word, opening/saving files, etc.).

Traditionally, this course is taught using a variety of different methods (e.g., brief lectures, small and large group discussions, activities). Because this version of EPSY 3264 will be conducted entirely online, it is important for all students to keep up with required readings and assignments. Students are also expected to be active participants in this course. Active participation includes asking and answering questions in assigned discussion groups, working with group members to complete various group assignments, and keeping up with other course activities and assignments.

Although attempts have been made to make this course as flexible as possible, there are weekly deadlines that are essential in order to ensure that everyone moves through the course at the same pace and completes the assignments in the same order. **This course is NOT going to be easier than the classroom version of the course and may in fact be MORE WORK than what you would have to do in the classroom (given that more reading and writing is involved in the online course).** Please think very carefully about this as you are attempting to decide if this course is the right fit for you.

TECHNOLOGY POLICY

Because this is a web-based course, it is expected that all students who enroll in the course have internet access and a basic understanding of computer use (e.g., using e-mail, sending attachments via e-mail, using web browsers, using word-processing software such as Microsoft Word). Also, it is expected that all students will regularly check their WebVista e-mail accounts (at least once every 48 hours). **If any changes need to be made in the class schedule, or if special announcements are necessary, the instructor will contact all students via e-mail and announcements in Web vista.**

It is also expected that all students will purchase and use the student-version of *Fathom* on their home computers. This software comes bundled with the course textbook (if purchased in the U of M bookstore), and it will be used extensively throughout the course. It runs on both the PC and the Mac.

To be able to view all components of the course website and download handouts, students should have Java installed on their computers. Students can download free Java software from <http://webct.umn.edu/browser/config-vista.shtml#3>. Students will also need Adobe Acrobat in order to view certain handouts. Links to the Java site and to the Adobe site will be posted on the **Additional Resources** page on WebVista.

REQUIRED MATERIALS

The following materials will need to be purchased:

- Moore, D. S., (2009). *The basic practice of statistics* (5th Ed.). New York: W. H. Freeman and Company
- *Fathom 2 Dynamic Data*TM Software (CD).

You can buy the textbook both online or at the University of Minnesota bookstore just make sure that you are buying the newest version of the book, the 5th edition

You should plan to purchase the software at the University of Minnesota bookstore or online. Make sure you get the correct materials; you just need the student version of *Fathom*. In the past, some students have been able to buy *Fathom* used on eBay at a good price.

You will also need a calculator with a square-root function.

THE WEBVISTA COURSE SITE

There are two ways you can log on to the WebVista course site for EPSY 3264. During the beginning of the semester, I encourage you to spend some time becoming familiar with the course site, and with WebVista (especially if you have never used WebVista before). The WebVista site will become available on September 6th.

Logging on to the site through myU:

- Go to <http://www.myu.umn.edu>
- You will be prompted to enter your U of M username and password. Enter that information and log in.
- Click on the **myU Space** link in the upper right-hand corner. You should then see a list of all the WebVista course sites you have access to. Click on the link for EPSY 3264.

Logging on to the site through WebVista:

- Go to <http://www2.webvista.umn.edu>
- Follow the directions to log in.
- Once you are logged in, click on the link to EPSY 3264.

Once you log in to WebVista, you will see the following homepage for the course site. I am including this here in order to point out the important components of the course site.

The screenshot shows the WebVista course site homepage for EPSY 3264. At the top, there is a red navigation bar with the MYU logo and 'UNIVERSITY OF MINNESOTA' on the left, and 'Accessibility | Help' on the right. Below this, there are tabs for 'Build', 'Teach', and 'Student View', with 'Student View' selected. The main content area has a purple header with 'College of Education and Human Development' and 'EDUCATIONAL PSYCHOLOGY'. Below the header, the course title 'EPSY 3264 Basic and Applied Statistics' is displayed. A sidebar on the left contains 'Course Tools' (Course Content, Mail, Discussions, Calendar, Assessments, Announcements, Assignments, Syllabus) and 'My Tools' (My Grades). The main content area features eight purple icons, each with a corresponding link: Syllabus, Meet your Instructor, Assignments and Resources, Additional Resources, Weekly Schedules, Meet your Teaching Assistant, Lecture Notes & Other Readings, and Discussion assignments & summaries.

The course site is divided into eight different sections.

