View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by James Madison University

James Madison University
JMU Scholarly Commons

Masters Theses | The Graduate School

Spring 2016

Interteaching: Types of prep guide questions and their effect on student quiz performance

Verena S. Bethke
James Madison University

Follow this and additional works at: https://commons.lib.jmu.edu/master201019


Part of the Experimental Analysis of Behavior Commons

Recommended Citation
Bethke, Verena S., "Interteaching: Types of prep guide questions and their effect on student quiz performance" (2016). *Masters Theses*. 84.
https://commons.lib.jmu.edu/master201019/84

This Thesis is brought to you for free and open access by the The Graduate School at JMU Scholarly Commons. It has been accepted for inclusion in Masters Theses by an authorized administrator of JMU Scholarly Commons. For more information, please contact dc_admin@jmu.edu.

Interteaching: Types of Prep Guide Questions and Their Effect on Student Quiz Performance

Verena Sema Bethke

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Arts

Psychology

May 2016

FACULTY COMMITTEE:

Committee Chair: Bryan K. Saville

Committee Members/ Readers:

Tracy Zinn

Daniel D. Holt

Acknowledgments

I would like to thank my committee chair, Dr. Bryan K. Saville, who guided me through this process and provided much expertise and support. I would also like to thank my committee members, Dr. Daniel D. Holt and Dr. Tracy Zinn, for their invaluable input throughout. In addition, thank you to Dr. David B. Daniel for his help during the planning phases of this project; to Amanda Woolsey for having an extra set of eyes on things and helping with whatever was needed; to our undergraduate pilot testers who shaped the final procedures used; and to Nicholas Dashnaw for his help in creating the materials for this study.

Table of Contents

List of Tables

List of Figures

Abstract

Although previous research indicates that interteaching is an effective alternative to more traditional teaching approaches, not many component analyses of the method exist. For example, although researchers have shown that the prep guide component contributes to the effectiveness of interteaching, no research has directly examined how the content of the prep guides affects learning. The current lab-based study investigated whether having prep guides consisting of lower-level or higher-level questions impacted students' subsequent quiz performance. We found no significant differences in quiz performance between the two conditions, but several extraneous factors may have impacted the results.

Introduction

Many researchers, particularly in recent years, have sought alternatives to the traditional lecture method of instruction (e.g., Felder & Brent, 2003; Fernandes, Mesquita, Flores, & Lima, 2014; Hmelo-Silver, 2004; Woods, 2014). Despite these efforts, lecture still appears to be the most common form of pedagogy (Benjamin, 2002). This is true even in fields like medicine (Krueger et al., 2004), despite the fact that institutions have changed their entire programs to be based on methods other than lecture (Maudsley, 1999).

Criticisms of lecture are by no means recent. For example, Osgood (as cited in Jones, 2007) suggested that lecture is no better than reading a book and therefore unnecessary when most people can read. Jones (2007) also focused on lecture's shortcomings, suggesting they fall into four general categories: boredom and student inattention, redundancy when different media are available, the tendency to produce surface rather than deep learning, and inability to compensate for the different ways in which students actually learn.

The research on traditional lecture as an effective instructional method is also not impressive. In a review of the literature in medical education (Krueger et al., 2004), lecture appeared to be as efficient as other teaching methods when measuring pure knowledge transfer. When examining retention, transfer to new situations, or problem-solving, however, lecture tended to be less effective than other methods. Krueger et al. (2004) also found that, when it came to the latter skills, students watching online lectures did just as well as students observing in-person lectures.

In a more general review, Dunkin (1983) pointed out that research on lecture has typically focused on two main issues: effectiveness of lecture compared to other teaching methods and differences between more effective and less effective lecturing. A common problem with comparison studies, as Dunkin pointed out, was that very few of the studies that examined lecture defined what it is. According to Dunkin, researchers do, however, agree on at least three things when comparing lectures to discussions: first, that lecture is no better or worse than discussion when it comes to learning facts; second, that discussions are more effective for higher-level learning; and, third, that discussions are more effective in promoting changes in attitudes.

B. F. Skinner (1954, 1974) also criticized traditional teaching methods. He described traditional approaches as mainly punitive, in that students study and read to avoid aversive consequences. This kind of aversive control, however, is difficult to justify. According to Skinner, instructors should instead reinforce students' appropriate behaviors. In addition to his criticisms, Skinner (1974) also suggested some solutions. He suggested replacing contrived aversives with contrived reinforcers instead of the natural ones educators often attempt to use. He also advocated for making the consequences of reading more conspicuous and immediate by allowing for self-pacing, by making students respond to the information they read, and by making sure that what students just learned helps them with the next set of material. Both programmed instruction (Skinner, 1954) and Keller's (1968) Personalized System of Instruction—which fall under the umbrella of *behavioral teaching methods*—do these things well (Skinner, 1974).

Behavioral Teaching Methods

Programmed Instruction

Skinner (1954) developed programmed instruction to address the shortcomings he saw in traditional teaching methods. He suggested that programmed machines could present lessons in small steps and then provide learners with immediate reinforcement after each successful step. These teaching machines, then, would serve to eliminate traditional contrived, punitive methods and replace them with reinforcement. These machines would also enable students to move at their own pace, while the teacher could move around and assist students when necessary. Because of how the machines presented the lessons, it would also ensure that students mastered the material before moving on.

Programmed instruction, however, has not caught on. Moreover, research on programmed instruction has proved inconclusive in secondary school education (Kulik, Schwalb, & Kulik, 1982). It does not appear to be more effective than other teaching methods; however, Kulik et al. (1982) found a positive correlation between the year in which a study was conducted and the effect that programmed instruction had on achievement, such that programmed instruction appears to be more effective in recent implementations than in earlier ones. It also appears to be more effective in social science than in natural science and math classes. In higher education (Kulik, Cohen, & Eberling, 1980), about half of the studies failed to find a significant difference between programmed instruction and other methods. The majority of studies that did find a significant difference, however, showed programmed instruction to be more effective, albeit with a small overall effect sizes. Students in programmed instruction classes also generally spent less time on learning activities, but only a handful of studies have investigated this. Finally, Kulik et al. (1980) found a positive correlation between the year in which a study occurred and programmed instruction's effect on achievement.

Skinner (1984) also commented on why programmed instruction might not have caught on. He listed the popularity of both humanistic and cognitive psychology as possible factors, in addition to the assumption that teaching can properly be discussed using everyday language.

Personalized System of Instruction (PSI)

In 1968, Keller introduced his Personalized System of Instruction (PSI). PSI has five essential features: self-pacing, unit mastery, lectures and demonstrations for motivational purposes, emphasis on the written word, and use of undergraduate proctors (Buskist, Cush, & DeGrandpre, 1991; Keller, 1968). The self-pacing allows students to take quizzes in sequence, at their own pace. There are no arbitrary deadlines for these quizzes, and instead the instructor makes them available as students are ready for them. Unit mastery means that a student must master each unit (e.g., by getting a 90% or higher on the quizzes) before he or she may study the next unit. The instructor also provides study questions to aid students in identifying the "big picture" of each unit. An important component of mastery is that students may retake a unit quiz infinitely until they earn a high enough score, with no cost or penalty. Lectures or demonstrations should be kept short and only occur perhaps once a week, and attendance should not be required. Nor should quizzes or tests include any material from the lectures. Keller (1968) originally envisioned lectures as a motivational tool. Students generally communicate with both the instructor and the proctors in writing and also receive course materials in written form. In order to satisfy the "personalized" aspect of the system, the instructor enlists undergraduate proctors to provide individual attention to students. Proctors grade quizzes and respond to student questions throughout the course.

In numerous studies, PSI has produced improvements over traditional lecture in terms of student performance (Buskist, Cush, & DeGrandpre, 1991; Fox, 2004). In addition, research suggests that each of PSI's components seem to provide significant contributions to its overall effectiveness (Buskist et al., 1991; Fox, 2004). PSI, however, is not widely used. This may be for a few reasons. First, behavioral teaching methods do not mesh well with traditional academic practices (the semester structure and a wide distribution of student grades). Second, instructors may be reluctant to implement an approach so different from the approaches to which they are accustomed. Third, students are often resistant to such alternative teaching methods, which may make instructors reluctant to implement them. Fourth, PSI requires instructors to give up much of the control they normally have in their classrooms, which may make them uncomfortable. And finally, there is a general misunderstanding of behavioral principles, on which these methods are based (Fox, 2004; Saville, Lambert, & Robertson, 2011).

