James Madison University
JMU Scholarly Commons

Senior Honors Projects, 2010-current | Honors College

Spring 2015

The effects of in-class application questions on academic behaviors

Julia Ricotta
James Madison University

Follow this and additional works at: https://commons.lib.jmu.edu/honors201019

 Part of the Applied Behavior Analysis Commons, Educational Methods Commons, and the Higher Education Commons

Recommended Citation
Ricotta, Julia, "The effects of in-class application questions on academic behaviors" (2015). *Senior Honors Projects, 2010-current*. 88.
https://commons.lib.jmu.edu/honors201019/88

This Thesis is brought to you for free and open access by the Honors College at JMU Scholarly Commons. It has been accepted for inclusion in Senior Honors Projects, 2010-current by an authorized administrator of JMU Scholarly Commons. For more information, please contact dc_admin@jmu.edu.

The Effects of In-Class Application Questions on Academic Behaviors

An Honors Program Project Presented to

the Faculty of the Undergraduate

College of Health and Behavioral Studies

James Madison University

by Julia Mary Ricotta

May 2015

Accepted by the faculty of the Department of Psychology, James Madison University, in partial fulfillment of the requirements for the Honors Program.

FACULTY COMMITTEE:

Project Advisor: Tracy Zinn, Ph.D.,
Associate Professor, Psychology

Reader: Bryan Saville, Ph.D.,
Associate Professor, Psychology

Reader: Krisztina Jakobsen, Ph.D.,
Assistant Professor, Psychology

HONORS PROGRAM APPROVAL:

Philip Frana, Ph.D.,
Interim Director, Honors Program

PUBLIC PRESENTATION

This work is accepted for presentation, in part or in full, at the Department of Psychology's symposium on April 20, 2015.

The Effects of In-Class Application Questions on Academic Behaviors

Julia Ricotta
James Madison University

Author Note:

Julia Ricotta, Department of Psychology, James Madison University.

I would like to thank Dr. Zinn, Dr. Saville, and Dr. Jakobsen for all their help with this project. I would also like to thank my classmates for their help in collecting data.

Correspondence concerning this article should be addressed to Julia Ricotta, Department of Psychology, James Madison University, Harrisonburg, VA 22807.

E-mail: ricottjm@dukes.jmu.edu

Table of Contents

List of Figures

Tables

Figures

Abstract

Interteaching is a behavioral method of teaching college courses, where students take a more active approach to learning. The current study manipulated interteaching preparation guides and studied the effects on exam scores, attendance, duration of discussion, and ratings of the discussions. Both groups received the same materials in different formats. The control group completed full prep guides at home, which included factual and application questions. The intervention group completed revised prep guides at home, which included factual questions, and application questions in class. During discussions, the control group discussed answers to the full prep guides, and the intervention group composed answers to the in-class application questions. There were significant differences between the control and intervention groups on duration of the discussions. The two groups were not significantly different on exam scores, ratings of discussions, or number of absences.

The Effects of In-Class Application Questions on Academic Behaviors

Many undergraduate courses revolve around lectures. Although lectures are an excellent way to pass on information, students may become passive members of the classroom (Smith & Cardaciotto, 2011). As Boyce and Hineline (2002) mentioned in their study of interteaching, the goal of lectures is to convey information rather than to teach skills. Lectures do not give students time to practice the new skills they've been taught, which reinforces their role as a passive learner (Smith & Cardaciotto, 2011). Active learning is a method to increase student engagement and facilitate changes in student behavior (Boyce & Hineline, 2002). Active learning can help students develop the skills they will need to be successful after college.

Researchers have studied how passive learning affects students' understanding of course material through Bloom's taxonomy. Bloom's taxonomy is a way to conceptualize the different levels of learning behaviors. It was developed as a continuum of educational objectives with six different dimensions, which include knowledge, comprehension, application, analysis, synthesis, and evaluation (Krathwohl, 2002). These concepts range from concrete to more abstract and complex categories. Bloom's taxonomy has been revised to include knowledge and cognitive dimensions, with six similar objectives: remember, understand, apply, analyze, evaluate, and create (Krathwohl, 2002). According to a study by Kunen, Cohen, and Solman (1981), many students are only reaching the knowledge level of Bloom's taxonomy, associated with minimum understanding and the lowest recall abilities. Most educators should have the goal of reaching these higher levels of understanding. More research is needed to discover how educators can promote higher levels of thinking. In a study by Kunen et al. (1981), the students with the highest understanding and recall abilities were those receiving synthesis prestudy questions. These students were becoming active participants in the learning process by reading information before

class and coming to class with an understanding of the material. Andre (1979) also found that higher order questions can promote learning.

Active learning includes many methods that help improve the current level of student engagement and critical thinking skills (Mathie et al., 1993). Active learning encourages the active participation of every student in the class. It includes methods such as discussions, demonstrations, debates, or interactive lectures (Mathie et al., 1993). Active learning allows students to reach higher levels of Bloom's taxonomy by having students practice synthesizing, evaluating, and analyzing material. Instructors have many methods of active learning to choose from, including team-based learning, collaborative learning, just-in-time teaching, personalized system of instruction, and interteaching. Each of these methods make use of different techniques and principles.

Team-based learning. Teachers have different active learning approaches to consider. One method of active learning is a team-based approach. Michaelson developed this model for a business class, but it has shown to be effective in a variety of courses. The instructor divides students into permanent teams to complete in-class activities, which require synthesis of complex concepts (Michaelson & Sweet, 2008). Michaelson and Sweet (2008) developed these assignments with regard to the "4 S's" criteria: (1) assignments should be significant to students, (2) all students should be working on the same problem, (3) students should be required to make a specific choice, and (4) groups should simultaneously report their answer choices. These teams are intended to develop throughout the semester as team members grow more comfortable with one another (Michaelson & Sweet, 2008). In addition, participation in the group counts for a part of final grades to hold students accountable.

A key component of team-based learning is the Readiness Assessment Process (RAP), where students read the text and complete the Individual Readiness Assessment Test (iRAT). These tests include multiple choice questions that ensure students understand the material before moving on to application components. In class, they complete the assessment again in groups, called the Team Readiness Assessment Test (tRAT). As a team, they use an Immediate Feedback Assessment Technique (IF-AT) card, which provides immediate feedback as students scratch off their answers (Michaelson & Sweet, 2008). Clair and Chihara (2012) adapted the traditional team-based approach for a statistics class, administering the RAP at the end of each unit. In addition, they included questions with the "4 S's." The researchers found instructor and student enjoyment to be high, but they did not assess student achievement. Webb (1985) researched learning in small groups, where students are given a set of problems to work on together in class, and composed a set of guidelines for team based learning. Webb found that sets of problems should include teacher feedback, such as an explanation to why an answer was incorrect, in order to be the most effective. Webb also found that pairs were the most effective team size.