- **Syllabus:** Go to this page anytime you want to review the syllabus and due dates for the assignments in the course
- **Weekly Schedules:** Here is where you should go each week to find out the scheduled assignments for the week and when important deadlines are. When you click on this link, you will be taken to a new page that has links to every week of the semester. All handouts and individual activities for any given week can be found on that week's page.
- **Meet your Instructor:** Here, you can find information about the instructor.
- **Meet your Teaching Assistant:** Here, you can find information about the teaching assistants
- **Assignments and Resources:** Here, you will find assignments for the course; copies of each lab assignment, graph and article critiques are posted in this folder. You can also find other valuable resources here (e.g., *Fathom* data sets, some *Fathom* tutorials, etc.).
- **Lecture Notes & other readings:** This can be thought of as the central location for lecture notes. Note that there are lecture notes for each week, they are seen as an attempt to introduce material to you, supplement and explain your textbook readings, provide extra examples, answer questions and give you hints about how to complete homework or other assignments. Lecture notes for the whole semester will be available from the first week of the semester
- **Additional Resources:** Here, I will post links to websites with data, links to websites where you can download free software, and other important resources.
- **Discussion assignments & summary:** Here you can find the discussion group list (which group you are in), a copy of each group assignment, summary and answers from group discussions and any material related to the group discussion. Group discussion summaries will be posted by **5.p.m. on Mondays** after each discussion

Note that on the left side of the WebVista page, you see links to other pages. The **Mail** link is where you should go to check your WebVista e-mail and send e-mail through WebVista. If you click on the **Discussions** link, you can get to a page that lists all the different discussion rooms. The **Calendar** link provides you with a calendar of important deadlines, and the **Assessments** link can lead you to links to different quizzes and surveys. Finally, the **My Grades** link will take you to pages where you can assess your progress and your grade in the course.

ASSIGNMENTS AND GRADING

Grades will be determined by (a) participation in small-group discussions, (b) performance on individual lab assignments, (c) performance on critiques (d) performance on quizzes, (e) performance on a midterm exam, (f) performance on a final exam, and (g) peer assessment from your group members. Each of these assignments is discussed in detail below.

Pre-test

Students must complete a pre-test at the beginning of the course to measure their statistical knowledge entering the course. Students will complete the test online, they will have up to 4 hours to complete it. Students will not receive a grade for the pre-test instead 10 points will be granted for those completing the pre-test.

Group Discussion

One important component of this course is **active student participation**. Although the textbook, course notes, and the instructor are important sources of information, so too are your classmates. To facilitate interaction among your classmates, the instructor will break you into small discussion groups at the beginning of the semester, and you will be required to interact with your discussion group frequently throughout the semester. Also, it is possible that minor changes may need to be made in discussion group composition around the beginning of the semester if there are changes in class enrollment. Any changes that are made will be announced to the class via announcements or the WebVista e-mail.

There will be 6 group discussions that you will be required to participate in this semester and these assignments are meant to take the place of activities you might work on in groups if you took this course in a classroom setting.

For each discussion, the group must attempt to talk about a particular topic/concept and answer certain questions about this topic/concept. As you answer these questions, you should not only share your own thoughts about the assignment, but you should respond to what your group members say. As you discuss different assignments, we want to see that you are not only taking time to post your own answers, but to reflect on what your group members have posted and to help each other learn the concepts. For each group discussion assignment, you should elect a group leader who will be willing to summarize the group discussion and submit the group summary via WebVista assignments, (see **instructions for submitting your assignment through course website**) no later than **at Midnight on Friday**. The group leader will receive 1 point of extra credit for compiling the group summary, and ideally, each student in each group should have the opportunity to be a group leader at least one time.

Each group discussion assignment will be worth **10 points**, for a total of 60 points. To receive full credit, each group member must post his/her own thoughts about the group assignment (worth 3 points) and must return twice to the discussion and post two more messages, both of which should be meaningful responses to messages posted by other students in the group (or to questions asked by the instructor or teaching assistant) (worth 2 points each total of 4). Please refer to the handout **Guidelines for Group Discussion** posted in the **Discussion assignments folder** for more information about what constitutes a meaningful response. You will receive 1 point if your group summary is submitted on time and if it contains answers to all the discussion questions in the assignment. You will receive 2 more points for the quality (1 point) and correctness (1 point) of the group summary.