Interteaching

Because a pure mastery-based approach like PSI can be problematic in traditional academic settings, some researchers have suggested newer, more structured approaches. One of these approaches is interteaching (Boyce & Hineline, 2002; see also Saville et al., 2011). Interteaching is an attempt to retain some of the features of earlier behavioral teaching methods, while removing the self-pacing elements and making the method easier to implement for instructors. In a typical interteaching class, the instructor first creates a preparation (prep) guide that covers a particular reading assignment and consists of questions covering a wide range of question formats. The instructor then distributes this prep guide to the students, who complete the prep guide before class. In class,

students first hear a clarifying lecture on material from the previous day's prep guide; this lecture lasts about one third of the class period (see below). After the lecture, students form pairs to discuss the prep guide with a partner, which takes up the other two thirds of the class period. During the discussions, the instructor and any teaching assistants, if available, move around the classroom to answer questions and guide discussions. After students complete their discussions, they complete a record sheet, which informs the instructor on how the discussions went and which material was difficult to understand. Based on these record sheets, the instructor creates a brief clarifying lecture focusing on the material deemed difficult by students. The instructor delivers this lecture at the beginning of the next class period. After the lecture, students discuss the next prep guide.

Additional components of interteaching include frequent testing (at least five per semester; Boyce & Hineline, 2002) and participation points for participating in each pair discussion (totaling approximately 10% of the course grade across the entire semester). In addition, Boyce and Hineline introduced the concept of quality points, which introduces a cooperative contingency into interteaching. For example, suppose that Students A and B discussed a prep guide with each other and then took an exam on the material, which happened to include, as an essay question, an item the students had discussed together in class. The instructor would examine their responses on the essay question, and the two students could earn a small number of "quality" points if they both received an A or B on that question (e.g., 4 or 5 points on a 5-point question). But neither would receive points if one or both of them received fewer than 4 points. Boyce and Hineline suggested that quality points should total approximately 10% of a student's overall course grade.

Research on Interteaching

Because of the well-established behavioral theories that lie at the heart of interteaching, it should, conceptually, be an effective teaching method. Until recently, however, researchers have not studied it systematically**.**

Comparison studies. In the first experimental analysis of interteaching, Saville, Zinn, and Elliott (2005) compared interteaching to lecture and reading (they also included a control condition). Students in the interteaching condition read a short article, answered prep-guide questions, and then heard a brief clarifying lecture. Students in the lecture condition simply heard a lecture over the same material. Students in the reading condition read the article. One week later, all students returned to take a short quiz on the material. Students in the interteaching condition performed better on the multiple-choice quiz than students in any of the other three groups. In addition, students in the lecture, reading, and control conditions did not significantly differ from one another.

In two subsequent, classroom-based studies, Saville, Zinn, Neef, Van Norman, and Ferreri (2006) compared interteaching to lecture. In the first study, participants in a graduate special education course took quizzes after alternating conditions of interteaching and lecture. Their scores were higher on quizzes following interteaching than on quizzes following lecture. In Study 2, Saville et al. alternated between interteaching and lecture in two sections of an undergraduate research methods course; they also counterbalanced the order of conditions across the two sections. Student test scores following interteaching units were higher than test scores following lecture units for both sections of the course. On the cumulative final exam that occurred at the end of the semester, students also answered more interteaching-based questions correct than

they did lecture-based questions. Finally, students in both studies reported a preference for interteaching.

Arntzen and Hoium (2010) subsequently compared a session of interteaching with a single session of lecture. They collected data on participants' self-ratings of how much knowledge on the given subject they had before and after each session. In this self-report survey, participants reported greater knowledge gains after interteaching sessions than after lecture sessions.

In a later study, Saville, Pope, Truelove, and Williams (2012) examined whether interteaching was more effective for low-, moderate-, and high-GPA students. In this study, students in a psychology of learning course alternated between traditional lecture and interteaching multiple times during the semester. Saville et al. split students into different GPA groups using a tertiary split. They then measured exam performance for all three groups. Though all students performed better on exams following interteaching, students in the low- and moderate-GPA groups showed greater improvement following interteaching as compared to exam scores following lecture.

Finally, interteaching also appears to improve long-term recognition over traditional lecture. In another lab-based study, Saville et al. (2014) assigned students to either an interteaching, a lecture, or a control condition. Students then took a multiple-choice quiz immediately after exposure to the material, 1 week after exposure, and 1 month after exposure. Overall, students in the interteaching condition performed better on every quiz than students in the lecture and control conditions. In addition, students in the interteaching condition had higher quiz scores after 1 month than students in the lecture condition did immediately after the initial teaching session.

Some researchers have also examined the effectiveness of interteaching outside of psychology. Goto and Schneider (2009) implemented interteaching in a nutrition science course and had students complete a class evaluation survey about the effects of interteaching on their perceived learning outcomes. Students commented that the prep guides were helpful, that the questions and discussions fostered critical thinking, that interteaching promoted understanding instead of memorization, that they felt more focused during lectures because they were already exposed to the material, and that they felt more motivated to do additional research on the topics that were being covered in class. Tsui (2010) implemented interteaching in sociology courses and, although she did not present empirical findings, reported the experience to have been quite successful in that it produced more focused class discussions and more focused lectures. Emurian and Zheng (2010) used a combination of interteaching and programmed instruction—another behaviorally based teaching method—to teach Java™ in a course. Students showed progressive improvements in test performance and software self-confidence. However, the authors noted that gains observed during interteaching did not always transfer to subsequent quizzes.

Zeller (2010), who already used the normal variety of interteaching in his classes, decided to try a session of interteaching without an instructor present. Although he did not provide any empirical data, he did report that it was quite successful and that students performed just as well as they did when he was present. Finally, Slagter and Scribner (2014) investigated an adapted version of interteaching in a political science classroom. Using this adapted version of the method, students were more likely to complete the reading for class when it was part of an interteach assignment instead of a more

traditional lecture assignment, more likely to retain information for which there was a prep guide, read more carefully when the reading was part of an interteach assignment, generally liked interteaching better—though they did not like the group discussions—and, on average, said that the quality of the classroom environment improved as a result of interteaching. It is important to note that, in their adaptation of interteaching, the instructors may have also changed some components that make, for instance, the student discussions more successful in the original interteaching approach.

Research on improving interteaching. Some researchers have investigated how interteaching might be improved. One such possible improvement was to capitalize on the testing effect by introducing post-discussion quizzes. The testing effect refers to the phenomenon that tests appear to enhance later retention of material more than additional study of the material, even when students never receive feedback on those tests (see Roediger & Karpicke, 2006). In a preliminary, lab-based study by Lambert and Saville (2012), students participated in a mock interteaching session: reading a short article, answering prep-guide questions, and hearing a subsequent clarifying lecture. However, some students completed short quizzes after discussing the prep-guides, while other students completed anagrams instead. One week later, students who had completed post-discussion quizzes performed significantly worse on a follow-up quiz than the students in the anagram control condition. Saville, Pope, Lovaas, and Williams (2012) conducted a systematic replication of Lambert and Saville's study in two sections of an undergraduate psychology of learning course. Instructors taught both of these classes using interteaching and alternated whether students experienced post-discussion quizzes for the unit preceding the test (i.e., Section 1 received quizzes for units 1, 3, and 5, while Section 2

received quizzes for units 2, 4, and 6). The researchers did not find a significant difference between exam scores of students who experienced interteaching with post-discussion quizzes and students who experienced interteaching without post-discussion quizzes. Together, these studies suggest that adding post-discussion quizzes to interteaching does not improve its efficacy. This may be because interteaching already includes elements that improve performance, and, thus, additional quizzing may be redundent.

Interteaching component analyses. The bulk of the research so far, then, has focused on the effectiveness of interteaching as compared to more traditional teaching methods and on ways to improve its efficacy. This is because interteaching, as originally proposed, is an entire "package" of components to be implemented together (Boyce & Hineline, 2002). But since the introduction of interteaching over a decade ago, researchers have also examined which of its components contribute to its efficacy.

The first component analysis of interteaching investigated the impact of quality points on exam scores in two sections of an introductory psychology course (Saville & Zinn, 2009). Both sections experienced interteaching with and without quality points, but the researchers counterbalanced the order of conditions across the two sections, with each section experiencing both conditions multiple times. In the quality points condition, students could earn extra points, if both they and their partners performed well certain exams question. Saville and Zinn found that the presence of quality points did not significantly impact student exam scores.

