INVESTIGATING THE VARIABLES IN A MOCK EXAM STUDY SESSION DESIGNED TO IMPROVE STUDENT EXAM PERFORMANCE IN AN UNDERGRADUATE BEHAVIOR MODIFICATION AND THERAPY COURSE

By

Wesley H. Dotson

Submitted to the graduate degree program in Applied Behavioral Science
and the Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy.

Chairperson _____

_____

Committee members _____

_____

_____

Date Defended: _____

The Dissertation Committee for Wesley H. Dotson certifies
that this is the approved version of the following dissertation:

INVESTIGATING THE VARIABLES IN A MOCK EXAM STUDY SESSION DESIGNED TO IMPROVE STUDENT EXAM PERFORMANCE IN AN UNDERGRADUATE BEHAVIOR MODIFICATION AND THERAPY COURSE

By

Wesley H. Dotson

Committee:

_____
Chairperson

_____
Chairperson

_____

_____

_____

Date Approved: _____

CONTENTS

List of Tables

List of Figures

Abstract

The purpose of the present study was to identify components of an optional mock exam review session (e.g. requiring students to write answers, providing students grading keys for questions) responsible for improvements in student performance on application-based short-essay exams in an undergraduate behavior modification course. Both within-subject and across-groups comparisons were made across three studies within the larger investigation. The primary dependent variable across studies was student accuracy on exam questions. Additional measures of extra credit earned, class attendance, mock exam attendance, and entering GPA were also gathered and analyzed using correlation and multiple linear regression analysis. Students attending mock exam sessions scored higher on exams than students who did not. Students did not score higher when required to write answers versus when not required to write answers. Students also did not score higher when required to discuss a question versus when asked to listen to the GTA discuss a question. A package of components involving discussion, evaluation, and correction of a sample answer to a question produced superior performance over providing students with copies of study materials about a question. Student GPA entering the course, number of mock exams attended, and amount of extra credit earned were significantly predictive of final grade earned in the course, but attendance at lectures was not significantly predictive.

Investigating the variables in a mock exam study session designed to improve student exam performance in an undergraduate behavior modification and therapy course

Introduction

Instructors use many methods to support and improve student learning in college courses. Methods of support include several kinds of procedures and contain a number of components, each of which may have individual or cumulative effects on how students learn material. Procedures involve redesigning course structures (e.g. programmed instruction, Personalized Systems of Instruction, Interteaching), introducing active responding to the classroom (e.g. clickers, daily quizzes, etc.), holding review sessions outside of class, providing materials outside of class (e.g. practice exams, study guides, lecture notes), and arranging different feedback contingencies within the course. These procedures contain a number of components including: providing students practice opportunities, giving students feedback on their performance, exposing students to course goals and learning objectives, asking students to evaluate samples of work, giving students products that could be taken home to be studied, having students actively participate in the class, and designing review opportunities that closely resemble the actual exam.

Some of the methods of support have been more effective than others, and some require more effort and restructuring of the course than others. For example, while courses offered in the Personalized System of Instruction (PSI) format have consistently produced higher rates of student learning compared to more traditional lecture courses (Keller, 1968; Kulik, Jaksa, & Kulik, 1978; Taveggia, 1976), the majority of instruction continues to be delivered using lecture format (NCES, 2002). One of the reasons most

often given for the disuse of the PSI format is the effort involved in creating materials and managing the course (Buskist, Cush, & DeGrandpre, 1991; Lloyd & Lloyd, 1986). For example, Lloyd & Lloyd (1986) surveyed professors who had published PSI research. Many of the people surveyed noted that they used PSI course structures less than they had previously, with the primary reasons for that decline being costs in time such as "training proctors, creating numerous sets of study and testing materials, extensive bookkeeping, [and] negotiating for extra space and assistants." This suggests that one reason for the lack of widespread adoption of PSI course structure is the substantial additional effort required to improve student performance. If so, then effort must be made to identify procedures or components of procedures that produce improvements in student learning that can also be implemented easily and within common course structures such as those relying primarily on lectures. The identification and evaluation of these components could lead to the development of techniques more likely to be adopted and used within lecture courses.

To begin that process, I reviewed the empirical literature on improving student exam and quiz performance to identify (a) the most and least effective procedures, (b) the components of common procedures most likely to improve performance, and (c) the variables deserving of more careful evaluation in future studies. I identified studies that involved inquiry about or manipulation of procedures and variables including: practice opportunities, review sessions, specifying course goals and learning objectives, asking students to evaluate samples of work, giving students products that could be taken home to be studied, and providing students with immediate feedback on their performance.

I identified an initial pool of studies by conducting keyword searches on the PsychINFO database. Keywords included: undergraduate education, exam performance, test performance, study session, review session, study guide, clicker, response card, essay exam, college education, lecture notes, crib sheet, practice exam, review activity, and quiz performance. From that pool, I identified further articles by reviewing the reference sections of studies found in the initial search. I identified over 200 articles in the initial search. This number was reduced by applying the following criteria for inclusion:

1. The article was empirical: I only included an article if the reported results included quantitative data about student performance. Thus, a case report or an article that simply reported an effective procedure without accompanying behavioral measures was not included in the review.
2. The article investigated the impact of an instructional intervention other than a course structure manipulation or type of assessment manipulation on student performance: I only included an article if the question answered involved a procedure within a course rather than the overall structure or assessment of a course. Articles that addressed interventions that could be conducted within an existing course (e.g. the effect of giving different types of feedback) were included for review. Interventions that targeted larger, course structure-related questions were not included. For example, studies comparing PSI versus lecture course arrangements or multiple-choice versus essay-based examinations were not included.

3. The article investigated the impact of an intervention on student quiz or exam performance: I only included an article that reported exam or quiz performance as a dependent variable of interest. Thus, I did not include studies in the review that evaluated only student behaviors such as preference for an instructional technique, participation in class, or attendance.