To find out which small discussion group you have been assigned to, please check out the **Discussion Group List** posted in the **Discussion assignments folder (this list will become available before Friday September 9th)**. You can go to the **Discussions** link in the upper right-hand corner of the WebVista page to be taken to the discussion room boards.

Lab Assignments

Students must complete a total of 6 lab assignments this semester, and each assignment will be worth **10 points** (for a total of 60 points). You can find all lab assignments if you go to the **Lab Assignments and Resources** page (see the link for this on the home page of the WebVista site). A PDF and Word copy of each lab assignment is posted, so hopefully, these assignments can be easily downloaded by everyone.

Each of the six labs consists of several homework-type problems that will help you prepare for the quizzes and exams and also teach you the basics of *Fathom*. **Lab assignments will always be due by midnight on Mondays.** For how to submit your labs see **instructions for submitting your assignment through course website** below.

Although you can certainly work on lab assignments with your peers and talk about these assignments with each other, you are each responsible for writing your own solutions and turning in your own individual work. It is not acceptable for two (or more) students to submit the exact same assignment, word-for-word.

IMPORTANT: When you save your lab assignments to send them to the teaching assistant, please save files as Word 97-2003 files if you have Word 2007. This will ensure that the teaching assistant is able to open the files from any computer. If you need help with this, please let the instructor or teaching assistant know.

As a student of statistics, working through the lab problems is an important piece in building a complete understanding of the concepts, as well as allowing you to practice doing statistics. Only by trying to apply the concepts can you be sure that you really understand them. Lab assignments should be regarded as a genuine “learning experience.” We urge you to form study groups to work on these problems and master the concepts. You should, however, be sure that the effort is truly collaborative. The best strategy for completing the assignment is to begin tackling the questions alone, then discussing with others, and finally writing up your answers by yourself. Feel free to consult the teaching assistant and instructor when you are stuck.

Students should read through the **Lab Guidelines** posted on the **Lab Assignments and Resources** page on WebVista before beginning the first assignment in order to learn more about what we expect to see when you submit your assignment (and how to earn the maximum number of points on each assignment). Note that it is very important, as you are working through each assignment, to make sure you consult the different resources we have available for learning how to use *Fathom*. Within each lab assignment, we have attempted to provide detailed instructions for how to use the software, and when necessary, handouts will be posted to help you learn *Fathom*. You are also **STRONGLY ENCOURAGED** to make use of the many *Fathom* movies that you can access through the Help menu when you open the *Fathom* program. You will learn more about accessing these movies very early in the course. There are also some links to some *Fathom* tutorials on the **Lab Assignments and Resources** page.

When lab assignments are graded, the teaching assistant will provide each student with individual feedback. Students are strongly encouraged to contact either the instructor or the teaching assistant if they are having difficulties with lab assignments or if they want to go over any lab problems that were missed. Students are also encouraged to ask for help on lab assignments in their discussion groups or by posting a message in the **Lab Questions** discussion room.

Article Critique

Students will complete one article critique during the semester. The critique will provide students feedback on their ability to effectively understand statistical ideas that are presented in research articles. The article critique is worth 10 points.

Graph Critique

Students will complete one graph critique during the semester. This critique will provide students with feedback on their ability to effectively read a graph presented in the media and evaluate its merit. The graph critique is worth 10 points.


Group tests

Over the course of the semester, three group tests will be administered. The tests are designed to help assess your statistical literacy and reasoning and to provide you with feedback about your learning of important concepts and ideas. The tests are conceptual in nature and are designed to test your ability to apply what you are learning about statistics. Each test will be relatively short and will be worth **20 points** (for a total of 60 points). Tests will consist of multiple-choice and short-answer questions. Everyone in the group should participate in order to get a score on the test. More information regarding the group tests will be sent out 2 weeks before each group test.

Tests will always be available (during test weeks) from Monday at 12 p.m. (noon) until Saturday at midnight.