Rosales, Soldner, and Crimando (2014) later investigated how instructors could make quality points more effective in interteaching. Because feedback on quality points

is delayed in their original implementation, Rosales et al. aimed to make the feedback more immediate. In this case, the study involved an alternating treatments design in a single class. The two conditions were the presence and absence of quality points. In both conditions, students discussed the material, completed a post-discussion quiz, completed a record sheet, and then heard a clarifying lecture at the beginning of the next class session. The post-discussion quizzes consisted of both fill-in-the-blank and short-answer questions, which required students to provide or apply a definition, recall information, or apply knowledge to novel examples. Students received quality points on a quiz if they and their discussion partner both scored 80% or higher on the quiz. If either student scored below 80%, neither received quality points. Quality points were not part of the students' final grades, as originally suggested by Boyce and Hineline (2002), and instead were extra credit that students could earn. Finally, students received immediate feedback on their quiz performance in the form of an answer sheet. They could then discuss this answer key with their discussion partner. Rosales et al. found that students' quiz scores were significantly higher when quality points were present compared to when they were absent. This changed version of quality points, with more immediate feedback, might thus be an effective component of interteaching.

Researchers have also investigated the effectiveness of the clarifying lectures in interteaching. Saville, Cox, O'Brien, and Vanderveldt (2011) assigned each of three sections of an undergraduate research methods course to a different lecture condition. One section experienced a delayed lecture at the beginning of the next class period (which was 2 or 5 days after discussing the prep guides with a partner). The second section experienced immediate lectures 5 minutes after discussing the prep guides with a

partner. Students in the third section did not experience any lectures. Overall, students in both lecture conditions earned more points during the semester than students in the control condition, but the two lecture conditions did not differ significantly from one another.

Felderman (2014) investigated whether giving students more frequent exams produced a performance difference in an introductory psychology course. One section completed 6 unit exams, while a second section completed 12. Students also completed pre-tests at the beginning of the semester and post-tests at the end. Overall, Felderman found no significant differences between the two sections on exam scores or the differences between pre- and post-test scores.

Cannella-Malone, Axe, and Parker (2009) conducted a study to investigate the prep guide component in a special education course. They compared student performance on quizzes depending on whether students generated their own prep guide questions or answered teacher-generated questions. In terms of overall quiz scores, writing questions led to only slightly higher quiz scores than answering questions. The difference between the two conditions appears to be bigger towards the end of the course, with the writing questions condition still producing higher quiz scores. Additionally, although participants scored slightly higher on multiple-choice questions in the writing condition, they performed substantially better on fill-in-the-blank questions in two quizzes in the answering condition. Student performance on factual short answer questions was slightly better in the answer questions condition, while performance on problem solving short answer questions was consistently better in the write questions condition. Finally,

students tended to prefer answering questions because they felt better prepared when doing that than when writing their own questions.

In a recent lab-based study, Saville, Bethke, Asdourian, and Cairns (2015) examined whether having prep guides impacted students' quiz performance. While one group experienced interteaching as normal, the other group experienced interteaching without the prep guide component. Students in the prep guide condition answered significantly more quiz questions correctly than students in the condition without the prep guide.

Interteaching and Prep Guide Questions

Other than Cannella-Malone et al.'s (2009) and Saville et al.'s (2015) studies, no other studies have examined the prep-guide component of interteaching. Boyce and Hineline (2002) suggested that the prep guides should contain different types of questions to occasion good discussions. Similarly, Saville et al. (2011) suggested that future research should examine whether the "level" of prep-guide questions has any effect on learning.

Determining "Levels" of Questions

A common approach for writing different "levels" of questions has followed the suggestions provided in Bloom's taxonomy (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956; see also Anderson, Krathwohl, & Bloom, 2001; and Krathwohl, 2002).

Although there are other approaches to constructing questions, interteaching researchers thus far have discussed prep guides in terms of Bloom's taxonomy (Cannella-Malone, Axe, & Parker, 2009; Saville et al., 2011; Zeller, 2010; Zinn & Saville, 2007). Bloom's taxonomy consists of six "levels" of questions: knowledge, comprehension,

application, analysis, synthesis, and evaluation. “Remember” refers to questions that require long-term remembering, which often require students to recognize, recall, locate, or identify. “Comprehension” refers to questions that require clarification, paraphrasing, representing, translating, illustrating, giving examples, classifying, constructing models, and so forth. “Application” refers to questions that require students to carry out or use a procedure in a given situation, whether familiar or unfamiliar. “Analysis” refers to questions requiring students to break concepts into constituent parts, determining how parts relate, differentiating between relevant and irrelevant, and so forth. “Synthesize” refers to questions requiring students to reorganize elements into new patterns or structures, generate, hypothesize, design, plan, construct, or produce. “Evaluate” refers to questions requiring students to make judgments based on criteria, check something, detect inconsistencies or fallacies in something, or judge or critique something (see Table 1).

Past research indicates that the types of questions students experience has an effect on how well they remember the information. In a series of early studies, Hunkins (1968, 1969) aimed to determine whether question type was related to student achievement in elementary school students. Two groups of students worked through a social studies text with sets of questions every day for a month. In one group, the questions required mostly knowledge, whereas in the other groups, the questions required mostly analysis and evaluation. Both types of quizzes consisted of short-answer questions. Hunkins found that, on the multiple-choice posttest, students in the higher-level group scored significantly better than students who answered the knowledge questions. Hunkins then broke the results down by type of posttest question and found

that the two groups did not differ in their performance on knowledge, comprehension, analysis, and synthesis questions. But the higher-level group did score significantly higher on application and evaluation questions.

More recently, McDaniel, Thomas, Agarwal, McDermott, and Roediger (2013) investigated the effects of question type in a middle-school science class. Students received quiz questions focused on either definitional information (lower level) or application of the principle (higher lever) throughout a unit of material. The questions focused on application increased student exam performance at the end of the unit for both definitional questions and application questions.

In another study by Jensen and colleagues (2014), students in two introductory biology courses experienced the same inquiry-based teaching method, but different levels of quiz and exam questions throughout the semester. One section received only questions that would be categorized as low-level in Bloom's taxonomy, whereas the other section received only high-level questions. The final exam for both sections encompassed questions of both types. Jensen et al. (2014) found that students exposed to high-level questions performed better on the final exam.

To bring Bloom's taxonomy into the realm of behavioral research, Crone-Todd, Pear, and Read (2000) created operational definitions for each level of the taxonomy (see Table 2). Crone et al. grouped questions from the first two levels of the taxonomy—knowledge and comprehension—together, as both of these require answers that can be found in the assigned material and require no extrapolation. Answers to knowledge questions may be memorized or closely paraphrased from the material, whereas answers to comprehension questions must be in the student's own words, while still using

appropriate terminology. Questions from the other four groups all require answers that go beyond the textual material and must be inferred or extrapolated. Answers to application questions require recognition, identification, or application of a concept or principle to a new situation that is not found in the material. Answers to analysis questions require breaking down concepts into their individual parts or identifying or explaining the necessary components of a concept, principle, or process. These kinds of questions might also require comparing and contrasting of concepts or explaining how an example illustrates a concept. Answers to synthesis questions require a student to put together parts to form a whole; for example, a question might require creating a new definition not identified in the material. Finally, answers to evaluation questions require presenting and evaluating reasons for and against a particular position. When Crone-Todd et al. tested these operational definitions, they found relatively high interobserver reliability for each level of question. In fact, Crone-Todd and Pear (2001) later used these definitions to aid in specifying learning objectives for a PSI course.

Previous research on Bloom's taxonomy, then, shows that the types of questions to which students are exposed affects learning. Students who are exposed to higher-level questions seem to, in later assessments, perform better on all types of questions. Interteaching researchers have noted the possible effects of question type and have referred to Bloom's taxonomy, but have not systematically studied whether the types of prep guide questions impact student learning. The purpose of the present study is to examine whether different question types (lower or higher level) affect learning in interteaching, as measured by multiple-choice quizzes consisting of mixed question types.

Method

Participants

The participants were students in various undergraduate psychology courses at James Madison University (JMU). Some students signed up through the Psychology Department's online participation pool (n = 7), and others were recruited using an in-class announcement (n = 65). All participated in exchange for partial course credit. The final sample consisted of 72 students (11 men, 61 women) whose average age was 20.19 years old (SD = 0.85). There were 3 freshmen, 20 sophomores, 45 juniors, and 4 seniors.

Materials

This study used a laboratory-based analogue, as seen in previous studies (e.g., Lambert & Saville, 2012; Saville et al., 2005; Saville et al., 2014). Students were assigned to one of two conditions: either higher-level prep guide (n = 36) or lower-level prep guide (n = 36).

Prep guide and quiz creation. Two prep guides, each consisting of seven questions, and a quiz, consisting of 20 questions, were created using Crone-Todd et al.'s (2000) operationalized version of Bloom's Taxonomy. Using an adapted version of Crone-Todd et al.'s decision-making flowchart using just 'higher level' and 'lower level' as categories (see Figure 1), we first created a number of lower-level and higher-level questions. For the purposes of this study, Categories I and II were considered lower level, and categories III through VI were considered higher level. To ensure that the questions were, in fact, higher or lower level questions as intended, four undergraduate research assistants, who were blind to the purpose of the study, used the flowchart and provided their independent assessment of the level of each prep-guide and quiz question. On all but

one question, the research assistants agreed that the question was of the intended level. The final question, on which there was not initial agreement, was rewritten to be at the intended level.