4. The article involved undergraduate or graduate level students.

A total of 40 articles met the inclusion criteria above and were reviewed. The general results of the review are presented below. I grouped the articles by the form of the procedure evaluated in the study because that represents the most common method of classifying procedures. The grouping of procedures was as follows: procedures to increase active responding during class periods, review sessions conducted outside of class periods, providing materials to be used or completed outside of class, and providing feedback to students about their performance on quizzes or exams.

I analyzed several dimensions of each study: (a) the specific independent and dependent variables (e.g. materials provided to students, the format of the exam), (b) the experimental design of the study, (c) the results of the study, (d) the components present in each procedure used (i.e., the degree of similarity between the educational support materials and the actual exam or quiz, the type of student response required, whether or not specific learning objectives or a question pool were provided, and whether or not sample answers or feedback were given to students), and (e) the amount of improvement in student performance as a result of using the intervention procedures (when it was possible to calculate the improvement). Not every reviewed article provided sufficient

information to address each component, and in those cases I did not analyze those components.

Whenever possible, performance of students was converted to a percentage correct measure. For example, if the study reported the average number of questions answered correctly on a 20-question quiz, the total number correct was divided by 20 and the result multiplied by 100 to calculate a percentage correct measure. The average performances across conditions, then, were compared with each other to determine a percentage of improvement for each intervention. While it would be preferable to use a more precise and established measure of effect size such as Cohen's *d*, there was rarely sufficient information presented in the articles to calculate such measures. The more easily determined conversion to percentage improvement did, however, allow a comparison of studies of many different types with many different assessment instruments.

I first reviewed procedures to increase active responding during class periods. I defined a procedure as involving active participation if it required students to engage in some behavior (e.g. taking a short daily quiz, holding up a response card, saying an answer, writing an answer) in response to a question during the class period. My literature search identified fifteen studies evaluating procedures designed to increase active responding during class periods: six involved the use of clickers or response cards (Clayton & Woodard, 2007; Kellum, Carr, & Dozier, 2001; Malanga & Sweeney, 2008; Morling, McAuliffe, Cohen, & DiLorenzo, 2008; Poirier & Feldman, 2007; Shabani & Carr, 2004), one involved adding interactive windows to class lectures (Huxham, 2005), two dealt with asking students to answer questions as they worked in class (Miller &

Malott, 1997, 2006), three involved giving in-class quizzes (Landrum, 2007; Nevid & Mahon, 2009; Padilla-Walker, 2006), and three evaluated requiring students to write during class periods (Drabick, Weisberg, Paul, & Bubier, 2007; Hautau et al., 2006; Simon, 2005).

The first type of educational support I reviewed was electronic response devices (clickers) and response cards. Response cards and clickers are often used to ask students multiple-choice questions. Typically, the instructor asks a question to the class and displays several answer options that the students choose from. If response cards are used, the students hold up the card with their answer to the question on it. If clickers are being used, the students press the button on the device corresponding to their answer. The instructor then displays the correct answer to the question to the entire class. Students thus get immediate feedback about the accuracy of their answers. Response cards and clickers typically involve asking questions in multiple-choice format and thus have a similar form to the quizzes and exams reported in these studies (all multiple-choice). While instructors use response cards and clickers to cover material important for students to know for assessments, the questions asked represent a very small proportion of the material actually covered.

In a study examining the effect of using response cards with one of two sections of undergraduates in an introductory psychology course, the authors reported a minimal effect on weekly multiple-choice quiz scores for the group using the response cards (Clayton & Woodard, 2007). The average improvement appeared to be about 4% for students using response cards. Another study (Morling, et al., 2008) evaluated the use of clickers within 2 of 4 large sections of an introductory psychology course and found only

a 1.5% average improvement for the sections using clickers. Researchers reported similar results in a study across two sections of an introductory psychology course (Poirier & Feldman, 2007), with a 1.31% average improvement on multiple-choice exam performance for students in the section using clickers versus students in the section not using clickers.

Another study (Kellum, et al., 2001) used an alternating treatments design to evaluate the effects of response cards on multiple-choice quiz performance. On different days, students either used or did not use response cards, and the instructors then compared students' performance on daily quizzes given at the end of each class. The authors reported a small improvement in performance on daily quizzes on response card days, but the actual scores were only represented in a graph, and no average percentage on improvement could be calculated. The authors did not report any longer-term measures such as performances on course exams. Another set of researchers attempted to replicate the findings of the Kellum et al. (2001) study and add an additional analysis of the effect of response card use on later multiple-choice exam performance (Shabani & Carr, 2004). The authors replicated the results of the Kellum, et al (2001) study for quiz performance, but found no difference on exam performance for material covered on response card days and non-response card days.

The studies reporting no or minimal improvements for using clickers or response cards did not compare the effects of their use to the effects of using different active methods. A study by Malang and Sweeny (2008) did so. The authors compared the use of response cards during class to asking students to write answers to study questions at the end of class and found students performed much better on multiple-choice quizzes

when required to write answers to study questions. In alternating weeks of an introductory behavior analysis course, students either used response cards to answer questions during lectures or wrote answers to 5 study questions at the end of each class period. They took a quiz at the end of each week. Study questions led to a 12.5% improvement in performance on the weekly quizzes when study questions were used and no improvement in performance when response cards were used. The results of the Malanga & Sweeney (2008) study suggest the type of active responding by students may also play a crucial role in achieving improvements in student responding. In that study, when students were completing the study questions, they were engaging in a behavior more similar to what they would be doing on the weekly quiz than holding up a response card or pushing a button on a clicker.