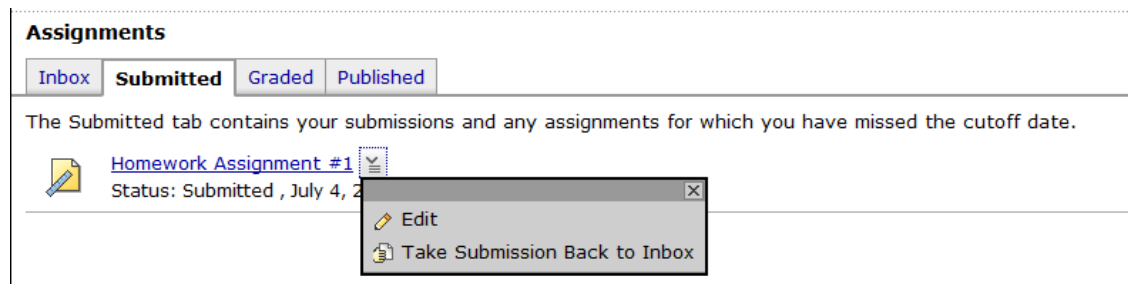
To prepare for each test, students should carefully read through course notes and review group discussion assignments and lab assignments.

Instructions for Submitting Your Assignments Through Course Website:

When you get ready to submit your homework assignment, be sure to first save the assignment as either an .rtf file or a Word 97-2003 file (this is especially important for those students who have Word 2007; we want the homework file to have either an .rtf or a .doc extension, not a .docx extension as you get when you do an automatic save in Word 2007). You should then click on the **Assignments** link  on the left side of the page. Once you click the **Assignments** link, be sure to:

- Click on the link of the assignment name in the **Inbox** area. For example, if you are submitting Homework Assignment #1, click on the link for Homework Assignment #1.
- You will then see a new screen. Under the box labeled “**Submission**” you will see an option to “**Add Attachments.**” Click on this box and then click on the **My Computer** icon that will appear in the new screen that pops up (this will appear on the left side of the screen). Search for the file you want to attach and then double click on it so that it opens and gets attached within the **Assignments** tool. Note there is no need to type in comments in the **Submission** box (or in the **Add Comment** box, unless you have something you want us to know prior to grading your assignment).
- Once you have attached your file, be sure to scroll down to the bottom of the page and click on the “**Submit**” button (and indicate that it is OK to submit the assignment).

- Note that once an assignment is submitted, you can view this assignment by clicking on the **Submitted** tab (within the **Assignments** tool). If you decide you want to edit and re-submit your assignment BEFORE the due date, go the **Submitted** tab, and click on the arrow key next to the name of the assignment. This will result in a drop-down menu like the one you see below. Click on “**Take submission back to Inbox**” so that the assignment returns to the **Inbox**. Then, click on the **Inbox** tab (to go back to the **Inbox**), and from there, you can delete the assignment you submitted and load a newer, revised one. If you ever do this, just remember to hit the **Submit** button after you submit a new assignment so the new assignment is submitted. Once the deadline for submissions has passed, you CANNOT revise your assignments.



- After the assignments are graded, the instructor or teaching assistant will e-mail the entire class (through WebVista) and you can then go back to the **Assignments** link to see your graded assignment (by clicking on the **Graded** tab).

If a student is not able to submit a homework assignment by the appropriate deadline, it is his or her responsibility to notify the instructor BEFORE the assignment is due in order to avoid losing points for submitting the assignment late. It is also the student's responsibility to make sure that homework (and other) assignments are submitted on time. If you see that your assignment appears in the **Submitted** area within the **Assignments** tool, this means the instructor or teaching assistant will be able to view it. If you are uncertain if the assignment was submitted properly, please contact the instructor or teaching assistant for help.

Midterm

Like the quizzes, the midterm exam will help assess your statistical literacy and reasoning. This exam will be worth **35 points** and you will take this exam during Week 9 of the semester. The midterm exam will consist of several multiple-choice and short answer questions designed to test your ability to apply the knowledge you gained by reading the assigned material, working on homework problems and participating in class activities and discussions. You will complete this exam online where and you will be allowed to take up to 4 hours to finish the exam.

Final

The cumulative final exam—like the tests and midterm exam--will help assess your statistical reasoning. This exam will consist of multiple-choice questions and will be worth **35 points**. You will have up to 4 hours to complete this exam online, during the last week of the semester.

Peer assessment

Your grade will also be based on peer assessment from your work in the discussion groups. Each member of your group will provide you with a grade (and a justification for this grade) on a 10-point scale, the grades given to you by all your group members will then be averaged to get the final grade for the peer assessment. So, for example, if you have three group members (in addition to yourself) and they all give you 10 points, you will get 10 points toward your grade on this assignment.