Collaborative learning. Collaborative learning is the process of working toward common academic goals in a small group (Bruffee, 1999). Key elements of collaborative learning include discussions, evaluating other students' ideas, and clarification of concepts (Gokhale, 1995). Gokhale (1995) found collaborative learning increased critical thinking skills in a critical thinking test developed by the researcher.

There are many models of collaborative learning, including peer learning, consensus group work, or collaborative project work (Bruffee, 1999). Peer learning is a type of collaborative learning that focuses on learning from and with other students without direct intervention from the instructor (Boud, Cohen, & Sampson, 1999). Students act as both teachers

and students in reciprocal peer learning (Boud et al., 1999). Peer learning helps to develop a range of skills, including teamwork, critical thinking, and communication skills. In addition, students often learn how they can manage their time more effectively (Boud et al, 1999).

Just-in-time teaching. Another teaching method is just-in-time teaching, which is "anywhere, anytime" learning that is individualized to specific learners (Bradenberg & Ellinger, 2003). Just-in-time teaching is an active model of learning because it is based on the assumption that students and workers are driven to learn as much as they can. Just-in-time teaching revolves around web-based warm-ups available before class (Novak, 2011). Students complete these assignments based on their current knowledge, and instructors use this information as they teach the lesson, keeping in mind students' current understanding. The warm-ups serve as a way for instructors to measure student knowledge (Novak, 2011). Technology allows resources to be accessed anywhere, and materials are provided preemptively (Bradenberg & Ellinger, 2003). Students are involved in their learning, shown from their use of extra resources and their application of concepts to their knowledge foundation (Novak, 2011).

Just-in-time teaching has been shown to improve academic performance, especially when paired with collaborative learning (Crouch & Mazur, 2011). In addition, Novak (2011) found that students' attendance dramatically increased with this method, leading to higher grades.

Personalized system of instruction. A behavior-based strategy that promotes active learning is Keller's (1968) personalized system of instruction (PSI). From a behavioral stance, professors should utilize active learning. From a behavioral perspective, learning is a change in behavior; therefore, teachers should encourage this change instead of simply passing on information (Boyce & Hineline, 2002). The typical classroom set up requires students to manage themselves with only long-term contingencies, such as midterm and final exams, in place. In

addition, the lecture set-up discourages student participation (Boyce & Hineline, 2002). PSI has five main components, including an emphasis on learning at one's own pace and mastery of a unit before moving on to new material (Keller, 1968). Instructors also use lectures as reinforcers so students are motivated to study more. Classes are used to clarify confusion and motivate students to learn. Students who master the course content may be recruited to be undergraduate teaching assistants for following semesters, which might also encourage students to learn the material.

PSI also emphasizes the written word through quizzes, reading materials, and professor-student interactions (Buskist, Cush, & DeGrandpre, 1991). This method has shown to be effective in increasing student's understanding of a subject; however, its use has declined in recent years (Eyre, 2007). One of the reasons for the decline is that PSI requires more work for the instructors as they develop course materials, train teaching assistants, and grade assignments (Eyre, 2007). Today, computer-aided personalized system of instruction (CAPSI) allows instructors to easily monitor student mastery and progress (Eyre, 2007). As a result, this method of active learning has become easier to implement.

Interteaching. Interteaching is another behavioral method of teaching that promotes student engagement and active learning behaviors. Interteaching is a "mutually probing, mutually informing conversation between two people" (Boyce & Hineline, 2002). This method has proven to be more effective than lecture-based classrooms, because it sets up contingencies that force students to prepare and study for class (Saville, Zinn, Neef, VanNorman, & Ferreri, 2006). Instructors assign preparation guides (prep guides) as homework to be completed before class. These prep guides serve as reading guides, including definition questions as well as critical thinking and application questions. In class, students engage in pair discussions, focusing on

understanding the material and overcoming challenging concepts. The instructor and teaching assistants facilitate these discussions and answer any questions. After their discussions, students complete a record sheet, indicating difficulties and suggesting concepts to review. At the beginning of the next class, the instructor prepares a clarifying lecture on difficult concepts indicated by the record sheets.

Like PSI, interteaching is based on behavioral principles. Students receive points for pair discussions, setting up a social contingency so students are encouraged to attend class (Boyce & Hineline, 2002). In addition to coming to class, students must engage in a discussion with only one other person. A pair discussion sets up a natural social contingency where partners must attend to each other and contribute to the discussion (Boyce & Hineline, 2002). In addition, students provide a self-rating on the quality of their pair discussion, which serves as a social contingency so students are motivated to prepare prep guides ahead of time. Teachers also prepare clarifying lectures to supplement interteaching discussions. Lectures are focused on difficult concepts that students will need to know for exams. Therefore, lectures become reinforcing consequences of reading instead of antecedents. In addition, students can receive quality points on exams if both partners perform well on the material relating to their pair discussions.

Previous research on interteaching. Studies have shown that interteaching is an effective method in terms of academic performance; however, there is still more to learn about this approach to teaching. Saville, Lambert, and Robinson (2011) suggested that researchers conduct component analyses to discover what exactly makes interteaching effective. An unpublished study compared interteaching with and without prep guides in a lab setting. The study found that the students who completed the prep guides scored significantly better on a quiz

one week later compared to students who were instructed to take notes. Goto and Schneider (2010) looked at prep guides as a critical aspect of the interteaching process, emphasizing that prep guides should include synthesis questions. According to students' self-ratings, these synthesis questions fostered critical thinking. Cannella-Malone, Axe, and Parker (2009) manipulated the format of prep guides. One group received a prep guide constructed by an instructor, and the other group read through chapters, writing their own questions and constructing their own prep guides. These researchers did not find significant differences between the two groups, suggesting that discussions may be a more crucial component of interteaching.

Researchers have also tried to improve pair discussions through quality points. Saville and Zinn (2009) investigated whether quality points affected test performance, where points were distributed depending on how well partners performed on an exam. For example, Partner A and Partner B worked together on a prep guide, and an exam essay focused on material from that prep guide. If both Partner A and Partner B received an A or a B on the essay, points were added to each of their exams. Quality points were intended to motivate students to have thorough discussions and engage in on-task behaviors during in-class discussions; however, they were found not to have a significant effect on exam scores. Saville and Zinn (2009) believed this result may be due to a delay in reinforcement because of the lengthy time interval between pair discussions and exams and more immediate consequences.

The current study. The current study focused on manipulating the length of prep guides and the presentation of application questions. The control group completed the full prep guides at home. For the intervention group, students completed a shorter, revised prep guide at home that served as a reading guide with questions about basic concepts and definitions. In class, students

in the intervention group completed application questions in pairs during their discussions. Both groups received identical questions, but at different times. By adding in-class application questions, students in the intervention group constructed answers to these questions during discussions. I hypothesized that this may have focused student discussions on relevant examples by creating a contingency so students have an assignment to complete in class. Boyce and Hineline (2002) manipulated when students received prep guides, distributing them on the day of the scheduled interteach session. As a result, prep guides were used as an outline for discussions. Saville, Lambert, and Robinson (2011) have found longer prep guides to be aversive to students. As a result, they often distribute additional practice problems to complete in-class, so students have longer discussions and receive additional practice.