The final seven questions on each prep guide (see Appendix A for lower-level prep guide and Appendix B for higher-level prep guide) aimed to cover some of the key information in a short chapter on small-*N* designs (Saville & Buskist, 2003), which was chosen as the reading material for this study because we initially assumed that most participants would be unfamiliar with its content. Each question on the lower-level prep guide corresponded with a question on the higher-level prep guide, in that the factual information from the same section of the reading needed to be used to answer both of them, albeit at different levels. The final quiz (see Appendix C) consisted of 10 lower-level and 10 higher-level questions, all based on topics covered on the prep guides (cf. Boyce & Hineline, 2002). Each lower-level question was paired with a higher-level question that required the same information from the reading to answer.

Procedure

During the first session, which lasted between 1.25 and 1.5 hrs, students read the book chapter on small-*N* designs and completed either the lower-level or higher-level prep guide. Participants had 50 min to complete the reading and prep guide[1] but took an average of 40.66 min (*SD* = 6.89) to do so. Next, participants paired up to discuss their answers to the prep guide items. Participants had up to 30 min to complete the discussions, but no pair took longer than 15 min (*M* = 6.60, *SD* = 2.76). Finally,

[1] Pilot testing revealed that 50 min was a sufficient amount of time for students to complete the reading and prep guide.

participants in both conditions completed a record sheet and then heard a clarifying lecture that lasted about 15 min.

All participants returned 1 week later for Session 2, which took no more than 30 min to complete. First, participants took the multiple-choice quiz. The participants also provided demographic information that included, among others: gender, age, GPA, and whether they had previously experienced interteaching (see Appendix D). Finally, participants were debriefed upon completing the session (see Appendix E).

Results

Preliminary Analyses

Analysis of the demographic data indicated that the two groups did not differ significantly on any of the given factors: gender, age, number of credits taken in the current semester, overall GPA, or employment status (all *p*s > .05; see Table 3). The groups did differ, though, on how many psychology courses they had taken, with the higher-level group taking fewer courses (M = 5.83, SD = 2.71) than the lower-level group (M = 7.39, SD = 3.89), $t(70) = -1.97$, $p = .05$, $d = 0.47$. Finally, the groups did not differ with regard to how many interteaching-based classes they had previously experienced ($p > .05$). It is noteworthy, though, that approximately 75% of participants in each group had experienced interteaching in the past.

As expected, participants in the higher-level prep guide group took longer to complete the reading and prep guide (M = 43.21, SD = 6.62) than participants in the lower-level prep guide group did (M = 38.11, SD = 6.23), $t(70) = 3.36$, $p = .001$, $d = 0.79$. Also, participants in the higher-level prep guide group took longer to complete the discussion (M = 8.52, SD = 2.58) than participants in the lower-level prep guide group (M

= 4.69, *SD* = 1.13), $t(70) = 8.15$, $p < .001$, $d = 1.92$. Neither prep guide completion time nor discussion completion time, however, were significantly correlated with any of the three percentage scores obtained from the participants (see Table 4). Accordingly, further analyses did not include these as potential covariates. The only two variables that were correlated with participants' scores were GPA and prep guide completion time, which were, accordingly, included as a covariate in the primary analysis.

Primary Analyses

A MANCOVA (controlling for number of psychology courses, GPA, and prep guide completion time) found no significant differences between conditions on any of the primary variables, Wilks' $\lambda = 0.97$, $F(2, 66) = 1.20$, $p = .31$, $\eta_p^2 = 0.04$. More specifically, the higher-level group (*M* = 78.61, *SD* = 13.07) did not significantly differ from the lower-level group (*M* = 81.67, *SD* = 11.34) on total percentage of correct answers, $F(1, 67) = 1.08$, $p = .30$, $\eta_p^2 = 0.02$; on percentage of higher-level correct answers (higher level *M* = 77.50, *SD* = 12.28; lower level *M* = 79.72, *SD* = 15.02), $F(1, 67) = 0.01$, $p = .94$, $\eta_p^2 < 0.01$; or on percentage of lower-level correct answers (higher level *M* = 79.72, *SD* = 17.97; lower level *M* = 83.61, *SD* = 13.13), $F(1, 67) = 2.24$, $p = .14$, $\eta_p^2 = 0.03$ (see Figure 2).

Discussion

Previous research has shown that interteaching tends to produce better student-learning outcomes than lecture (e.g., Saville et al., 2006) and that the prep guides seem to be an important component of interteaching (Saville et al., 2015). But no studies have examined whether the types of prep-guide questions impact learning. The purpose of this study was to compare quiz performance after students completed either a prep guide

consisting either of lower-level questions or higher-level questions. In sum, there were no significant differences between the two groups on total quiz score or on lower- or higher-level quiz questions.

Past research has indicated that the types of questions experienced by students can affect subsequent test performance. Specifically, students who initially answer higher-level questions tend to perform better on both lower- and higher-level test questions. For example, McDaniel and colleagues (2013) found that students who were quizzed using higher-level questions performed better on both lower- and higher-level exam questions than students who were quizzed using only lower-level questions. Similarly, Jensen and colleagues (2014) found that students who received higher-level quizzes throughout the semester performed better on both lower- and higher-level questions on a final exam. The present study, however, does not support these findings. There was no significant difference in quiz performance—total questions, higher-level questions, or lower-level questions—between students who answered higher-level prep guide questions and students who answered lower-level prep guide questions. This could be due to a variety of factors.

First, it is simply possible that the prep guide component may not be particularly important in interteaching. If so, the content of the prep guides would have no impact on how well students performed on the subsequent quiz. However, this seems unlikely due to past research indicating that students perform better in the presence of prep guides than in their absence. Saville et al. (2015), for example, compared interteaching with and without the prep guide component in a lab-based setting and found that students in the prep guide condition answered significantly more quiz questions correctly than students

who simply took notes instead of completing a prep guide. Cannella-Malone, Axe, and Parker (2009) also investigated the prep guide component and found some differences depending on whether students wrote their own prep guide questions or whether students answered teacher-generated prep guide questions. Together, these results suggest that the prep guides seem to contribute in some form to the overall efficacy of interteaching.

Another possibility is that the prep guides matter only because they cause students to contact course material in ways that might ultimately lead to higher-order types of thinking. For example, over the course of their discussions, students might find themselves discussing higher-level topics regardless of whether their prep guides contained lower- or higher-level questions. After reading a lower-level question, for instance, they may come up with their own examples, scenarios, or applications regardless of whether the prep-guide items specifically asked them to do so. This, however, also seems unlikely as an explanation for the present results. First, the discussion times were different between the two conditions, with the lower-level condition having significantly shorter discussions than the higher-level condition. If the discussion became "higher order" regardless of the prep guide items, one might expect similar discussion lengths between the two groups. Moreover, discussion length was relatively short overall (see Table 3), but particularly in the lower-level condition. Once again, if these students were discussing at a "higher" level, their discussion times would have likely been more similar to students in the higher-level prep guide condition.

The short discussion durations point to different but related possibility: that, in interteaching, it is the discussion and not the prep guide that affects student performance. The prep guides may only matter to the extent that they can impact and steer the

discussions. In other words, if a prep guide improves student performance, it does so because it generates a longer, deeper discussion than the alternative. Because students in both conditions engaged in relatively short discussions (approximately 7 min, on average), the discussions generated by the prep guides may have simply not been long enough to have a positive effect. This possibility, too, seems unlikely, though, given that both groups scored around 80% on the quiz

The fact that both groups scored around 80% on the quiz points to a possible ceiling effect. There is one likely reason for this observation. Only 7 of the 72 participants came from the standard participation pool, consisting of students who were enrolled in introductory psychology courses and likely had little familiarity with small-N research designs. The other students were psychology majors: 17 from a lower-level research methods course and 48 from an upper-level content course. It is likely that these students, particularly the latter 48, were not completely naïve to the subject matter covered in the reading. In fact, most of the students (about 75% in each group) also reported having experience with an interteaching-based class prior to participating in this study. Within the JMU Psychology Department, the faculty members who typically use interteaching in their courses are also the same faculty members who often teach their students about small-N research designs. This makes it a reasonable assumption that a large portion of the present sample may have learned about small-N designs before participating in the study. If students were already familiar with the material, the interteaching sessions would not have added much to their pre-existing knowledge, which may have ultimately been responsible for (or at least strongly contributed to) the relatively high quiz scores observed in both prep guide groups. Between pre-existing

knowledge about small-N designs and our use of a multiple-choice quiz (on which guessing may produce some correct answers simply by chance), students may have performed well despite the questions on their particular prep guide. In support of this notion, when the students who had prior interteaching experience were excluded from the sample, leaving 9 participants in each condition, the group means, although not significantly different (all *p*s > .25), trended in the hypothesized direction, with the higher-level condition scoring higher on each of the dependent variables (results not shown). Given the small remaining sample sizes, though, these results must be taken cautiously.