Extra credits

A few extra credits will be offered in the course, these activities mostly comprise of three surveys that you will be asked to complete in exchange for extra credits. These extra credit activities will be available in weeks one, two, 13 and 14 week of the semester. Announcements and email to your U of M email will be sent out to let you know of the extra opportunities.

Individual Assignments

In addition to graded assignments we will collect this semester, there may be some individual, non-graded assignments posted on the web site at various times throughout the semester. The purpose of each of these assignments is to provide you with additional practice and instruction using *Fathom*, and to provide you with more opportunities to reason about statistical concepts and apply the knowledge you are gaining about statistics. Keep in mind that although we will not grade or collect these assignments, working on them will help you better understand different concepts and should help you when it comes to working on your project or on other assessments (e.g., the quizzes, midterm, final, lab assignments, activities and group discussion assignments). We therefore **HIGHLY RECOMMEND** that everyone try to work through these activities in an effort to prepare for various assessments this semester. Answers to each individual activity will be posted at the end of each week, and students are encouraged to discuss these activities with their group members or to post questions or thoughts on these activities in the **General Questions and Answers** discussion room.

There are a total of **290 points** possible in this course.

Pre-test	10
Group Discussions	60
Lab Assignments	60
Group tests	60
Midterm	35

Final Exam	35
Article Critique	10
Graph Critique	10
Peer assessment	10
Total points	290

GRADING

Percentage Cutoff	Grade	Percentage Cutoff	Grade	Percentage Cutoff	Grade
92.5%	A	80.5%	B-	59.5%	D
89.5%	A-	76.5%	C+	Below 59.5%	F
86.5%	B+	72.5%	C		
82.5%	B	69.5%	C-		

IMPORTANT NOTE: A grade of C- is necessary in order to receive a mark of "S."

INCOMPLETES AND LATE WORK

- A grade of "I" (Incomplete) is assigned at the discretion of the instructor when, due to extraordinary circumstances (e.g., illness, hospitalization), a student is prevented from completing the work of the course by the end of the semester. To receive this grade, a written agreement must be completed between the instructor and the student. Notify the instructor as soon as possible if circumstances will prevent you from completing the course by the end of the semester.
- It is expected that all students will complete assignments by the appropriate deadlines. If assignments are turned in late, a total of 10% of the grade on the assignment will be deducted for each day the assignment is late. Contact the instructor ahead of time if there is a problem that will prevent you from turning in an assignment on time.

COURSE OUTLINE

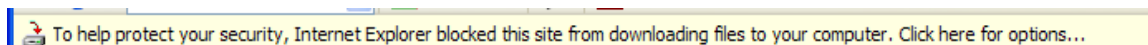
The following outline is an overview of the topics to be covered this semester. The general outline of the course is subject to change at the discretion of the instructor, but any changes will be announced well ahead of time to students.

Week	Date	Topic	Reading	Assignment Due	Extra credits assignments
1	5 -11 Sept	Overview/Introduction to the Course			SATS-36 Pre survey
2	12-18 Sept	Introduction to Modeling and Simulation/Surveys	CH1: p.4-p.19 Fathom video clip	Lab Assignment #1 Pre-test Group Assignment #1	SATS-36 Pre survey
3	19-25 Sept	Random Sampling/Experiments	CH.8 CH.9	Lab Assignment #2	

4	27-2 Oct	Distributions	CH.2	Lab Assignment #3 Group Assignment #2	
5	3-9 Oct	Measures of Center	CH.2	Group test #1	
6	10-16 Oct	Variability	CH.2	Group Assignment #3	
7	17-23 Oct	Reasoning about Variability		Group test #2	
8	24-30 Oct	Scatterplots/Correlation	CH.4	Graph critique	
9	31-6 Nov	Normal Distributions	CH.3	Lab Assignment #4 Midterm	
10	7-13 Nov	Sampling Variability/Sampling Distributions	CH.11	Group Assignment #4	
11	14-20 Nov	More Sampling Distributions		Lab Assignment #5/ Group test #3	
12	21-27 Nov	Introduction to Inference	Supplementary reading at week 12 CH.14	Article Critique Thanksgiving Relax	
13	28-4 Dec	Inference: Confidence Intervals	Supplementary reading at week 13 CH.15	Group Assignment #5	SATS-36 Post survey SPCT survey
14	5-11 Dec	Inference: One-Sample and Two- Sample t-test	CH.17 CH.18	Group Assignment #6	SATS-36 Post survey SPCT survey
15	12-18 Dec	Review		Lab Assignment #6 Peer rating Final exam	