As noted previously, the current study focused on the length of the prep guides and presentation of the application questions. I compared the exam scores, time on discussions, and self-reported ratings of quality of pair discussions of the control and intervention groups to see if the intervention led to a difference in the understanding of course material. I also looked at the number of missed days across sections.

The purpose of this intervention was to manipulate prep guides and the behaviors students are engaging in during discussions. The control group, students using the full prep guides, discussed their answers to questions they had already completed at home. The intervention group constructed answers together during discussions as they worked through the in-class application questions. With the intervention group, the full prep guides were split into the revised at-home prep guide and the in-class application questions. As a result, students in the intervention group had less work to finish at home, which may make them more likely to come to class prepared.

I hypothesized that manipulating prep guides would influence the behaviors of students before and during pair discussions. The quality of student discussions may be affected by presenting the application questions in class. Instead of discussing, students would work in pairs to construct their answers. This is similar to the in-class activities in collaborative and team-based learning (Michaelson & Sweet, 2008). I also hypothesized that the control and intervention groups would score differently on exams, because of the different discussion behaviors. Although the in-class application questions were not graded, I was interested to see if the application questions would affect student understanding and exam performance. Barnett and Francis (2012) found that quizzes with questions that required higher-order thinking resulted in significantly better test performance than quizzes with factual questions. The researchers hypothesized that higher order thinking questions may result in deeper thinking about the material and increased familiarity, as students reviewed and rethought the material (Barnett & Francis, 2012). The factual questions may have led to the use of rote memory or a reliance on the book.

In addition, I predicted that the two classes would spend different amounts of time on pair discussions. By presenting application questions in class, students may have longer, more in-depth discussions as they solve problems. Alternatively, students may rush through their in-class assignment so they could leave. Saville (2013) discussed 10 tips for implementing interteaching, finding that some students may see the in-class discussions as a way to leave class early. Saville (2013) suggests presenting supplementary materials during the discussion and using higher-level questions on prep guides to facilitate discussions. The current study addressed this idea.

I also hypothesized that the control and intervention groups will have different ratings of in-class discussions. The intervention group answered application questions in class, and the

control group answered these questions at home and discussed their answers in class. As a result, the discussions may be different qualities, shown through differences in self-ratings. Number of absences may also differ across sections. Students were evaluated on whether they completed the assignment at home. As a result, students in the intervention group may be more likely to come to class, because they had less work to complete at home. The two classes may have been differentially motivated to come to class.

Method

Participants

The current study used two class sections of a psychology statistics course, Psychological Statistics and Measurements, taught by Dr. Tracy Zinn. Students were declared psychology majors, and this course was their first required course in the major. Students in both sections learned the same material. One class was taught from 8:00 to 9:10 AM with 19 students (Section 1: 18 women, 1 man), and the other class was taught from 9:50 to 11:00 AM with 21 students (Section 2: 14 women, 7 men). Section 1 had a wider spread of year, with 10 sophomores, five juniors, and three seniors (average age = 19.83). Section 2 had 12 sophomores and seven juniors (average age = 19.74). Section 1 had three transfer students, and Section 2 had nine transfer students.

Materials

Prep guides. Questions from the prep guides were taken from a previous semester's course prep guides, which included definitions, factual questions, and application questions. In the control condition, students completed full prep guides. For the intervention condition, I split these prep guides into two different activities. The prep guide completed at home included definitions and factual questions (revised prep guides), and the application questions were

completed in class. These application questions included questions that related to examples. Many times application questions required calculations and interpretations and applied to real examples or studies. In general, application questions required a deeper understanding of the concepts. Both groups were receiving identical questions just at different times. Appendix A shows an example of a full prep guide and how we edited these prep guides into two different activities. The length of these application questions varied depending on each prep guide.

Record sheets. To measure the quality of discussions, I used self-reported ratings and duration of discussion sessions, obtained from record sheets. I measured these variables across units, which corresponded to the prep guides for each test. On record sheets, students recorded the time they completed their discussions and their partner's name. Teaching assistants recorded the time discussions started and calculated the length of discussions for each pair.

Students also rated their agreement with six statements regarding the quality of discussions and their (and their partner's) preparation for the discussion. Statements included "I was prepared for the discussion," "My partner was prepared for the discussion," "We discussed all questions thoroughly," "We were able to explain difficult concepts without looking at the book." A full record sheet is included in Appendix B. Students rated their agreement with the statements on a Likert scale, from strongly disagree (1) to strongly agree (5). I used a Chronbach's alpha to analyze the reliability of the six questions comprising the self-rating measure, $\alpha = .854$. As a result, I analyzed the composite quality score based on the average of these six responses. I called this students' self-ratings of discussion, because it was based on their own perceptions of the quality of their individual discussions.

Students also indicated which concepts were difficult to understand, as well as which concepts they thought they understood well. The record sheets also served as a way to measure attendance.

Exams. I used exam scores to measure students' understanding of the material, including unit exams. There were five unit exams; however, most of the exams included some material covered initially in previous units. Each exam consisted of more points than the previous exam. Exams consisted of one to two essay questions, approximately five short answer questions with multiple parts, and approximately 10-15 multiple-choice questions. The exact number of each of these questions varied with each exam. The essay questions revolved around important concepts from articles or overarching themes of the unit. The short answer questions involved hypothesis testing and interpreting results, including effect size, power, and confidence intervals. The multiple choice questions focused on concepts from the unit. All of this material was covered in the prep guides. Students had opportunities to complete test corrections to earn back half of the points that they missed. Only students' raw scores, without corrections, were included in this study.

Final exam. The final exam, or the sixth exam, was cumulative and consisted of all multiple choice questions. The final exam consisted of mainly application questions. Many questions on the final exam gave a real scenario, or study, and asked which test statistic would be most appropriate to implement.

Demographic questionnaire. Before students completed the final exam, they were asked to complete a demographic questionnaire, which included questions about gender, grade point average (GPA), major/minor, and age. Students also indicated if they were transfer students.

Procedure

Both class sections began the semester with the full prep guides to measure initial differences across sections. Section 2 was randomly assigned to the intervention condition. After the first test, Section 1 continued to complete the full prep guides at home, and Section 2 completed the shorter revised prep guides at home with in-class application questions. Students' understanding of the material was measured through five exams and a cumulative final. Exam 5 and the cumulative final were taken together, but analyzed separately. I also analyzed the cumulative points for each class.

Section 1. After the first exam, Section 1 continued to complete the full prep guides at home. Prep guides were available online on Canvas, a course management platform accessible online. In class, students participated in the clarifying lecture and discussed the whole prep guide, focusing on challenging questions or difficult concepts. The instructor and teaching assistants facilitated the discussion by answering questions. Students turned in a printed record sheet at the end of class, which ensured they got credit for coming to class. Section 1is referred to as the "Long PG condition."