There are other limitations that may have contributed to the present results as well. This study was conducted in a lab setting, while previous studies on question types occurred in true classroom settings (McDaniel et al., 2013; Jensen et al., 2014). Although previous lab-based interteaching studies have provided a good analogue of interteaching (e.g., Lambert & Saville, 2012; Saville et al., 2005; Saville et al., 2014), lab-based studies have inherent limitations compared to classroom-based studies. For instance, students in real classrooms are, presumably, motivated to perform well because they are receiving grades for their performance. And although students in this study received course credit (participant pool completion or extra credit) contingent upon their participation in the study, their quiz performance had no impact on whether they would receive the points. Although students in both conditions performed well on the quiz, this lack of motivation might have also contributed to the relatively short discussion times, which may have indirectly limited the impact of our independent variable.

Finally, in part due to the relatively small effect sizes, this study had low statistical power. The MANCOVA, controlling for number of psychology courses, GPA, and prep guide discussion time, provided only 25.3% power. The power for each dependent variable separately did not exceed 31.5% for any of them. As such, low power may also be an explanation for the lack of significant results.

Future research should, first and foremost, attempt to address the limitations present in this study. A naïve sample—including demographic questions that attempt to identify whether participants are naïve to the material—would be useful. It may also be beneficial to investigate question types while holding discussion and prep guide completion time constant instead of allowing them to vary between participants and controlling for them statistically. Finally, it is likely that experiencing lower- or higher-level questions for the duration of a unit, and then being tested on that unit, is different than experiencing one reading and prep guide, and so implementing the manipulation in a setting closer to a real classroom may be useful. Regardless, future researchers should continue to examine the prep guide component of interteaching and determine to what extent its contents contribute to student-learning outcomes.

Table 1

Bloom's Original Taxonomy

Categories	Subcategories (if any)	
Knowledge	Knowledge of specifics	Knowledge of terminology Knowledge of specific facts
	Knowledge of ways and means of dealing with specifics	Knowledge of conventions Knowledge of trends and sequences Knowledge of classifications and categories Knowledge of criteria Knowledge of methodology
	Knowledge of universals and abstractions in a field	Knowledge of principles and generalizations Knowledge of theories and structures
Comprehension	Translation Interpretation Extrapolation	
Application		
Analysis	Analysis of elements Analysis of relationships Analysis of organizational principles	

Synthesis	Production of a unique communication
	Production of a plan, or proposed set of operations
	Derivation of a set of abstract relations
Evaluation	Evaluation in terms of internal evidence
	Judgments in terms of external criteria

Note. Adapted from Krathwohl (2002).

Table 2

A Behavioral Adaptation of Bloom's Taxonomy

Overall level	Description	Categories	Description
Lower level (Categories I and II)	Answers will always be found in the assigned material; require no extrapolation	Knowledge	Answers may be memorized or closely paraphrased from material
		Comprehension	Answers must be in student's own words but still using appropriate terminology
Higher level (Categories III, IV, V, and VI)	Answers go beyond textual material: they must be inferred or extrapolated from the information in the text Require "processing" of information	Application	Answers may require recognition, identification, or application of concept/principle learned at "comprehension" in a new situation or to solve a new problem Question presents/requires examples not found in assigned material
		Analysis	Answer requires breaking down concepts into constituent parts, or identification/explanation of essential components of concepts/principles/processes

that is not already performed in assigned material		May require students to compare/contrast, or explain how an example illustrates a concept/principle/etc.
	Synthesis	Answer requires putting together parts to form a whole (opposite of "analysis") May require generating definitions not identified in assigned material or explaining how to combine principles/concepts to produce something new
	Evaluation	Answer requires presenting and evaluating reasons for/against a position and to come to a conclusion regarding the validity of that position Most important part: justification/rationale for conclusion Involves use of all preceding levels

Note. Adapted from Crone-Todd, Pear, and Read (2000).

Table 3

Demographics Information by Condition

	Total		Lower level		Higher level		*p*-value
	M	*SD*	*M*	*SD*	*M*	*SD*	
Prep guide completion time (in min)	40.66	6.89	38.11	6.24	43.21	6.62	.001
Discussion completion time (in min)	6.60	2.76	4.69	1.13	8.52	2.58	<.001
Gender	15.3% male 84.7% female		13.9% male 86.1% female		16.7% male 83.3% female		.74
Age	20.19	0.85	20.22	0.87	20.17	0.85	.78
GPA	3.24	0.42	3.28	0.35	3.19	0.47	.38
Number of psychology courses	6.61	3.42	7.39	3.89	5.83	2.71	.05
Credits taken this semester	15.13	2.77	15.22	2.28	15.03	3.21	.77
Experienced interteaching	74.6% yes 25.4% no		74.3% yes 25.7% no		75% yes 25% no		.95
Have a job	50% yes		44.4% yes		55.6% yes		.35

	50% no		55.6% no		44.4% no		
If job, how many hours	12.44	8.10	10.50	8.28	14.00	7.81	.20

Table 4

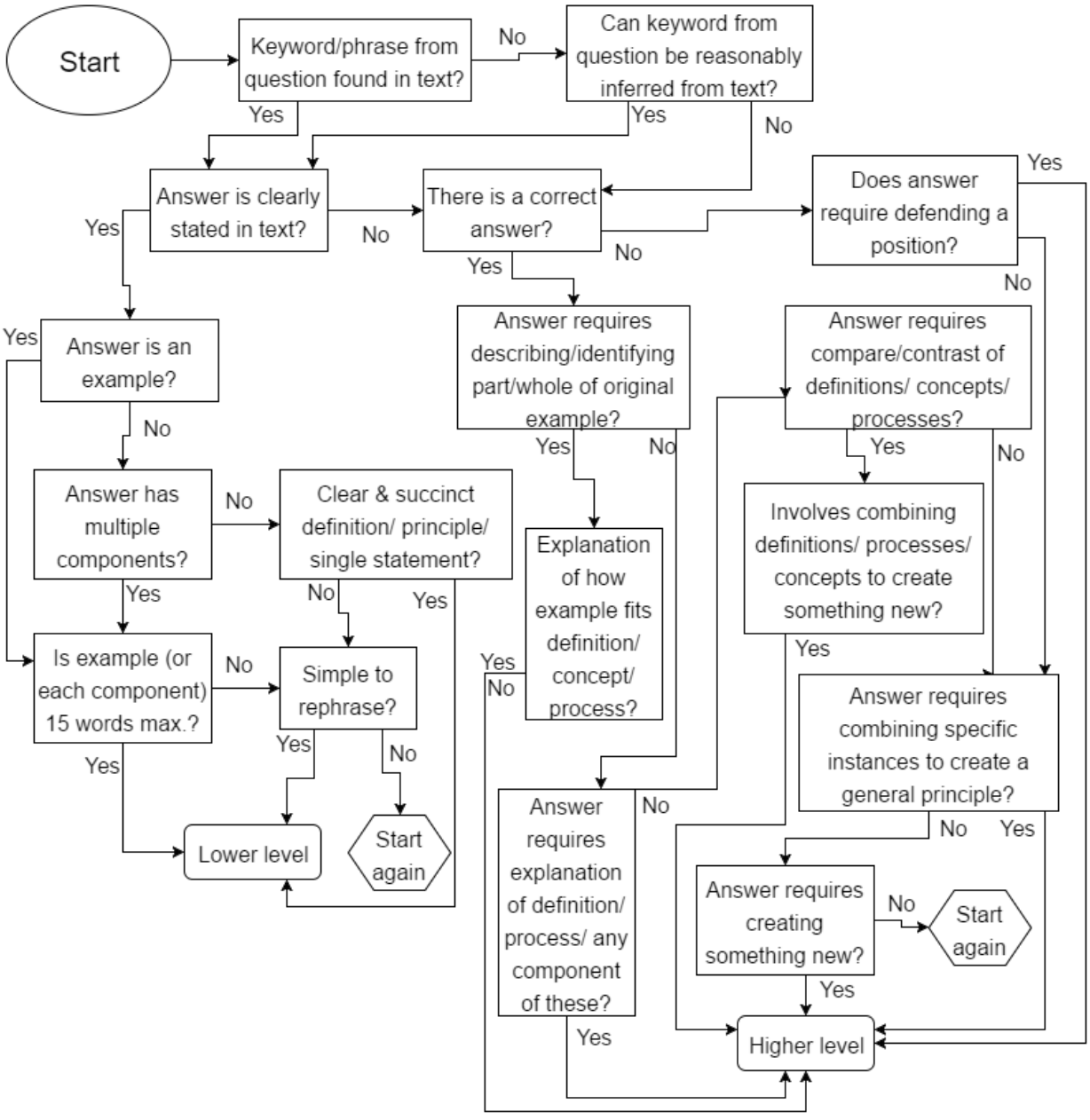
Correlations of Potential Covariates with Dependent Variables