IMPORTANT NOTE ON DOWNLOADING HANDOUTS

To ensure that all students can download different assignments for the course, I have saved most assignments in PDF and Word format. When you click on links for particular assignments, you should see a screen that will ask you if you want to open the assignment or save the assignment (or the assignment should open right away). Sometimes, based on the way WebVista works, you are not automatically prompted to open or save assignments when you click on them. Also, occasionally, you may click on an assignment and see nothing but a blank screen. If this happens, look in the upper left-hand corner of your computer screen. You may see a tool bar that looks like the following:



If you see this tool bar, click on the icon that has the red arrow over it (in the left corner). You will then get a menu like the one below. Click on **Download File**.



When you do this, you may actually be taken out of the WebVista course site and taken back to the main WebVista menu. From the main menu, click again on the link for the 3264 course site, and go back to the page that contains the file you want to download. If you now click again on the link for the file, you should see a menu that will allow you to either open the file or save it to a disk or your hard drive.

If you ever encounter other problems downloading or opening documents, please contact the instructor immediately and try to be as specific as you can about what problem you are experiencing.

A FEW WORDS OF ADVICE

- Many students approach mathematical material with apprehension. It would be dishonest to claim that statistics employs no math, but this course requires only the most elementary mathematics -- arithmetic and very simple algebra. Do not be put off by this minimal math: You can do it!
- It is a bad idea to fall behind in any course, but it is fatal to do so in this course: The course teaches skills and techniques, and the material is cumulative. Log on to course website on webvista regularly and do the readings and assignments on time. If you skip assignments, and cram for the exams, you will almost surely not do well.
- Although you are encouraged to work with other students on the class activities, the lab assignments, critiques and exams must be your own work. Academic dishonesty will be treated very seriously (see section on Scholastic Misconduct). Do not put yourself *and* another student in jeopardy by cheating.
- Do not hesitate to get in touch with the instructors or TAs if you are experiencing problems, need help, or have any questions or other course-related concerns. You can contact any of them via an email message or by coming to office hours.

Only by trying to apply the concepts can you be sure that you really understand them. Lab assignments should be regarded as a genuine “learning experience.” Feel free to consult the teaching assistant and instructors when you are stuck – but try not to ask for more help than you need to get started.

MISSION STATEMENTS

Quantitative Methods in Education (QME)

The Quantitative Methods in Education (QME) track offers educational opportunities in both quantitative and qualitative methods with a broad array of introductory and advanced coursework. Students who choose QME as their track within educational psychology may specialize in any of four areas: *measurement, evaluation, statistics, and statistics education*. The goal of QME is to provide students with broad but rigorous methodological skills so that they may conduct research on methodologies, may help to train others in methodology, or will have the skills necessary to conduct research in related fields.

Psychological Foundations of Education Program Mission Statement

To apply and generate knowledge of psychological processes and methodological procedures involved in learning and teaching for the betterment and improvement of humans in a wide range of situations.

Department of Educational Psychology Mission Statement

Educational psychology involves the study of cognitive, emotional, and social learning processes that underlie education and human development across the lifespan. Research in educational psychology advances scientific knowledge of those processes and their application in diverse educational and community settings. The department provides training in the psychological foundations of education, research methods, and the practice and science of counseling psychology, school psychology, and special education. Faculty and students provide leadership and consultation to the state, the nation, and the international community in each area of educational psychology. The department's scholarship and teaching enhance professional practice in schools and universities, community mental health agencies, business and industrial organizations, early childhood programs, and government agencies. *Adopted by the Dept. of Educational Psychology faculty October 27, 2004.*

College of Education & Human Development Mission Statement

The new College of Education and Human Development is a world leader in discovering, creating, sharing, and applying principles and practices of multiculturalism and multidisciplinary scholarship to advance teaching and learning and to enhance the psychological, physical, and social development of children, youth, and adults across the lifespan in families, organizations, and communities.