Section 2. Section 2 started the intervention after the first exam. Section 2 completed the shorter revised prep guides at home, which were also available on Canvas. In class, students would participate in the clarifying lecture. During discussions, students would work together on application questions in class. These application questions were printed handouts, but they would be available on Canvas after the in-class discussion. The instructor and teaching assistants would be available to answer any questions during the discussion. Students would also turn in printed record sheets for attendance purposes. Section 2 is referred to as the "Short PG condition." Table 1 shows a table of this experimental design.

Results

Comparison of two sections. I used the first exam as a way to control for differences across groups. The Long PG and Short PG conditions were not significantly different, $t(38) = .674$, $p = .504$, CI_{95} [-.04, .08]. Group means and standard deviations for exams 1 through 5 are found in Table 2. Therefore, exam 1 grades were not used as a control variable for the remaining analyses. In addition, I analyzed grade point average (GPA) across the sections, finding no significant differences, $t(31) = -1.005$, $p = .323$, [-.37, .12]. From this information, I was able to determine that our groups were relatively equal before the intervention in terms of demographics.

Exam scores. I analyzed differences in exam scores across sections using *t* tests for each exam. Contrary to my hypothesis, both classes scored similarly on all six exams (all ps > .27). I also analyzed differences in total points of exam 2 through exam 5, finding no significant differences between the sections, $t(28.55) = 1.303$, $p = .203$, $d = .416$. I eliminated low outliers from the sections, because some students did not complete all of the exams, either for health reasons or because they dropped the class. Table 2 shows the means, standard deviations, and confidence intervals for each exam and the total points. Figure 1 illustrates this trend.

In addition, I ran *t* tests to look at differences in demographic information, such as gender and transfer status, on total points. I found no significant gender differences on total points, $t(37) = .007$, $p = .994$, [-.11, .11]. I also compared transfer and native students on total points. Transfer students' total points (.70, .16) were marginally lower than those of native students (.70, .09), $t(14.60) = 1.822$, $p = .089$, [-.02, .20]. Therefore, I subsequently ran an ANCOVA with transfer status as the covariate. Transfer was a significant correlate, $F(1, 36) = 3.714$, $p = .062$, and PG condition was not a significant factor, $F(1, 36) = .130$, $p = .721$.

Duration of pair discussions. As hypothesized, I found a significant difference of discussion time across sections (Long PG and Short PG conditions) and units (5 exams) using a factorial ANOVA. There was a main effect of units on discussion times, $F(4, 674) = 12.744$, $p < .001$, $\square\square^2 = .093$. In addition, there was a main effect of the class section on discussion times. The Long PG condition had longer in-class discussions than the Short PG condition, $F(1, 674) = 16.986$, $p < .001$, $\square^2 = .018$. There was also an interaction between sections and units on discussion times, $F(4, 674) = 2.977$, $p = .019$, $\square\square^2 = .019$. The length of discussion duration for either section depended on the unit and the difficulty of the material. Using a Bonferroni correction, I found there was a significant difference between the Long PG and Short PG conditions on Unit 1, $t(122.75) = 3.188$, $p = .002$, CI_{95} [2.06, 8.79], d = .575. These results show the Long PG condition spending more time on discussions than the Short PG condition. Table 3 shows a table with the means, standard deviations, and confidence intervals of each section and unit on minutes spent discussing; Figure 2 illustrates this trend.

Self-ratings. There were no significant differences in the self-reported ratings of pair discussions across the Long PG ($M = 4.50$, $SD = .56$) and Short PG ($M = 4.53$, $SD = .53$) conditions. I found no significant differences on self-ratings when using a factorial ANOVA with units and section as our grouping variables (all ps > .10). There was also no interaction between units and sections, $F(4, 725) = .882$, $p = .474$.

Absences. I also measured the number of missed days across sections. The Long PG condition ($M = 2.05$, $SD = 1.70$) and Short PG condition ($M = 2.10$, $SD = 1.55$) showed no significant differences in number of missed days throughout the semester, $t(37.67) = -.097$, $p = .923$, CI_{95} [-1.09, .99].

Discussion

These results provide some evidence that completing the application questions in class may not influence student understanding of the material or the quality of the pair discussions. Presenting the application questions in class did not significantly affect the behaviors of the Short PG condition, such as attendance, or their own perceptions of their understanding, as shown by their self-ratings. In addition, the intervention did not seem to significantly affect their exam scores. These results illustrate that the two sections of the class engaged in similar academic behaviors regardless of the intervention.

Exam scores. To ensure that the two groups were equal before the intervention, both sections experienced interteaching with the full prep guides for the first unit. The two groups scored about two points differently on the first exam, which was not a significant difference. The two sections continued to score similarly on tests, with no significant differences between class sections on the five unit tests or the cumulative final. Though the sections were not significantly different on total points, there was a large effect size, possibly indicating there was not enough power in the study to see the differences.

Though results were not statistically significant, the analysis shows the Short PG condition scored lower on the second, fourth, and fifth exams, as well as total points. In Figure 1, the reader can see a visual depiction of this trend on the exam graph. Although this difference was not significant, I wanted to include possible explanations for this consistent trend. Some of the in-class application activities may have been too long for the amount of time given. As a result, some of the students may not have finished all the problems in class. There was no contingency to finish these questions outside of class, so students may not have worked through these problems after class. Similarly, it may be more beneficial to have more repetitions of the

material by completing the whole prep guide at home and then discussing the whole prep guide in class. Many students in the Short PG condition did not discuss the concepts on the prep guide, but focused solely on the in-class questions. As a result, students in the Short PG condition may not have discussed all concepts in the prep guide. Finally, students with the full prep guides had to "work through" the harder application questions. As a result, they may have resolved their problems by looking for answers in the book and grown more familiar with the material. This may have led to a small increase in knowledge in the Long PG condition.

Discussion duration. I found significant differences in discussion times across class sections and units. On average, there was a main effect of unit on discussion time and section on discussion time. The main effect of the different units is reasonable, because the material for some units was more challenging than other units. This affected the length of lectures and the time allocated to pair discussions. For example, in Unit 2, the concepts of power and effect size were introduced. Students found these concepts challenging and asked more questions, resulting in a longer lecture and less time for pair discussions. There was also a main effect of section on discussion duration with the Long PG condition having longer discussions.

There was also an interaction of section and unit on the amount of time spent on discussions, illustrated in Figure 2. The duration of discussions of each class section depended on the difficulty of the material in the unit. The Long PG condition spent more time on discussions than the Short PG condition, which may be influenced by a variety of variables. For example, the Long PG condition may have had more material to discuss because they had already completed the full prep guide at home. The Bonferroni correction found this interaction was significant in Unit 1, before the intervention started. As a result, it is unclear if the intervention impacted the duration of discussions.