Variable	Total quiz score	HL quiz score	LL quiz score	Prep guide completion time	Discussion time
HL quiz score	.805**				
LL quiz score	.857**	.384**			
Prep guide completion time	.117	-.017	.196†		
Discussion time	-.027	-.013	-.031	.345**	
Age	-.030	-.073	.018	.123	-.057
GPA	.355**	.378**	.224†	-.073	-.102
Number of psyc courses	.185	.193	.119	-.057	-.160
Credits taken this semester	.049	.195	-.092	-.007	.132

† $p < .10$, ** $p < .01$, * $p < .05$

Figure 1

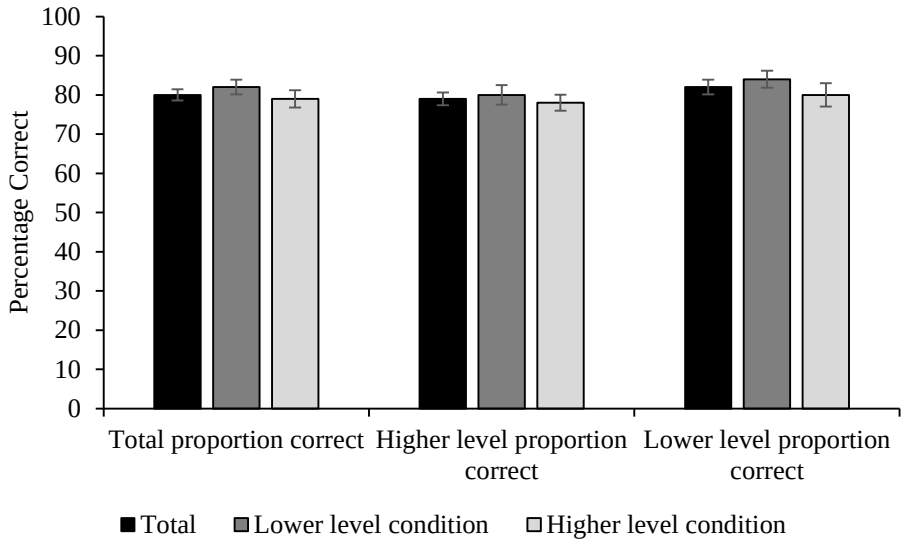
Adaptation of the Flowchart of the Operational Definitions of Bloom's Taxonomy



Note. Adapted from Crone-Todd, Pear, and Read (2000).

Figure 2

Mean Percentage Correct on Each Quiz Score by Condition



Note. Error bars represent standard error of the mean.

Appendix A

1. Why do researchers use Small-N research designs?
2. In what five primary ways do small-N designs differ from large-N designs? Briefly describe the differences for each of the five ways.
3. In small-N research designs, why is it important to continually monitor the dependent variable (DV)?
4. What is the purpose of establishing a baseline in an ABA design?
5. What are two reasons why it's useful to add a second treatment condition (B) in an ABAB design?
6. In the example on page 78, the authors describe a small-N design involving two independent variables (IVs). Why did the researchers introduce a design with more than one IV?
7. In a multiple-baseline design, the treatment is introduced at different times for each subject. Why?

Appendix B

1. Suppose that you want to investigate whether a dog's barking decreases after you start scolding it for barking. Why might you use a small-N research design to investigate this?
2. There are five ways in which small-N and large-N research designs differ. Imagine you wanted to study whether caffeine significantly affects how many words you can remember from a word list. Describe how your study would look if you used a large-N design. Now describe how your study would look if you used a small-N design.
3. *A study assessed whether the presence of a vibrating pager caused three autistic teenagers to eat more slowly. Participants were taught to take a bite only when the pager was vibrating. An ABAB design indicated that the vibrating pager slowed the pace of eating for all 3 participants (Anglesea, Hoch, & Taylor, 2008).* What was the dependent variable (DV) in this study, and how did continually monitoring it help the researchers determine whether the pager had a positive effect?
4. The article gives an example on page 76 about a study investigating human competition in a laboratory, using what the authors call "reinforcement schedules." In this specific example, how does the baseline help the researchers determine the ways in which their independent variable (IV) affects the participants?
5. *A study assessed whether the presence of a vibrating pager caused three autistic teenagers to eat more slowly. Participants were taught to take a bite only when the pager was vibrating. An ABAB design indicated that the vibrating pager slowed the pace of eating for all 3 participants (Anglesea, Hoch, & Taylor, 2008).* What were the baseline and treatment conditions in this example? Why would the researchers choose an ABAB design over an ABA design for this study?
6. Choose a behavior of yours that you would like to change. Create a small-N design using two treatments (or independent variables) to test how you could change that behavior.
7. Look at the graph below. Is this a multiple-baseline design or not? How do you know? Explain your answer.

Appendix C

1. Which of the following might be best studied using a small-N research design? (Higher)
    a. **Investigating whether a student's class participation increases after a teacher starts praising him.**
    b. Investigating whether smokers and non-smokers have different personality traits.
    c. Investigating whether cat owners and dog owners have different IQs.
    d. Both B and C are correct.

2. Why do researchers use small-N research designs? (Lower)
    a. To understand why people behave the way they do.
    b. To understand how groups of people differ, on average.
    c. To understand and change maladaptive behaviors.
    d. Both A and C are purposes of the small-N design approach.

3. Which of the following is **NOT** one of the five ways in which small-N and large-N designs differ? (Lower)
    a. The methods they use to analyze their data.
    b. Whether they allow researchers to conclude that a treatment caused a change in behavior.
    c. How many subjects they use in their study.
    d. How many levels of the independent variable (IV) the subjects experience.

4. Read the following example and determine which type of design is being used. *The researchers attempted to investigate whether turning studying into a game might impact students' quiz performance. Eight participants all experienced two conditions in an ABAB pattern: traditional studying and a studying game. The researchers analyzed their data using visual analysis (adapted from Neef, Perrin, Haberlin, & Rodrigues, 2011).* (Higher)
    a. It is a large-N design.
    b. It is a medium-N design.
    c. It is a small-N design.
    d. It is a correlational design.

5. Which of the following examples is most characteristic of a small-N research design? (Higher)
    a. A study in which researchers wish to use a statistical analysis to analyze the data they will obtain from the pretest and posttest that their 40 participants completed.

b. A study in which researchers wish to examine how the note-taking behavior of seven students changes as they experience three different classroom conditions.
c. A study in which researchers want to examine how 50 physics majors eat pizza differently than 50 psychology majors.
d. A study in which researchers want to examine a large group of randomly selected smokers and generalize the findings to the general population of smokers based on that single study.

6. Which of the following statements is **FALSE**? (Lower)
   a. Small-N designs analyze data using visual analysis whereas large-N designs analyze data using statistical analysis.
   b. Participants in small-N designs experience all levels of the independent variable (IV) while participants in large-N designs experience only one level of the IV.
   c. Large-N designs allow researchers to determine whether a treatment (or IV) had an effect on the DV. Small-N designs do not.
   d. Small-N designs rely on replications to generalize their findings while large-N designs rely on random selection and random assignment of subjects.

7. Why is the dependent variable (DV) continually monitored in small-N designs? (Lower)
   a. Because doing so helps control for confounding variables by keeping the DV stable.
   b. Because doing so makes small-N designs more similar to large-N designs.
   c. Because doing so helps the researcher identify other confounding variables that may influence the DV.
   d. Because doing so means that the researcher can be sure that the IV caused any change in the DV.

8. *A study assessed whether the presence of a vibrating pager caused three autistic teenagers to eat more slowly. Participants were taught to take a bite only when the pager was vibrating. An ABAB design indicated that the vibrating pager slowed the pace of eating for all 3 participants (Anglesea, Hoch, & Taylor, 2008).* Identify the DV in this study. (Higher)
   a. How quickly the teenagers ate.
   b. The presence of the vibrating pager.
   c. Autism.
   d. The length of the pre-determined intervals.