The next type of procedure I reviewed was the use of interactive windows. One study (Huxham, 2005) examined the effect of brief "interactive windows" inserted into lectures on students' performance on multiple-choice course quizzes and two short-essay questions on the final exam. The course was lecture-based, and the instructor inserted several brief opportunities for students to discuss ideas or problems from the lecture among themselves. The opportunities for students to discuss ideas were called "interactive windows." The instructor compared student performance on exam questions that were about material discussed during one of the interactive windows with performance on questions about material not discussed in an interactive window but covered during the lectures. The author reported a slight, but significant, difference in performance, with students doing better on questions about the material covered by the interactive windows than they did on questions about material covered only in lecture.

Next, I reviewed requiring students to answer questions while reading course material. Two studies (Miller & Malott, 1997, 2006) explored the effect of requiring students to engage in different types of responses while completing computerized course activities. In both studies, students alternated between conditions that either required them to type answers to study questions while completing an online module or that required them only to read passages about course material. In both studies, the questions were similar in form and content to the questions on multiple-choice section exams. Students consistently performed better on exams in the conditions in which the active answering of questions was required. The average improvement in performance for the active condition was between 11% and 16%.

Another procedure reviewed was the use of short, in-class quizzes. Three studies investigated the effect of giving short quizzes on exam performance (Landrum, 2007; Nevid & Mahon, 2009; Padilla-Walker, 2006), and all three reported slight positive effects of giving the quizzes. Padilla-Walker (2006) conducted a correlation analysis of student participation and accuracy in completing daily, short-answer, extra-credit quizzes about assigned readings in an advanced psychology course and found that student performance on those quizzes was highly correlated with multiple-choice exam performance in the course. Nevid & Mahon (2009) administered a brief, one-question multiple-choice quiz at the beginning and end of six class periods of an introductory psychology course and then analyzed student performance on multiple-choice exam questions in three categories: exam questions about material included in one of the short quizzes, exam questions about material covered on a quiz day, and exam questions about material covered on a non-quiz day. Students performed best on questions about material

that had been included on one of the short quizzes. Their next best performance was on questions about material covered on a quiz day. Students performed least well on questions about material covered on non-quiz days. The differences between groups were statistically significant.

Another study reported similar results (Landrum, 2007). In each week of an introductory psychology course, students took a multiple-choice quiz. On the multiple-choice final exam, students answered questions of three kinds: those appearing exactly as they had on weekly quizzes, those appearing on a weekly quiz but with scrambled answer options, and those not appearing before on a weekly quiz. Students performed better on questions that they had seen before than on questions they had not seen. Students performed equally as well on questions with and without scrambled answer options. The authors reported that the differences were statistically significant.

The last type of procedure I reviewed for this section was asking students to write in class. Other procedures required students to engage in writing activities during class periods. One study (Drabick, et al., 2007) examined the impact on multiple-choice exam scores of asking students to complete brief writing exercises about a topic versus asking them to spend the same amount of time "thinking about" another topic. Even when all other individual factors had been controlled (e.g. gender, entering GPA), students consistently performed better on multiple-choice exam questions about material from the brief-writing exercises than questions about material from the thinking-only exercises. Students had an average of 4.5% improvement on exam questions about material covered in the brief writing exercises compared to questions about material covered in the thinking-only exercises.

In another investigation of brief writing exercises and the impact of different feedback contingencies on their effectiveness (Hautau, et al., 2006), the investigators found that students performed best when they received credit for every writing assignment than when they only got credit for random assignments. During some class periods of an introductory psychology course, students were asked to write for several minutes about a topic covered during lecture. In one section of the course, students received credit for participating in every writing assignment for each unit. In another, students received credit for one randomly selected writing assignment for each unit. In the third section students did not receive any credit for the writing assignments. Students performed the least well when no credit was given. Students receiving credit for every writing assignment had an average improvement on multiple-choice exam performance of 9.7% versus students receiving no credit. When receiving credit for a random sample of writing assignments, the average improvement was 3.7% versus receiving no credit.

Another investigator explored the impact of required written rehearsal and study guide completion on student performance on fill-in-the-blank exams and found that students did the best on questions for which they had both engaged in the in-class written rehearsal and completed the optional study guide (Simon, 2005). Both the study guide and written rehearsal required students to answer questions in the same format as that used on the exam and both also covered similar content as the actual exams. Students received feedback about their performance for the written rehearsal but not on their answers to the study guide questions. When the study guide alone and the study guide with written rehearsal were compared in a later phase of the same study, however, the written rehearsal did not produce any improvements in performance on exam questions

relative to questions for which students only completed the study guide. Since students did very well on the exams, there may have been a ceiling effect where either type of practice produced enough improvement in performance to make it difficult to see any differences between the procedures.

When looking at the results of all of the studies investigating procedures to increase active responding during class periods, participation in class activities alone does not appear to guarantee students will show more than a minimal improvement (less than 5%) on exams, and more realistic or demanding activities may be necessary to produce larger improvements in students' exam-taking. It is not clear that the use of response cards or clickers produces improved exam performance for the majority of students, since the results of many of the studies report either no or minimal improvements for using active response devices on assessments. An interesting finding across the studies is that all of the studies investigating the effectiveness of clickers or response cards report increased rates of student responding during class periods, and the increased rate of student participation is often offered as one of the primary benefits of the use of such systems. That such increased participation does not appear to be correlated with equally significant improvements in the performance of students on course assessments, however, suggests that merely being more active in a class is not a sufficient condition to improve student performance. The results of the Malanga & Sweeney (2008) study suggest the type of active responding by students may also play a crucial role in achieving improvements in student responding. In that study, when students were completing the study questions, they were engaging in a behavior much more similar to what they would be doing on the weekly quiz than just holding up a response card or pushing a button on a

clicker. More realistic practice opportunities were more effective (Malanga & Sweeney, 2008; Miller & Malott, 1997, 2006; Nevid & Mahon, 2009; Padilla-Walker, 2006), and in the most successful interventions, students wrote answers to questions rather than only talking or holding up a response card.