It is important to note that time on task may differ from the actual length of the discussions. The Long PG condition's class started at 8:00 AM, which could have influenced the alertness of the class. Though the Long PG condition had longer discussions, they may not have engaged in on-task behaviors for the full time recorded. The in-class application questions may have impacted the Short PG condition's discussion times. Some of the prep guides had shorter sets of application questions. The Short PG condition may have only completed in-class questions and left immediately after they finished.

Self-ratings. There were no significant differences between sections for self-ratings. The similar self-ratings could be due to the nature of grading one's own understanding. Students from both sections may rate themselves highly if they are nervous it will factor into their grade. Also, it may be possible that students tend to rate themselves higher when using Likert scales. Boyce and Hineline (2002) had students discuss and determine a rating of their discussion from 1 to 10; they also had to justify why they deserved this grade. Future research could include justifying the rating in order to understand the true quality of the discussions. In addition, students may not know what constitutes a "good discussion." Future studies can provide an example of a high quality discussion and even create a discussion rubric (Saville, 2013).

Attendance. In addition, we recorded attendance to see if either section was more likely to come to class. There were also no significant differences across sections. At first, I determined that the application questions should only be available in class; however, I realized that this may change the availability and exposure of material for the two groups, which was not the purpose of the experiment. As a result, I posted the application questions online after the pair discussions. Because all of the material was available regardless of attendance, this may have led to the similarity in number of missed days across class sections.

Possible explanations. I was examining the effects of the manipulation of the prep guides and timing of the application questions on discussions and other academic behaviors. The application questions were intended to shape discussion behaviors to see if there was a difference in understanding when students constructed answers together in class or discussed the full prep guides which they completed at home. As Saville and Zinn (2009) concluded, the immediate social consequences of the discussions are critical. In-class application questions may not be necessary to facilitate on-task behaviors during discussions; however, the discussions themselves are critical so students can explain and grow familiar with the material. Future research could measure the effectiveness of discussions by eliminating the discussion component for one class section and comparing academic performance to a control section.

Application questions affected student understanding when these questions were included in quizzes (Nguyen & McDaniel, 2015). Application questions may only be effective when there is a contingency for completing them accurately. Saville, Pope, Lovaas, and Williams (2012) studied the testing effect in an interteaching classroom. They found that post-discussion quizzes did not affect students' exam performance. They hypothesized this result may be due to the nature of interteaching, which already contains methods to improve learning (Saville et al., 2012). Similarly, application questions may not be necessary when other active learning strategies are in place.

From the results, including application questions in class rather than before class may not enhance student learning. With interteaching, students are already completing prep guides at home and discussing difficult concepts in class. This in-class component may not be needed. Research has shown that manipulating the composition of prep guides results in minimal nonsignificant differences (Cannella, Axe, & Parker, 2009). In addition, I may be able to

conclude that the presentation of the application questions is not important, simply including application questions in prep guides may be enough.

Future directions. Future research can focus on improving the quality of pair discussions. Researchers could set up a contingency to ensure that students completed all of the in-class application questions, similar to adding frequent quizzing in the testing effect (Lambert & Saville, 2012). To further examine how the application questions affect the quality of discussions, observers could measure the frequency of certain on-task behaviors in class, such as active listening or asking clarifying questions.

I am replicating this study this semester with a new group of students. I reversed the intervention so the 8:00 AM section is receiving the in-class application questions. I have made minor changes to the procedures to account for the possible limitations of the current study. For example, I am holding the application questions until the pairs have discussed their completed prep guides. This allows the discussion to focus on concepts as well as practice problems. This is to increase the repetitions in the intervention group so they are receiving similar exposure to the material as the control group.

Composing and distributing these application questions were somewhat time-consuming. We made copies of the in-class application questions, which required more preparation and money. The researchers even received a small grant from the department to make all the copies for every class. These results suggest teachers may not need this extra component. Alternatively, the results show that the Short PG condition had shorter prep guides and spent less time in discussions, yet they still earned similar grades on the exams as the Long PG condition. Further research is needed to replicate these results and measure student enjoyment. If both of these

methods produce similar academic results, it may be necessary to look at student and instructor satisfaction to determine the method instructors choose to use.

Another future consideration could include additional practice problems. It is possible that the application questions did not promote critical thinking. A future study could design questions that make use of the "4 S's" found in team-based learning (Michaelson & Sweet, 2008). This criteria was developed to promote critical thinking when constructing in-class assignments. Further component analyses are needed to discover what makes interteaching effective and how to make the discussion component more productive.

References

Andre, T. (1979). Does answering higher-level questions while reading facilitate productive learning? *Review of Educational Research, 49*(2), 280-318.

Barnett, J. E., & Francis, A. L. (2012). Using higher order thinking questions to foster critical thinking: A classroom study. *Educational Psychology, 32*(2), 201-211.

Boud, D., Cohen, R., & Sampson, J. (1999). Peer learning and assessment. *Assessment & Evaluation in Higher Education, 24*(4), 413-426.

Boyce, T. E., & Hineline, P. N. (2002). Interteaching: A strategy for enhancing the user-friendliness of behavioral arrangements in the college classroom. *The Behavior Analyst, 25*(2), 215-225. Retrieved from RIS Format UTF-8 database.

Brandenburg, D. C., & Ellinger, A. D. (2003). The future: Just-in-time learning expectations and potential implications for human resource development. *Advances in Developing Human Resources, 5*(3), 308-320. doi:10.1177/1523422303254629

Bruffee, K. A. (1999). *Cooperative learning* (2nd ed.). Baltimore, MD: The Johns Hopkins University Press.

Buskist, W., Cush, D., & DeGrandpre, R. J. (1991). The life and times of PSI. *Journal of Behavioral Education, 1*(2), 215-234. doi:10.1007/BF00957005

Cannella-Malone, H., Axe, J. B., & Parker, E. D. (2009). Interteach preparation: A comparison of the effects of answering versus generating study guide questions on quiz scores. *Journal of the Scholarship of Teaching & Learning, 9*(2), 22-35. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,url,cpid,uid&custid=s8863137&db=ehh&AN=45408843&site=eds-live&scope=site

Clair, K. S., & Chihara, L. (2012). Team-based learning in a statistical literacy class. *Journal of Statistics Education, 20*(1). Retrieved from www.amstat.org/publications/jse/v20n1/chihara.pdf

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

Eyre, H. L. (2007). Keller's personalized system of instruction: Was it a fleeting fancy or is there a revival on the horizon? *The Behavior Analyst Today, 8*(3), 317-324.