UNIVERSITY OF MINNESOTA POLICIES AND PROCEDURES

Diversity: It is the University Policy to provide, on a flexible and individualized basis, reasonable accommodations to students who have disabilities that may affect their ability to participate in course activities or to meet course requirements. Students with disabilities are encouraged to contact me when possible to discuss their individual needs for accommodations.

University Grading Standards

- A achievement that is outstanding relative to the level necessary to meet course requirements.
- B achievement that is significantly above the level necessary to meet course requirements.
- C achievement that meets the course requirements in every respect.
- D achievement that is worthy of credit even though it fails to meet fully the course requirements.
- S achievement that is satisfactory, which is equivalent to a B- or better.

F (or N) Represents failure (or no credit) and signifies that the work was either completed but at a level of achievement that is not worthy of credit, or was not completed and there was no agreement between the instructor and the student that the student would be awarded an I.
I (Incomplete) Assigned at the discretion of the instructor when, due to extraordinary circumstances, e.g., hospitalization, a student is prevented from completing the work of the course on time. *Requires a written agreement between instructor and student.*

Scholastic Misconduct: Academic integrity is essential to a positive teaching and learning environment. All students enrolled in University courses are expected to complete coursework responsibilities with fairness and honesty. Failure to do so by seeking unfair advantage over others or misrepresenting someone else's work as your own, can result in disciplinary action. The University Student Conduct Code defines scholastic dishonesty as follows:

Scholastic Dishonesty. Scholastic dishonesty means plagiarizing; cheating on assignments or examinations; engaging in unauthorized collaboration on academic work; taking, acquiring, or using test materials without faculty permission; submitting false or incomplete records of academic grades, honors, awards, or professional endorsement; or altering, forging, or misusing a University academic record; or fabricating or falsifying of data, research procedures, or data analysis.

Within this course, a student responsibility for scholastic dishonesty can be assigned a penalty up to and including "F" or "N" for the course. If you have any questions regarding the expectations for a specific assignment or exam, ask.

Credits and Workload Expectations: Generally, when a one-credit course is taken, an average of three hours of learning effort per week (over a full semester) is necessary to achieve an average grade. A student taking a three-credit course that meets for three hours a week should expect to spend an additional six hours a week on coursework.

Additional Statements: This publication/material is available in alternative formats upon request. Please contact Psychological Foundations Program, Education Sciences Building 250, 612-624-0042.

The University of Minnesota is an equal opportunity employer and educator.

Appendix C

Correspondence to Students

C-1 Initial Email to Students

Hello,

You are receiving this email because you are enrolled in the online version of EPSY-3264 Basic and applied statistics. I am inviting you to participate in a study I am conducting as part my dissertation research in Statistics education in the Quantitative Methods in Education program in the Department of Educational Psychology at the University of Minnesota.

The study is about exploring the use of different models of collaborative tests in online introductory statistics courses. The research on statistics courses taught online is limited. This research will add needed information about best practice in offering statistics course online. Your help with this is greatly appreciated. By agreeing to participate, you would give me permission to use your test scores and discussion posts that are part of your work in the online EPSY-3264 Basic and Applied Statistics course. I would ask you to complete the following instruments: the Pre- SATS-36 survey at the first week of the semester and at week 13 the post SATS-36 survey and the Students Perception on Collaborative Tests (SPCT) survey.

If you decide to participate in the study, no specific action is needed from you at this point. If you decide not to participate, please reply to this email saying no and you will not be included in the study. Not participating in this study will not affect your grade in EPSY-3264 in anyway. Participation is voluntary.

Attached you will find the consent form for this study, please review it before you decide about your participation in this study.

If you have any questions or concerns regarding this study, please contact me at bjrns001@umn.edu.

Sincerely

Audbjorg Bjornsdottir

Doctoral candidate in Educational Psychology (QME)

Department of Educational Psychology

Room 197 Educational Sciences building

56 East River Road

Minneapolis, MN 55455

C-2 Consent Form

INFORMATION SHEET FOR RESEARCH

Evaluating the use of two different models of collaborative tests in an online introductory statistics course

You are invited to be in a research study where the aim of this study is to explore the impact of using two different formats of collaborative tests in an online statistics course on students learning. You were selected as a possible participant because you are enrolled in the course EPSY-3264 Basic and applied statistics online version. We ask that you read this form and ask any questions you may have before agreeing to be in the study.

This study is being conducted by Audbjorg Bjornsdottir, Department of Educational Psychology at the University of Minnesota.