The next type of educational support procedures I reviewed were those that involved offering review sessions. For the purposes of this review, review sessions were defined as any review activity (study session, administration of a practice exam, question and answer opportunities with instructors) that occurred outside of the regular class meeting times and was mediated by a course instructor or teaching assistant.

Study sessions to prepare for an exam are common on college campuses but have rarely been empirically evaluated. Two studies explored the effectiveness of a study session at improving student performance on a multiple-choice final examination in different sections of an introductory psychology course (Aamodt, 1982a, 1982b). In the first study (Aamodt, 1982b), students who attended a study session involving a review of important material and a question and answer period by the course graduate teaching assistant scored higher on a cumulative final exam than students who did not attend. The average improvement for the students attending the session was 8.9%. In an attempt to determine which aspect of the study session was most helpful, Aamodt conducted a second study in which he offered two study sessions: one structured like the first study session and the other only involving the question and answer portion but not the review of important material (Aamodt, 1982a). He found students who attended the study session where key information was reviewed scored better on the final exam than students who did not attend, and students attending the question and answer only session

did not score any higher than students not attending either session. The average improvement for the students attending the study session with the review component was 6.5%, but for those attending the question and answer session the average improvement was 0%. These results suggest the graduate teaching assistant reviewing key information with students (thus making them aware of the instructor's expectations of what they should know) the night before an exam was responsible for the improved performance of the students who attended the session.

Researchers reported mixed results in studies comparing the effectiveness of two different types of study sessions (one a games-based session, and the other a question and answer session) in the same graduate education course (Neef et al., 2007). In the first of two studies in the report, students who participated in the games-based review session performed better on multiple-choice assessments than students who did not attend the review sessions. The average improvement was 9.5%. In the second study, attendance at the games-based session led to only a 2% average improvement. The games-based review sessions involved students answering trivia questions about the quiz material and receiving feedback about their answers, but the questions were not in the format of the actual quizzes and did not comprehensively cover the material that would appear on the quiz. In both studies, the question and answer sessions did not lead to any improvement in student performance. The authors report that one reason for the mixed results may have been that the review sessions were held immediately before class periods in which quizzes were to be given, and the graduate students taking the course reported doing the majority of their studying long before the review sessions began.

A study by Rust, Price, and O'Donovan (2003) reported a positive and persistent effect of study sessions on student performance on end-of-course essay assessments. Across two years of a large-enrollment, undergraduate business course, the authors offered a 90-minute study session four weeks before the final, open-ended assessments were turned in by students. Every student in the course received a set of grading criteria and two sets of sample answers to the assessment questions they would be completing at the end of the course. One week later the instructors offered an optional 90-minute study session. Students attending the session were asked to evaluate the sample answers according to the grading criteria and bring the completed evaluations to the optional study session. Once at the session, students worked in small groups to discuss their grading, shared their grading with the larger group, listened to an instructor/grader describe how the question would be graded, discussed their grading again in light of that description, and then finally viewed and discussed the instructor/grader's specific evaluation of the same sample answer. There were no significant differences between the two groups of students (those attending and those not attending the study session) on their performance in a prior course, suggesting there were not differences in ability and motivation between the two groups even though they were self-selected. Following the study session, however, there were statistically significant differences in course performance between the two groups, with those attending the study session scoring higher in the business course than those who did not attend the study session. The average improvement for the students attending the session was 12% for one cohort of students and 14% for the other. Those differences in performance persisted in a third business course students took a year later, with students who had attended the study

sessions in the second course scoring higher on average in the third course than those students who did not attend the session.

A later study by the same authors reported an attempt to transfer the management of a similar review session over to the undergraduates with less successful results (Price, O'Donovan, & Rust, 2007). In this study, students worked together in small groups and engaged in a peer review of answers each of them had written to the assessment questions. No sample answers or explanation of the grading criteria were provided. Students only reviewed each other's work; no grader from the course explained evaluation criteria or showed sample answers being evaluated. It appeared the removal of the guidance of the grader in the process to explain and model correct and thorough evaluation severely impacted the effectiveness of the program, as there was only a 1.6% average improvement in performance on the end-of-course assessment between students who attended the peer-review workshop and those who did not.

Another type of review activity that has been evaluated involved offering students an opportunity to complete a practice exam. Practice exams typically allow students to answer questions about course material in a format that resembles the actual exam.

One study compared a completion versus an accuracy contingency for a practice exam and evaluated the effect on student performance on multiple-choice course exams (Oliver & Williams, 2005). Both the completion and accuracy contingency groups received extra credit for answering practice exam questions that were similar in form (multiple-choice) and content to the actual exam, but in the accuracy group, the students had to answer the questions correctly to get the extra credit. Both groups received a review session led by the instructor in which the correct answers to all questions were

reviewed and the most common mistakes discussed. The students in the accuracy contingency group scored higher on the course exams than the students in the completion groups (on average, a 6% improvement). The smaller effect, however, may have been a result of the review session activities of the instructor, where even students who may not have answered practice exam items correctly had an opportunity to identify their mistakes with time to study and correct them for the actual exam. It would have been helpful to have included a comparison group who did not have access to the practice exam at all to evaluate the overall effectiveness of the practice exam.

A similar study (Balch, 1998) compared the final exam performance of students completing a practice exam with the performance of students attending a more traditional review session. Two groups of student volunteers each met separately in the days before a final exam in an undergraduate psychology course. One group completed a practice exam (similar in form and content to the actual exam) under test-like conditions and then graded and discussed their answers with the guidance of the course instructor. The other group also viewed the practice exam, but instead of completing the answers, they filled out a rating sheet asking if they thought the questions were relevant and likely to be on the actual exam. That group also then reviewed the correct answers to the practice exam and discussed them with the guidance of the course instructor. Students who had completed the practice exam under exam-like conditions scored higher on the multiple-choice final exam than students who attended the more traditional review with an average improvement of 4%. The practice exam made the biggest difference for students who had performed the worst in the course leading up to the final exam, suggesting more realistic practice may be most beneficial to struggling students.