Gokhale, A. A. (1995). Collaborative learning enhances critical thinking. Journal of Technology Education, 7(1). Retrieved from http://scholar.lib.vt.edu/ejournals/JTE/v7n1/gokhale.jte-v7n1.html?ref=Sawos.Org

Goto, K., & Schneider, J. (2010). Learning through teaching: Challenges and opportunities in facilitating student learning in food science and nutrition by using the interteaching approach. *Journal of Food Science Education, 9*(1), 31-35. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,url,cpid,uid&custid=s8863137&db=eric&AN=EJ867443&site=eds-live&scope=site; http://dx.doi.org/10.1111/j.1541-4329.2009.00087.x

Keller, F. S. (1968). Good-bye teacher. *Journal of Applied Behavior Analysis, 1*(1), 79-89.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212.

Kunen, S., Cohen, R., & Solman, R. (1981). A levels-of-processing analysis of Bloom's taxonomy. *Journal of Educational Psychology, 73*(2), 202-211. doi:10.1037/0022-0663.73.2.202

Lambert, T., & Saville, B. K. (2012). Interteaching and the testing effect: A preliminary analysis. *Teaching of Psychology, 39*(3), 194-198. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,url,cpid,uid&custid=s8863137&db=eric&AN=EJ1004927&site=eds-live&scope=site; http://dx.doi.org/10.1177/0098628312450435

Mathie, V. A., Beins, B., Benjamin Jr., L. T., Ewing, M. M., Hall, C. C. I., Henderson, B., et al. (1993). Promoting active learning in psychology courses. (pp. 183-214). Washington, DC, US: American Psychological Association. doi:10.1037/10126-007

Michaelson, L. K., & Sweet, M. (2008). The essential elements of team-based learning. *New Directions For Teaching & Learning, 2008*(116), 7-27.

Nguyen, K., & McDaniel, M. A. (2015). Using quizzes to assist student learning in the classroom: The good, the bad, and the ugly. *Teaching of Psychology, 42*(1), 87-92.

Novak, G. M. (2011). Just-in-time teaching. *New Directions For Teaching & Learning, 2011*(128), 63-73. doi:10.1002/tl.469

Saville, B. K. (2013). Interteaching: Ten tips for effective implementation. *Observer, 26*(2). Retrieved from http://www.psychologicalscience.org/index.php/publications/observer/2013/february-13/interteaching-ten-tips-for-effective-implementation.html

Saville, B. K., Lambert, T., & Robertson, S. (2011). Interteaching: Bringing behavioral education into the 21st century. *Psychological Record, 61*(1), 153-165. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,url,cpid,uid&custid=s8863137&db=bth&AN=57636710&site=eds-live&scope=site

Saville, B. K., Pope, D., Lovaas, P., & Williams, J. (2012). Interteaching and the testing effect: A systematic replication. *Teaching of Psychology, 39*(4), 317-324.

Saville, B. K., & Zinn, T. E. (2009). Interteaching: The effects of quality points on exam scores. *Journal of Applied Behavior Analysis, 42*(2), 369-374. doi:10.1901/jaba.2009.42-369.

Saville, B. K., Zinn, T. E., Neef, N. A., VanNorman, R., & Ferreri, S. J. (2006). A Comparison of Interteaching and Lecture in the College Classroom. *Journal of Applied Behavior Analysis, 39*(1), 49-61.

Smith, B., & Cardaciotto, L. (2011). Is active learning like broccoli? Student perceptions of active learning in large lecture classes. Journal of Scholarship of Teaching and Learning, 11(1), 53-61. Retrieved from http://eds.a.ebscohost.com/eds/pdfviewer/pdfviewer?sid=08b322a6-aca6-4b89-8fec-dd6ccbe6176a%40sessionmgr4002&vid=4&hid=4103

Webb, N. M. (1985). Verbal interaction and learning in peer-directed groups. *Theory into Practice, 24*(1), 32. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,url,cpid,uid&custid=s8863137&db=bth&AN=5199892&site=eds-live&scope=site&authtype=ip,uid

Tables

Table 1

Experimental Design

	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5
Long PGs	Interteaching	Interteaching (full prep guide at home) →			
Short PGs	Interteaching	Shorter prep guide at home + in-class application questions			

Note. Both classes began the semester with interteaching, which includes a full length prep guide completed at home. After Unit 1, the Short PG condition began the intervention, which included completing a revised prep at home and application questions during pair discussions in class.

Table 2

Average Exam Scores of the Long PG and Short PG conditions

	Long PG	Short PG	
	M (SD)	*M (SD)*	95% CI
Exam 1	.78 (.08)	.76 (.11)	[-.04, .08]
Exam 2	.81 (.13)	.77 (.17)	[-.05, .14]
Exam 3	.81 (.09)	.83 (.13)	[-.08, .06]
Exam 4	.81 (.08)	.70 (.27)	[-.01, .25]
Exam 5	.68 (.14)	.63 (.16)	[-.05, .14]
Final	.76 (.10)	.79 (.12)	[-.10, .04]
Total Points	.79 (.08)	.74 (.15)	[-.02, .13]

Note. CI = confidence interval

Table 3

Average Duration of Discussions in Minutes of the Long PG and Short PG conditions

	Long PG	Short PG	
	M (SD)	*M (SD)*	95% CI
Unit 1	29.72 (11.05)	24.30 (8.19)	[2.05, 8.79]
Unit 2	25.83 (8.29)	23.81 (5.53)	[-.09, 4.13]
Unit 3	28.19 (5.55)	28.91 (7.89)	[-3.05, 1.59]
Unit 4	24.35 (5.55)	22.69 (6.28)	[-.32, 3.65]
Unit 5	22.55 (5.31)	20.74 (5.31)	[-.45, 4.06]

Note. CI = confidence interval

Figure 1.