Procedures:

If you agree to be in this study, we would ask for permission to use your test scores and discussion posts that are part of your assessment in the EPSY-3264 Basic and Applied Statistics course. And for you to complete the following instruments: the Pre- SATS-36 survey at the first week of the semester and at week 13 the post SATS-36 survey and the Students Perception on Collaborative Tests (SPCT) survey.

Confidentiality:

The records of this study will be kept private. In any sort of report we might publish, we will not include any information that will make it possible to identify a subject. Research records will be stored securely and only researchers will have access to the records.

Voluntary Nature of the Study:

Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with the University of Minnesota. If you decide to participate, you are free to not answer any question or withdraw at any time without affecting those relationships.

Contacts and Questions:

The researcher conducting this study is Audbjorg Bjornsdottir. You may ask any questions you have now. If you have questions later, you are encouraged to contact her at Room 250 EdSciB, 56 E River Road, Minneapolis, MN 55455, phone: 612-624-6083, bjrn001@umn.edu. Or her academic advisor Professor Joan B. Garfield, phone: 612-625-0337, email: jbg@umn.edu

If you have any questions or concerns regarding this study and would like to talk to someone other than the researcher(s), you are encouraged to contact the Research Subjects' Advocate Line, D528 Mayo, 420 Delaware St. Southeast, Minneapolis, Minnesota 55455; (612) 625-1650.

You will be given a copy of this information to keep for your records.

C-3 Invite Email for the The Survey Of Attitudes Toward Statistics (Pre-SATS-36)

Hello,

You are receiving this email because you are enrolled in the online version of EPSY-3264 Basic and applied statistics.

Following this email, there is a link to an online survey **The Survey Of Attitudes Toward Statistics (SATS-36)** that asks a variety of questions about your attitude towards statistics. I am asking you to look over the survey and, if you choose to do so, complete the questionnaire. It has 54 items and should take about 10 to 20 minutes to complete.

If you choose to participate, your responses will not be identified with you personally and you will receive five extra credits in the course.

Remember, that your participation in this survey is voluntary. But I hope you will take the time to complete this survey. The survey will be available from September 7th until the 18th 2011.

Sincerely,

Auðbjörg Björnsdóttir

C-4 Invite Email for the The Survey Of Attitudes Toward Statistics (Post-SATS-36)

Hello,

You are receiving this email because you are enrolled in the online version of EPSY-3264 Basic and applied statistics.

Following this email, there is a link to an online survey **The Survey Of Attitudes Toward Statistics (SATS-36)** that asks a variety of questions about your attitude towards statistics. I am asking you to look over the survey and, if you choose to do so, complete the questionnaire. It has 46 items and should take about 10 to 20 minutes to complete.

If you choose to participate, your responses will not be identified with you personally and you will receive five extra credits in the course.

Remember, that your participation in this survey is voluntary. But I hope you will take the time to complete this survey. The survey will be available from November 28th until December 11th 2011.

Sincerely,

Auðbjörg Björnsdóttir

C-5 Invite Email for the Students Perception on Collaborative Tests (SPCT)

Hello,

You are receiving this email because you are enrolled in the online version of EPSY-3264 Basic and applied statistics.

Following this email, there is a link to an online survey called **Students Perception on Collaborative Tests (SPCT)** that asks a variety of questions about your experience taking collaborative tests in this course. I am asking you to look over the survey and, if you choose to do so, complete the questionnaire. It has 20 items and should take about 5 to 10 minutes to complete.

If you choose to participate, your responses will not be identified with you personally and you will receive five extra credits in the course.

Remember, that your participation in this survey is voluntary. But I hope you will take the time to complete this survey. The survey will be available from November 30th until December 11th 2011.

Sincerely,

Auðbjörg Björnsdóttir

C-6 Thank you Email for The Survey Of Attitudes Toward Statistics (SATS-36).

Thank you for taking the **The Survey Of Attitudes Toward Statistics (SATS-36)**.

Forward this email to your instructor bjrns001@umn.edu to receive the 5 extra credits

C-7 Thank you Email for The Students Perception on Collaborative Tests (SPCT)

Thank you for taking the **The Students Perception on Collaborative Tests (SPCT)**.

Forward this email to your instructor bjrns001@umn.edu to receive the 5 extra credits