Another study (Bol & Hacker, 2001) also explored the effect of a practice exam on student exam performance. The experimenters compared the multiple-choice and short-answer exam performance of graduate students across two conditions: a practice exam condition where students reviewed by completing a practice exam of similar form to the actual exam, and a traditional review condition where students reviewed the course material with the course instructor. It is not clear exactly what each condition specifically involved, but the authors report students performed equally as well on course exams regardless of what type of session they attended.

Overall, the literature on outside-of-class-time review sessions and practice exams reported mixed results. Study sessions that included a review of performance expectations and realistic practice (Rust, Price, & O'Donovan, 2003) were the most effective, followed by the sessions involving only a review of performance expectations (Aamodt, 1982a, 1982b). Study sessions that included only question and answer opportunities (Aamodt, 1982a; Neef, et al., 2007) or peer feedback and grading without instructor or grader support (Price, et al., 2007) did not produce large improvements in exam performance. Practice exams appeared to be effective, but the improvements were generally small to moderate. The least successful interventions involved graduate students as participants. Characteristics of those students (e.g., increased time spent studying independently and advanced ability to record and organize key material from lectures) probably made it unlikely review sessions would show much benefit. Additionally, the review sessions in the Neef et al. (2007) study were held immediately before the class period in which the quizzes were given. That arrangement made it difficult for students to study in an effort to improve any areas the review sessions

indicated the students had not mastered. As in the interventions to increase active responding during class periods, the studies reporting the largest improvements generally involved review activities that closely resembled the actual exam materials, required students to write answers, and in which the instructor or GTA provided an explanation of the performance expectations for the course (Balch, 1998; Oliver & Williams, 2005; Rust, et al., 2003).

The next type of educational support procedure I reviewed involved providing students with materials they could use to study. For the purposes of this review, interventions in this category involved providing students with materials such as a list of desired instructional objectives or study questions designed to help the students prepare for course exams or providing lecture notes to students for the same purpose. Any provided materials had to be available for students to access outside of class periods.

One type of provided material is learning objectives. One study investigating the effect of providing learning objectives (Jenkins & Neisworth, 1973) found a positive effect for providing accurate learning objectives to students. The authors randomly divided students in an introductory educational psychology class into two groups. Each group received 10 instructional objectives for an upcoming multiple-choice quiz. For each group, five of the objectives were accurate; they accurately predicted material to be included on an exam question (i.e. they provided questions that would be asked on the quiz). The other five objectives were inaccurate; they did not predict accurately material to be covered on the exam (i.e. they provided questions that would not be asked about on the quiz). The accurate versus inaccurate objectives were counterbalanced across the two groups. Students performed significantly better on questions for which they received

accurate instructional objectives (61% correct versus 34% correct, on average), with an average improvement of 27% for accurate objectives over inaccurate objectives. When students knew what to expect, they answered those questions more accurately.

A less-successful attempt to improve student exam performance by increasing knowledge of instructor expectation was undertaken by Fleming (Fleming, 2002). Students in an introductory psychology course were given 6 short lectures at the end of the first 6 class periods discussing general study tips and ideas for doing well in the course. In addition, students were instructed in how to set academic goals. When compared to students in another section of the same course, students who had received the study tips and goal setting instruction did not perform any better on multiple-choice course exams. As contrasted with Jenkins and Neisworth (1973), Fleming only provided general, non-specific recommendations early in the course –at a time when students were not likely to be motivated to attend to them.

Another resource sometimes provided to students in an effort to help them prepare for or take exams involved letting them make or use crib sheets. Crib sheets are pieces of paper (typically note cards) on which students are allowed to write information that might help them on an exam. Typical information included on crib sheets includes key points from notes, definitions of terms, or mathematical equations. Instructors usually allow students to make their own crib sheets and use them during exam periods. Because they are supposed to contain prompts about and examples of important materials to be covered on the exam, crib sheets were assumed to be similar in content to the actual exams. In two studies, Dickson and colleagues explored the effect of allowing students in introductory courses to construct and use crib sheets on exams (Dickson & Bauer, 2008;

Dickson & Miller, 2006). In one study (Dickson & Bauer, 2008), students were allowed to prepare crib sheets for an upcoming exam. Before the exam was administered, students were asked to complete a brief quiz containing some of the exam questions, but they were not allowed to use their crib sheets. Students then took the exam and were allowed to use the crib sheets. Students did significantly better on the exam questions when allowed to use the crib sheets than they did on the same questions on the quiz when they were not allowed to use them. The percentage of improvement when crib sheets were allowed ranged from 7-19%. The results suggest students anticipating being able to use crib sheets may not prepare as thoroughly as students not expecting to use crib sheets. In the second study (Dickson & Miller, 2006), students were allowed to prepare and use crib sheets for two of four multiple-choice course exams. For the other two exams, students were told they could not use crib sheets, but at the time exams were handed out the instructors gave students a copy of a prepared crib sheet created by the instructors. Students performed better on the two exams on which they used the instructor-prepared crib sheets, with an average improvement of 5.5% over the exam performance when they used the crib sheets they prepared. The results demonstrate that students do not prepare crib sheets that are as useful as those produced by the course instructors. Overall, both studies suggest that crib sheets hinder students learning material as well as they might if they knew they would not be allowed to use crib sheets.