9. In the article you read, the authors described a study with an ABA design, in which researchers attempted to investigate human competition in a laboratory,

using what the authors call "reinforcement schedules." How did the baseline help the researchers obtain their results? (Higher)

a. By comparing responses during baseline with responses during the treatment condition, the researchers could assume that the treatment had caused the change in behavior.

b. Because the baseline ensured that the participants did not know which condition they experienced, the researchers could assume that the treatment caused the change in behavior.

c. Because the baseline ensured that the participants did not get bored, the researchers could assume that the treatment caused the change in behavior.

d. Because during the baseline, researchers were able to test a few different interventions, they could assume that the treatment caused the change in behavior.

10. What is the purpose of establishing a baseline in an ABA design? (Lower)

a. The baseline (which measures how participants respond "normally") lets researchers generalize the findings of the study to the general population.

b. The baseline (which measures how participants respond "normally") lets researchers use statistical methods to analyze the data from the study.

c. The baseline (which measures how participants respond "normally") lets researchers treat participants in a more ethical manner.

d. The baseline (which measures how participants respond "normally") lets researchers be fairly certain that the independent variable (IV) caused changes in the dependent variable (DV).

11. Which of the following is an advantage of adding a second treatment condition (B) in an ABAB design? (Lower)

a. It allows the researcher to test whether two different independent variables (IVs) work together to produce a change in the dependent variable (DV).

b. It allows the researcher to be even more confident that the IV caused the change in DV.

c. It allows the researcher to test the effects of two different treatments (or independent variables).

d. It requires less time to conduct an ABAB design than it does to conduct an ABA design.

12. In the example in question 8, the researchers chose an ABAB design to investigate the effect of a vibrating pager on eating speed. What factor could have led the researchers to choose an ABA design instead? (Higher)

a. Because the researchers only have access to two participants instead of three.

b. Because the researchers had access to ten participants instead of three.

c. Because instead of having access to the participants for a month, the researchers only have access to them for two weeks.
d. Because the change in eating speed caused by the vibrating pagers is permanent.

13. Researchers want to increase the time a student spends doing reading assignments for his classes. They have two treatments they want to try: (a) giving the student a chocolate reward for spending at least 20 minutes every day on his readings, and (b) fining the student a small amount of money for not spending at least 20 minutes every day on his readings. They also want to see what effect those two treatments have when implemented at the same time. Which of the following would be the correct arrangement of conditions for a two-IV (factorial) small-N design like this? (Higher)
    a. Baseline | Presence of a chocolate reward | Presence of monetary fine | Presence of both chocolate reward and monetary fine | Baseline
    b. Presence of a chocolate reward | Baseline | Presence of monetary fine | Baseline | Presence of both chocolate reward and monetary fine
    c. Baseline | Presence of a chocolate reward | Baseline | Presence of both chocolate reward and monetary fine | Baseline
    d. Baseline | Presence of a chocolate reward | Baseline | Presence of monetary fine | Baseline | Presence of both chocolate reward and monetary fine | Baseline

14. Why would researchers use a research design with more than one IV? (Lower)
    a. So they can learn about more than one IV without having to conduct multiple studies.
    b. To be more certain about the first IV's effects on the DV.
    c. To learn about what happens to the dependent variable (DV) when you mix treatments (or IVs) together.
    d. Both A and C are reasons to introduce a design with more than one IV.

15. What are the disadvantages of using a design with more than one IV? (Lower)
    a. It requires additional time and effort because it has more conditions.
    b. It requires additional participants because it has more conditions.
    c. It requires additional researchers because it has more conditions.
    d. It requires additional dependent variables because it has more conditions.

16. Suppose that researchers wanted to investigate the interaction between the two IVs, as in Question 13 (with the chocolate reward and monetary fine), but don't want to use the normal arrangement for such a design because of the disadvantages associated with this type of design: What could they do to investigate the interaction between the two treatments (IVs)? (Higher)

a. They could use an ABA design and a separate ACA design.
b. They could use an ABAB design and a separate ACAC design.
c. They could use a multiple baseline design.
d. None of these designs would be appropriate for investigating the interaction between two IVs.

17. Imagine the following scenario. *In this study, researchers attempted to investigate the effects of praise given by a principal on the attendance record of three elementary school students. During baseline, each student receive no praise, and the researchers simply collect how often the students attend school. The baseline period lasts 7 weeks for Student 1, 10 weeks for Student 2, and 12 weeks for Student 3. At the end of the baseline period for each student, the principle starts giving praise for each day that the students are in school. The researchers ultimately find that the attendance rates for each student increase after the principal starts praising that student (Copeland, Brown, & Hall, 1974).* Which of the following best describes the type of design the researchers used? (Higher)

a. This study used multiple-baseline design because the results showed that each student's behavior only changed after the principal started praising that student.
b. This study used a multiple-baseline design because the researchers started with a baseline and then varied when the treatment was introduced for each.
c. This study did not use a multiple-baseline design because the researchers never returned to baseline for any of the students.
d. This study did not use a multiple-baseline design because multiple-baseline designs only measure different behaviors, and this study included three participants with the same behavior.

18. Why in a multiple-baseline design are the treatments introduced for each participant at different times (after different baseline lengths)? (Lower)

a. Because it allows researchers to determine if the treatment (IV) is the true cause of changes in the dependent variables (DV).
b. Because it allows researchers to make sure that participants cannot anticipate when the treatment will be implemented.
c. Because it keeps the researchers unaware of which treatment each participant is experiencing.
d. Because it keeps the participants unaware of which treatments they are in.

19. Which of the following statements is/are true about multiple-baseline designs? (Lower)

a. Multiple-baseline designs can be used be used when examining multiple behaviors, multiple settings, or multiple participants.

b. The first behavior in a multiple-baseline design does not necessarily have to include a baseline measure, as long as the others do.
c. The last behavior in a multiple-baseline design does not necessarily have to include a baseline measure, as long as the others do.
d. Multiple-baseline designs require a return to baseline, just like ABA designs do.

20. Which of the following examples is **NOT** a correctly designed multiple-baseline study? (Higher)
    a. In the first session, the researchers began collecting baseline data for two participants, Nathan and Jessica. For Nathan, they continued collecting baseline data until Session 5 and then implemented the treatment. For Jessica, they continued collecting baseline data until Session 10 and then implemented the treatment.
    b. In the first session, the researchers began collecting baseline data for John. For John, they continued collecting baseline data until Session 5 and then implemented treatment the following session. In Session 6, they also started measuring baseline for their second participant, Sarah. They collected baseline data for Sarah until Session 13, after which they implemented treatment.
    c. In the first session, the researchers began collecting baseline data for how often Drew was out of his seat and for how often he interrupted others when they were talking. For "out-of-seat behavior," they collected baseline data until Session 10, and then implemented treatment. For "talking-disruption behavior," they collected baseline data until Session 15 and then implemented treatment.
    d. In the first session, the researchers started collecting baseline data on how much time Rachel spent reading and how much time she spent playing the clarinet. For time spent reading, they collected baseline data until Session 4 and then implemented treatment. For time spent practicing clarinet, they collected baseline data until Session 10 and then implemented treatment.

Appendix D

Demographics Questionnaire
Instructions: Please answer the following questions as accurately as possible.

1. Gender: Male Female Other

2. Age: _____

3. Current Year in School: Freshman Sophomore Junior Senior

4. Cumulative grade point average (if unknown, give best approximation): _____

5. Number of psychology classes you have had so far: _____

6. Number of credit hours you are taking this semester: _____

7. Have you experienced "interteaching" in any of your courses so far?

8. Do you currently have a job? Yes No

If **Yes** to #8, how many hours per week do you work, on average? _____

Appendix E

Debriefing Form

Title of Project: Improving Educational Outcomes in College Students Using Study Guides

Investigator: Verena Bethke (email: bethkevs@dukes.jmu.edu; phone: 518-986-7235)

One of the primary goals of education is to get students to remember material from their courses. Unfortunately, a good amount of research shows that students quickly forget much of the material they learn in their courses; in fact, some studies suggest that, after 3 months, students forget nearly 90% of the material they learn. With regard to learning and remembering, considerable research has shown that alternative teaching methods tend to produce better results than lecture. Interteaching is a new teaching method that is based on well-established psychological principles. In this study, we wanted to know if higher level prep guide questions produce better learning than lower level prep guide questions. To do this, we assigned students to one of two conditions: an **interteaching with higher level questions** condition (where students read a short article, completed a study guide consisting only of higher level questions, discussed the material with another student, and then heard a brief lecture) or an **interteaching with lower level questions** condition (where students read a short article, completed a study guide consisting only of lower level questions, discussed the material with another student, and then heard a brief lecture). We then had students from each condition take a quiz 1 week later, which consisted of both lower and higher level questions. This will allow us to determine whether the question type on the prep guide influences student performance on the different question types in the quiz.