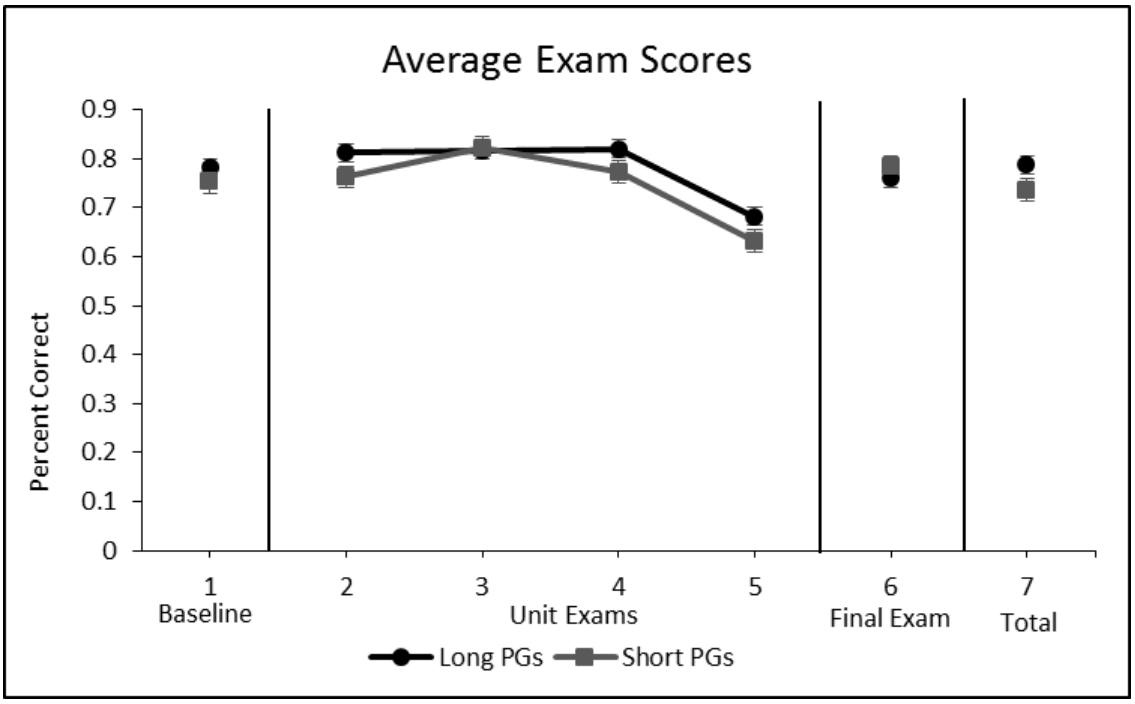


Figure 1. This graph shows the average exam scores of the Long PG and Short PG conditions. The Long PG condition scored higher on exams 2, 4, and 5; however, this difference was not significant. Standard error bars are included.

Figure 2.

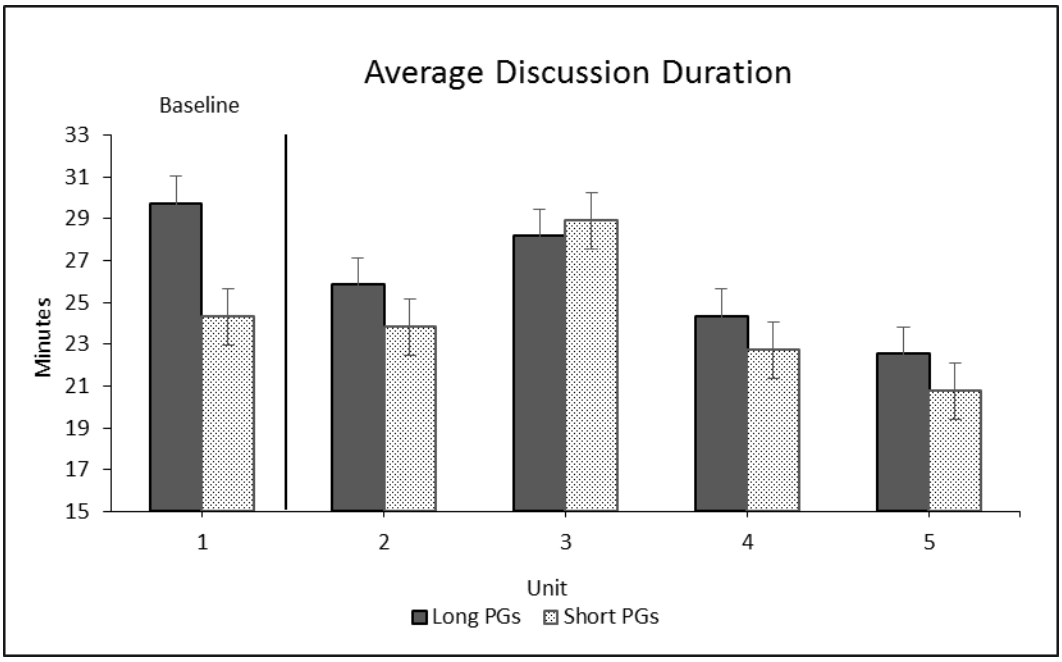
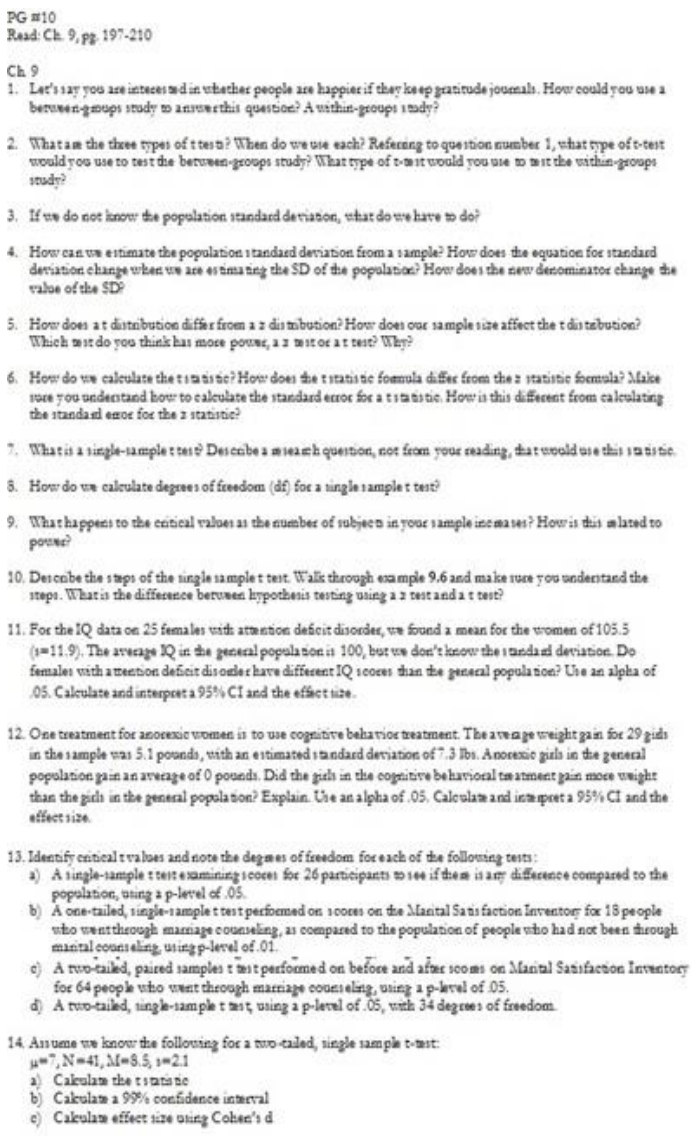


Figure 2. This graph shows the interaction between unit and section on duration of discussions (in minutes). There is a significant difference in discussion duration between the Long PG and Short PG condition on Unit 1. In general, the Long PG condition had longer discussions. Standard error bars are included.

Appendix A



PG #10
Read: Ch. 9, pg. 197-210

Ch. 9

1. Let's say you are interested in whether people are happier if they keep gratitude journals. How could you use a between-groups study to answer this question? A within-groups study?

2. What are the three types of t tests? When do we use each? Referring to question number 1, what type of t-test would you use to test the between-groups study? What type of t-test would you use to test the within-groups study?