There are a number of studies exploring the effects of providing study guides or sets of study questions on student performances on exams and quizzes. In two studies, Dickson and colleagues explored the effect of study guides on student exam performance (Dickson, Devoley, & Miller, 2006; Dickson, Miller, & Devoley, 2005). In the first

study (Dickson, et al., 2005), students in two sections of an introductory psychology class both had access to an online study guide. In one section, students were required to complete the study guide as part of their course grade. Students in that section performed better on course exams than students in the section that did not require the completion of the study guide. There was only a 2.8% difference between the two groups, with the students in the required completion section performing better. This small difference may have been because the study guide contained many questions about material not on the actual exam (only 21% of exam questions came from material in the study guide), and the questions in the study guide were not of the same format as the exam students took in class (study guide questions were in short-answer, true/false, multiple-choice, and fill-in-the-blank format while the actual exam was only in multiple-choice format). The second study (Dickson, et al., 2006) explored the second question raised in the 2005 study by comparing student performance on exams when a provided study guide only contained questions in the format of the actual exam versus when the study guide contained questions of many different forms. The experimenters found no difference in performance on exams between the two study guide conditions.

In a study similar to Dickson et al. (2005), Flora and Logan (1996) evaluated the effect of a completion contingency for a computerized study guide (a practice exam) on multiple-choice exam performance as compared to simply making the practice exam available but not requiring students to complete it. Students who completed the practice exam received feedback about the accuracy of their answers when they submitted them to the course website (delivered automatically). Students who were required to complete the practice exam performed an average of 2% better on the multiple-choice course

exams than those who were not required to do so. Like the Dickson et al. (2005) study described above, however, the small effect size may have occurred because the study guide bore little resemblance to the actual exam, either in the form of the questions asked or the content of the guide.

Another study, reporting a much more effective intervention, evaluated the effectiveness of study guides at improving student performance on multiple-choice exams (Miles, Kibler, & Pettigrew, 1967). The authors provided students in a beginning educational psychology course a list of 80 study questions with which to help prepare for a multiple-choice exam. On the actual exam, half of the questions came from the study guide, and the other half did not. Students performed much better on the same questions they had seen before on the study guide, with an average improvement of 14% on the questions that appeared on the study guide.

Semb and colleagues (1973) also evaluated the effect of providing study questions on quiz and exam performance in an undergraduate human development course. They gave students a set of study questions, and like the Miles et al. (1967) study above, included some of those questions on multiple-choice quizzes and a multiple-choice final. The researchers also included additional questions not found in the pool of study questions on the quizzes and final exam. On the final examination in the course, student performance was evaluated on four types of questions: those that appeared on the study guide and had been on an earlier quiz, those that had only appeared on the study guide, those that had only appeared on an earlier quiz, and those that had not appeared in either the study guide or on an earlier quiz. Students performed best on the questions they had seen twice before (33.5% average improvement versus questions not seen before). They

performed better on questions that had only been on the study guide (28% average improvement) than on questions that had only been on an earlier quiz (19.25% average improvement) versus questions they had not seen before.

While the results of studies evaluating providing study guides have been mixed, positive effects occurred most often when the study guide or study questions contained materials that were both relevant and presented in a way closely resembling the format of the actual exam or quiz (Miles, et al., 1967; Semb, Hopkins, & Hursh, 1973).

Providing lecture notes is another way students have been exposed to course expectations and important information. Providing lecture notes provides students information about the important content to be learned in the course and presumably helps them study more effectively. Several studies have investigated providing notes to students and the effects on their multiple-choice exam performance. In one study (Grabe & Christopherson, 2008), the experimenters conducted a correlational analysis between students accessing lecture notes made available online and performance on exams related to the notes and found a slight positive correlation between the two. Because no actual exam scores were reported, it was not possible to calculate a percentage of improvement as a result of providing lecture notes in this study.

Another study (Hove & Corcoran, 2008) also explored the effects of providing lecture notes online and how doing so affected student attendance and performance on course exams. Two groups of students in two sections of an introductory psychology course participated. One section was given access to lecture notes during the semester, and the other section was not. The authors report students in the section receiving lecture notes performed significantly better on course exams than students who did not have

access to lecture notes, but no actual exam scores were reported. Additionally, even though there was a strong correlation between attendance and the final grade earned in both sections of the course, the actual rates of attendance were no different between the two sections, suggesting students did not attend class periods less often due to the availability of lecture notes online.

Two researchers (Cornelius & Owen-DeSchryver, 2008) explored the effects of providing two different types of lecture notes to students in four sections of an introductory psychology course. Two sections of the course were given full lecture notes each class period, and students in the other two sections received partial lecture notes. On the last two of four course exams, students who received partial notes did better on the exams than students receiving full notes. The difference was 2.5% between those receiving partial notes and those receiving full notes, with those receiving partial notes scoring higher. Unfortunately, no comparison was done with a section of students receiving no lecture notes to provide an estimate of the effect of providing any type of notes on exam performance. As in the Grabe and Christopherson (2008) and Hove and Corcoran (2008) studies, no negative relation between providing lectures and attendance in class was found.

In general, providing materials of any kind to students produced positive effects on student exam and quiz performance, with the largest effects in cases where the provided materials contained predominately relevant material and explicit descriptions of performance goals or expectations (Jenkins & Neisworth, 1973; Miles, et al., 1967; Semb, et al., 1973). In the cases in which providing supplementary materials was not helpful, it appeared the materials provided did not contain a large proportion of relevant

material (Dickson, et al., 2005), or the recommended tips were not specific enough to benefit students on individual exams (Fleming, 2002).

The last type of educational support reviewed was the arrangement of feedback provided to students. Because the importance of delivering feedback is well-established as necessary for learning to occur, the review in this area dealt only with studies exploring how different types and timing of feedback affected student quiz or exam performance. For the purposes of this review, a study was defined as involving arrangement of feedback if the question being asked evaluated different levels or timing of feedback on student performance.