Your participation is now complete. Thank you for your participation. We ask that you do not share any of the details of this experiment with anyone else because we are still collecting data. If you have any additional questions about the study, please feel free to contact the investigator listed above.

References

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of educational objectives*. New York, NY: Longman.

Arntzen, E., & Hoium, K. (2010). On the effectiveness of interteaching. *The Behavior Analyst Today*, *11*, 155–161.

Benjamin, L. T. Jr. (2002). Lecturing. In S. F. Davis & W. Buskist (Eds.), *The teaching of psychology: Essays in honor of Wilbert J. McKeachie and Charles L. Brewer* (pp. 57-67). Mahwah, NJ: Erlbaum.

Bloom, B. S., Englehart, M. B., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives, the classification of educational goals.* New York, NY: Longmans, Green.

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*, 215–226.

Buskist, W., Cush, D., & DeGrandpre, R. J. (1991). The life and times of PSI. *Journal of Behavioral Education*, *1*, 215–234.

Cannella-Malone, H. I., 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 and Learning*, *9*, 22 – 35.

Crone-Todd, D. E., Pear, J. J., & Read, C. N. (2000). Operational definitions for higher-order thinking objectives at the post-secondary level. *Academic Exchange*, 99–106.

Crone-Todd, D. E., & Pear, J. J. (2001). Application of Bloom's Taxonomy to PSI. *The Behavior Analyst Today*, *2*, 204–210.

Dunkin, M. J. (1983). A review of research on lecturing. *Higher Education Research & Development*, *2*, 63–78.

Emurian, H. H., & Zheng, P. (2010). Programmed instruction and interteaching applications to teaching Java™: A systematic replication. *Computers in Human Behavior*, *26*, 1166–1175. doi:10.1016/j.chb.2010.03.026

Felder, R. M., & Brent, R. (2003). Designing and teaching courses to satisfy the ABET engineering criteria. *Journal of Engineering Education*, *92*, 7–25.

Felderman, T. A. (2014). Preliminary analysis of interteaching's frequent examinations component in the community college classroom. *Journal of College Teaching & Learning*, *11*, 149-156.

Fernandes, S., Mesquita, D., Flores, M. A., & Lima, R. M. (2014). Engaging students in learning: Findings from a study of project-led education. *European Journal of Engineering Education*, *39*, 55–67. doi:10.1080/03043797.2013.833170

Fox, E. J. (2004). The Personalized System of Instruction: A flexible and effective approach to mastery learning. In D. J. Moran & R. W. Malott (Eds.), *Evidence-based educational methods* (pp. 201–221). San Diego: Elsevier Academic Press.

Goto, K., & Schneider, J. (2009). Interteaching: An innovative approach to facilitate university student learning in the field of nutrition. *Journal of Nutrition Education and Behavior*, *41*, 303–304. doi:10.1016/j.jneb.2009.02.003

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review, 16,* 235–266. doi:10.1023/B:EDPR.0000034022.16470.f3

Hunkins, F. P. (1968). The influence of analysis and evaluation questions on achievement in sixth grade social studies. Educational Leadership, 1, 326-332.

Hunkins, F. P. (1969). Effects of analysis and evaluation questions on various levels of achievement. *The Journal of Experimental Education, 38,* 45–58.

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test . . . or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review, 26,* 307–329. doi:10.1007/s10648-013-9248-9

Jones, S. E. (2007). Reflections on the lecture: outmoded medium or instrument of inspiration? *Journal of Further and Higher Education, 31,* 397–406. doi:10.1080/03098770701656816

Keller, F. S. (1968). "Good-bye, teacher...". *Journal of Applied Behavior Analysis, 1,* 79–89. doi:10.1901/jaba.1968.1-79

Kulik, C. C., Schwalb, B. J., & Kulik, J. A. (1982). Programmed instruction in secondary education: A meta-analysis of evaluation findings. *The Journal of Educational Research, 75,* 133–138.

Kulik, J. A., Cohen, P. A., & Ebeling, B. J. (1980). Effectiveness of programmed instruction in higher education: A meta-analysis of findings. *Educational Evaluation and Policy Analysis, 2,* 51–64.

Krathwohl, D. R. (2002). A revision of Bloom's Taxonomy: An overview. *American Journal of Psychology*, *41*, 212–218. doi:10.1207/s15430421tip4104

Krueger, P. M., et al. (2004). To the point: Reviews in medical education teaching techniques. *American Journal of Obstetrics and Gynecology*, *191*, 408–411. doi:10.1016/j.ajog.2004.02.003

Lambert, T., & Saville, B. K. (2012). Interteaching and the testing effect: A preliminary analysis. *Teaching of Psychology*, *39*, 194–198. doi:10.1177/0098628312450435

Maudsley, G. (1999). Do we all mean the same thing by "problem-based learning"? A review of the concepts and a formulation of the ground rules. *Academic Medicine*, *74*, 178–185.

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology*, *27*, 360–372. doi:10.1002/acp.2914

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.

Rosales, R., Soldner, J. L., & Crimando, W. (2014). Enhancing the impact of quality points in interteaching. *Journal of the Scholarship of Teaching and Learning*, *14*, 1–11. doi:10.14434/josotlv14i5.12746

Saville, B. K., Bethke, V., Asdourian, D., and Cairns, B. (2015). Interteaching: The impact of prep guides on quiz performance. Manuscript submitted for publication.

Saville, B. K., & Buskist, W. (2003). Traditional idiographic approaches: Small-N research designs. In S. F. Davis (Ed.), *Handbook of research methods in experimental psychology* (pp. 66–81). Malden, MA. doi:10.1002/9780470756973.ch4

Saville, B. K., et al. (2014). Interteaching and lecture: A comparison of long-term recognition memory. *Teaching of Psychology, 41*, 325–329. doi:10.1177/0098628314549704

Saville, B. K., Cox, T., O'Brien, S., & Vanderveldt, A. (2011). Interteaching: the impact of lectures on student performance. *Journal of Applied Behavior Analysis*, *44*, 937–41. doi:10.1901/jaba.2011.44-937

Saville, B. K., Lambert, T., & Robertson, S. (2011). Interteaching: Bringing behavioral education into the 21st century. *The Psychological Record*, *61*, 153–166.

Saville, B. K., Pope, D., Lovaas, P., & Williams, J. (2012). Interteaching and the testing effect: A systematic replication. *Teaching of Psychology*, *39*, 280–283. doi:10.1177/0098628312456628

Saville, B. K., Pope, D., Truelove, J., & Williams, J. (2012). The relation between GPA and exam performance during interteaching and lecture. *The Behavior Analyst Today*, *13*, 27–31.

Saville, B. K., & Zinn, T. E. (2009). Interteaching: The effects of quality points on exam scores. *Journal of Applied Behavior Analysis*, *42*, 369–74. doi:10.1901/jaba.2009.42-369

Saville, B. K., Zinn, T. E., & Elliott, M. P. (2005). Interteaching versus traditional methods of instruction: A preliminary analysis. *Teaching of Psychology, 32,* 161–164.

Saville, B. K., Zinn, T. E., Neef, N. A., Van Norman, R., & Ferreri, S. J. (2006). A comparison of interteaching and lecture in the college classroom. *Journal of Applied Behavior Analysis, 39,* 49–61.

Skinner, B. F. (1954). The science of learning and the art of teaching. *Harvard Educational Review, 24,* 86-97. (Reprinted in *Cumulative record, definitive edition,* pp. 179-191, by B. F. Skinner, 1999, Cambridge, MA: B. F. Skinner Foundation).

Skinner, B. F. (1974). Designing higher education. *Daedalus, 103,* 196–202.

Skinner, B. F. (1984). The shame of American education. *American Psychologist, 39,* 947–954. doi:10.1037/0003-066X.39.9.947

Slagter, T. H., & Scribner, D. L. (2014). Interteach and student engagement in political science. *Journal of Political Science Education, 10,* 37–41. doi:10.1080/15512169.2013.835562

Tsui, M. (2010). Interteaching: Students as teachers in lower-division sociology courses. *Teaching Sociology, 38,* 28–34. doi:10.1177/0092055X09353887

Woods, D. R. (2014). Problem-oriented learning, problem-based learning, problem-based synthesis, process oriented guided inquiry learning, peer-led team learning, model-eliciting activities, and project-based learning: What is best for you? *Industrial & Engineering Chemistry Research, 53,* 5337–5354. doi:10.1021/ie401202k

Zeller, B. E. (2010). “We learned so much when you weren’t there!”: Reflections on the interteach method and the acephalous classroom. *Teaching Theology and Religion*, *13*, 270–271.

Zinn, T. E., & Saville, B. K. (2007). Interteaching: A new approach to peer-based instruction. *Psychology Teacher Network*, *17*, 19–22.