3. If we do not know the population standard deviation, what do we have to do?

4. How can we estimate the population standard deviation from a sample? How does the equation for standard deviation change when we are estimating the SD of the population? How does the new denominator change the value of the SD?

5. How does a t distribution differ from a z distribution? How does our sample size affect the t distribution? Which test do you think has more power, a z test or a t test? Why?

6. How do we calculate the t statistic? How does the t statistic formula differ from the z statistic formula? Make sure you understand how to calculate the standard error for a t statistic. How is this different from calculating the standard error for the z statistic?

7. What is a single-sample t test? Describe a research question, not from your reading, that would use this statistic.

8. How do we calculate degrees of freedom (df) for a single sample t test?

9. What happens to the critical values as the number of subjects in your sample increases? How is this related to power?

10. Describe the steps of the single sample t test. Walk through example 9.6 and make sure you understand the steps. What is the difference between hypothesis testing using a z test and a t test?

11. For the IQ data on 25 females with attention deficit disorder, we found a mean for the women of 105.5 (s=11.9). The average IQ in the general population is 100, but we don't know the standard deviation. Do females with attention deficit disorder have different IQ scores than the general population? Use an alpha of .05. Calculate and interpret a 95% CI and the effect size.

12. One treatment for anorexic women is to use cognitive behavior treatment. The average weight gain for 29 girls in the sample was 5.1 pounds, with an estimated standard deviation of 7.3 lbs. Anorexic girls in the general population gain an average of 0 pounds. Did the girls in the cognitive behavioral treatment gain more weight than the girls in the general population? Explain. Use an alpha of .05. Calculate and interpret a 95% CI and the effect size.

13. Identify critical t values and note the degrees of freedom for each of the following tests:
    a) A single-sample t test examining scores for 26 participants to see if there is any difference compared to the population, using a p-level of .05.
    b) A one-tailed, single-sample t test performed on scores on the Marital Satisfaction Inventory for 18 people who went through marriage counseling, as compared to the population of people who had not been through marital counseling, using p-level of .01.
    c) A two-tailed, paired samples t test performed on before and after scores on Marital Satisfaction Inventory for 64 people who went through marriage counseling, using a p-level of .05.
    d) A two-tailed, single-sample t test, using a p-level of .05, with 34 degrees of freedom.

14. Assume we know the following for a two-tailed, single sample t-test:
    μ=7, N=41, M=8.5, s=2.1
    a) Calculate the t statistic
    b) Calculate a 99% confidence interval
    c) Calculate effect size using Cohen's d

Figure A1. An example of a full-length prep guide, used for the control group.

PG #10
Read: Ch. 9, pg. 197-210

Ch. 9

1. Let's say you are interested in whether people are happier if they keep gratitude journals. How could you use a between-groups study to answer this question? A within-groups study?

2. What are the three types of t tests? When do we use each? Referring to question number 1, what type of t-test would you use to test the between-groups study? What type of t-test would you use to test the within-groups study?

3. If we do not know the population standard deviation, what do we have to do?

4. How can we estimate the population standard deviation from a sample? How does the equation for standard deviation change when we are estimating the SD of the population? How does the new denominator change the value of the SD?

5. How does a t distribution differ from a z distribution? How does our sample size affect the t distribution? Which test do you think has more power, a z test or a t test? Why?

6. How do we calculate the t statistic? How does the t statistic formula differ from the z statistic formula? Make sure you understand how to calculate the standard error for a t statistic. How is this different from calculating the standard error for the z statistic?

7. What is a single-sample t test? Describe a research question, not from your reading, that would use this statistic.

8. How do we calculate degrees of freedom (df) for a single sample t test?

9. What happens to the critical values as the number of subjects in your sample increases? How is this related to power?

10. Describe the steps of the single sample t test. Walk through example 9.6 and make sure you understand the steps. What is the difference between hypothesis testing using a z test and a t test?

Figure A2. An example of a revised prep guide the intervention group completed at home.

PG #10 In-Class Questions
Ch. 9, Part 1

1. For the IQ data on 25 females with attention deficit disorder, we found a mean for the women of 105.5 (s=11.9). The average IQ in the general population is 100, but we don't know the standard deviation. Do females with attention deficit disorder have different IQ scores than the general population? Use an alpha of .05. Calculate and interpret a 95% CI and the effect size.

2. One treatment for anorexic women is to use cognitive behavior treatment. The average weight gain for 29 girls in the sample was 5.1 pounds, with an estimated standard deviation of 7.3 lbs. Anorexic girls in the general population gain an average of 0 pounds. Did the girls in the cognitive behavioral treatment gain more weight than the girls in the general population? Explain. Use an alpha of .05. Calculate and interpret a 95% CI and the effect size.

3. Identify critical t values for each of the following tests:
   a) A single-sample t test examining scores for 26 participants to see if there is any difference compared to the population, using a p-level of .05.
   b) A one-tailed, single-sample t test performed on scores on the Marital Satisfaction Inventory for 18 people who went through marriage counseling, as compared to the population of people who had not been through marital counseling, using p-level of .01.
   c) A one-tailed paired-samples t test performed on before and after scores on Marital Satisfaction Inventory for 18 people who went through marriage counseling, using a p level of .05.
   d) A two-tailed, paired samples t test performed on before and after scores on Marital Satisfaction Inventory for 64 people who went through marriage counseling, using a p-level of .05.
   e) A two-tailed, single-sample t test, using a p-level of .05, with 34 degrees of freedom.

4. Assume we know the following for a two-tailed, single sample t-test
   $\mu = 7$, $N = 41$, $M = 8.5$, $s = 2.1$
   a) Calculate the t statistic
   b) Calculate a 99% confidence interval
   c) Calculate effect size using Cohen's d

Figure A3. An example of the application questions the intervention group completed in class.

Appendix B

RECORD OF PAIR DISCUSSION

Name: _____ Partner: _____

Guide # _____ Date: _____

Time completed: _____ Was sufficient time provided? ____

Indicate your agreement with the following statements:

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
1. I was prepared for the discussion. (Circle a number)	1	2	3	4	5
2. My partner was prepared for the discussion.	1	2	3	4	5
3. We discussed all questions thoroughly.	1	2	3	4	5
4. We were able to provide original examples.	1	2	3	4	5
5. We were able to explain difficult concepts without looking at the book.	1	2	3	4	5
6. We had good two-way communication.	1	2	3	4	5

(Continued on back)

Please be specific about the questions to which you are referring (Can just put Q numbers).

7. What topics were easiest to understand?

8. What topics gave you difficulty?

9. Other comments or suggestions?

10. What is your gem for the day? (Question or most interesting thing you learned) **You need to have something here for credit.**

Figure B.1. A full record sheet used to obtain information on the quality of pair discussions by recording the self-ratings, duration of discussions, and attendance of each student.