In a study evaluating the effects of delivering feedback at different times for performance on multiple-choice exams (Brosvic & Epstein, 2007), the experimenters found more immediate feedback on unit exams led to improved performance on a final exam and improved retention of material at 3-, 6-, 9-, and 12-month follow-up assessments. The authors compared accuracy rates on multiple-choice questions between 4 conditions: immediate feedback after each question, feedback at the end of the exam, feedback 24 hours after the exam, and feedback during the next class period after the exam. The immediate feedback after each question condition produced superior performance across all assessment conditions, with an average improvement of 10-15% across conditions on the cumulative final exam. Questions in the feedback at the end of the exam condition and the feedback 24 hours after the exam condition were answered more accurately than those from the next class period after exam condition on the final exam, but the effect was small (3-4% average improvement) and did not show any improved retention in any of the follow-up assessments.

Another study (Buzhardt & Semb, 2002) also explored differences in student performance between an item-by-item feedback condition and an end-of-exam feedback condition when answering multiple-choice questions on a cumulative final exam. They did not find any differences in student performance on questions between conditions. This result may have occurred because students had the opportunity to take the final exam twice, and only the higher of the two scores was used for the analysis. The authors reported most students took the exam twice. Thus, students probably studied the material they did the worst on between attempts. This studying between attempts likely eliminated any differences between conditions, if they existed, because students may have studied the questions from whatever condition produced the least learning during the initial instructional exercises. Any differences between the two feedback conditions most likely would have appeared in an analysis of only the first exam attempt of each student. Because such an analysis was not done, the failure to find a difference cannot be interpreted as indicating the equivalence of the two feedback conditions.

Two investigators (Dorow & Boyle, 1998) looked at the effect of different feedback contingencies on writing behavior and the effect of those conditions on a post-test writing exam. Several undergraduate students in an introductory English class were selected to participate in the study and were randomly placed in one of three groups: a specific feedback group (grading criteria given, total number of points earned noted, and errors marked in writing), a non-specific feedback group (grading criteria given and total points earned noted, but paper not marked for specific errors), and a no feedback group (only told how many total points earned but no grading criteria given or specific errors marked). They then completed several short writing tasks (receiving the different types

of feedback) before taking a post-test writing exam. Compared to their performance on a pre-test writing exam, students in the specific feedback condition had significantly fewer spelling and grammatical errors as a proportion of the total words written and wrote longer answers. Students in the non-specific feedback condition made fewer errors but did not write longer answers. Students in the no-feedback condition performed the same on the post-test as they did on the pre-test. It was not possible to calculate the average improvement across the different conditions because the measures used did not involve a fixed potential amount of improvement or performance.

The studies investigating the arrangement of feedback contingencies on undergraduate exam and quiz performance have found more immediate and detailed feedback led to improved performances on assessments (Brosvic & Epstein, 2007; Dorow & Boyle, 1998). The one study whose results did not support that conclusion had a potential confound that made the results difficult to interpret (Buzhardt & Semb, 2002).

Analysis of Components

The review suggested procedures of several kinds, including increasing active participation in class, scheduling outside of class review activities, providing study materials, and arranging feedback effectively sometimes led to increased student performance on exams and quizzes in college courses. No one type of procedure, however, produced consistently positive and large improvements in student performance. Instead, the amount of improvement varied across and within procedures. Such variability suggested that characteristics of individual procedures, rather than the general type of procedure used to support student learning, were responsible for improvements in student performance. In order to identify the components of procedures associated with

larger improvements in student learning, I also conducted an analysis of all of the studies for which the percentage of improvement relative to a control condition could be determined.

Twenty of the forty reviewed studies provided enough information to determine both the percentage of improvement in student performance and the components of the procedures used. The components identified for analysis were: (a) the degree of similarity between the review activity and the assessment, (b) whether or not specific performance criteria or a question pool were provided to students, (c) the type of response required of the students during the review activity, and (d) whether or not sample or correct answers to the review activity were shown and feedback given to students about their answers to review questions. If it was not clear from the description provided in the article how a procedure should be scored along a dimension of a particular component, then that procedure was not included in the analysis of that component. Some studies contained multiple procedures, and each procedure was categorized separately for the analysis. For example, in Aamodt's second study (1982b), two types of study sessions, each with different components, were evaluated. Thus, both types of study sessions were scored separately, according to the criteria below, as were their percentages of improvement. In other studies, replications were reported. Each replication was scored separately. A summary of the twenty studies included in the analysis and the components contained in each is presented in Table 1.

Similarity of content and format. Each procedure was scored for similarity of content and form to the actual assessment. For a review activity to be classified as having similar content or format to the actual assessment there had to be a more than 50%

overlap between the two along that dimension. For example, to be scored as having a similar content, more than 50% of the content of the review activity must have been on the actual assessment and more than 50% of the assessment content must have been included in the review activity. Thus, if 100% of the content of the review activity appeared on the exam, but that content was only 10% of the material covered on the exam, the activity would not be scored as having a similar content to the actual assessment. The same requirement was used for the format dimension.

Provision of specific competencies or question pool. An activity was scored according to whether or not students were provided specific descriptions of the competencies they were expected to possess for the specific exam or whether they were given access to a question pool containing potential assessment questions. To count as providing specific competencies, the procedure had to give students a clear idea of a specific performance requirement (e.g., "Be prepared to define and provide an example of extinction") as opposed to general tips (e.g., "Exams will be multiple-choice and include material from lecture"). To count as having provided a question pool, the students had to receive a set of study or review questions from which more than 50% of the exam or quiz questions were selected. For example, if questions from daily quizzes were used again as 50% of the final exam questions, then that would count as the students receiving a question pool. If the same daily quiz questions were used on the final exam, but were only 10% of the total exam questions, then it would not count as providing the students a question pool.

Type of response required. Each procedure was scored as "requiring a written response," "requiring some other kind of response," or "not requiring any kind of

response" from the student. For a procedure to be scored as requiring a written response, students had to be required to write something as part of the activity. The requirement for writing is the defining characteristic. An activity such as providing guided notes (on which the students could write) would not count as "requiring a written response" unless the instructor placed some contingency on writing such as requesting that students do so or giving credit for writing notes. For an activity to be scored as "requiring some other kind of response," students had to be required to engage in some behavior as part of the activity (e.g. raising a response card, answering questions verbally, etc.). Again, only those activities that required responses, rather than simply provide the opportunity to respond, were included in this category. A score of "no response" required ranking included all activities that did not have an overt response requirement of any kind. This included such procedures as providing a study guide, handing out lecture notes, and holding question and answer sessions in which individual students did not have to make a response.

Provision of sample/correct answers and/or feedback. An activity was scored according to whether or not students viewed either a sample answer or a correct answer to review activity questions and also whether or not they received feedback about their performance on review activities. For example, some studies reported procedures where the instructor asked students to complete a practice exam. Students were then shown the grading key for the questions and the correct answer described. Such a procedure would count as providing both a correct answer and performance feedback. Feedback can take many forms, including specific, individualized, learner feedback and showing a group who each answered a question the answer key without the instructor individually

commenting on a single student's paper. The majority of studies reviewed here did not involve the instructor giving individual students individualized feedback. Other procedures involved simply showing students correct answers to study questions in the absence of any requirement that students answer the questions. In that case, the procedure was classified as showing a correct answer but not providing feedback to the students about their specific answer.

Tables 2-5 summarize the results of the component analysis. Each table represents one component and presents the average improvement for all studies in each component category.

Similarity of content and format. Table 2 summarizes the analysis of the effects of different degrees of similarity on improvements in student performance. The highest average amount of improvement was when procedures were similar to the actual exam or quiz in both form and content. The next highest average was when there was similarity of content but not form. The lowest averages were when the procedures and exams were dissimilar in content, even if they were similar in form.

Provision of specific competencies or question pool. Tables 3, 3a, and 3b summarize the analysis of the effects of providing specific competencies or a question pool on improvements in student performance. Procedures that provided specific competencies or question pools produced a much higher average improvement than procedures that did not. Providing a question pool led to larger average improvements than providing specific competencies. Procedures that provided a question pool for multiple-choice exams produced the highest average improvement.

Type of response required. Table 4 summarizes the analysis of the effect of the type of response required of students on improvements in student performance. Procedures requiring a written response produced the highest average improvement.

Provision of sample/correct answers and/or feedback. Table 5 summarizes the analysis of providing sample/correct answers or performance feedback on improvements in student performance. Procedures that provided both sample or correct answers and feedback to students about their individual answers produced the largest average improvement in student performance.

Discussion of component analysis. Of the four components analyzed, the similarity of the review procedure to the actual quiz or exam appeared to be the most effective. Of the ten largest improvements seen in student performance, nine of them (Jenkins & Neisworth, 1973; Miles, et al., 1967; Miller & Malott, 1997; Rust, et al., 2003; Semb, et al., 1973) involved a procedure that was similar in both content and form to the actual assessment. The tenth (Malanga & Sweeney, 2008) involved procedures similar in content. Six (Aamodt, 1982a; Dickson & Miller, 2005; Fleming, 2002; Neef, et al., 2007) of the ten least effective procedures were not similar in content or format to the actual assessment, and eight (Aamodt, 1982a; Dickson & Miller, 2005; Fleming, 2002; Morling, et al., 2008; Neef, et al., 2007; Shabani & Carr, 2004) of the ten least effective procedures were not similar in content.

Providing specific criteria and/or a question pool appeared to be the next most effective component. Eight (Jenkins & Neisworth, 1973; Malanga & Sweeney, 2008; Miles, et al., 1967; Rust, et al., 2003; Semb, et al., 1973) of the ten largest improvements in student performance were associated with procedures that provided criteria or a

question pool, and nine (Aamodt, 1982a; Fleming, 2002; Malanga & Sweeney, 2008; Neef, et al., 2007; Poirier & Feldman, 2007; Shabani & Carr, 2004) of the ten least effective procedures did not provide either of those things.

Determining the effectiveness of requiring different types of responses and providing sample/correct answers and/or feedback to students is more difficult. This is because most of the procedures that showed an improvement in student performance also provided students with specific competencies or a question pool and were similar to the actual exam in content. Thus, the presence of the additional components (both associated with improvements in performance) obscures the effects of the other two components. When the procedures containing other components are removed from the calculations for the type of response required and the provision of answers and feedback, however, the remaining procedures show smaller improvements in student performance.

The number of studies allowing a comparative analysis was small, however, and the conclusions drawn are necessarily tentative. Because the components of procedures, rather than their general form, appeared most related to the effectiveness of the procedures, it is important that research begin to explore the effects of individual components of educational supports in an effort to identify which ones are most associated with improvements in student learning and the circumstances under which they are most effective.

Discussion of reviewed literature

In addition to the need for more careful evaluation of individual components of procedures used to support student learning, there were three procedural limitations in much of the reviewed literature that should also be addressed by future research. First,

only three of the forty studies reviewed explored the effect of interventions on student completion of open-ended essay assignments (Huxham, 2005; Price, et al., 2007; Rust, et al., 2003). Essay exams are typically more demanding than multiple-choice exams and often require students to engage in more complex responding (e.g. constructing a coherent, written argument versus circling a correct answer). Only five others (Bol & Hacker, 2001; Dorow & Boyle, 1998; Malanga & Sweeney, 2008; Padilla-Walker, 2006; Simon, 2005) involved short-answer or fill-in-the-blank questions. Interventions that showed small effects on multiple-choice exam performance may possibly produce larger effects on more open-ended assessments or vice-versa.

Second, only seventeen of the forty reviewed studies used within-subject comparisons between conditions. Of those, only six (Jenkins & Neisworth, 1973; Landrum, 2007; Miller & Malott, 2006; Nevid & Mahon, 2009; Semb, et al., 1973; Simon, 2005) used a within-subject comparison of conditions on the same assessment. None of the studies used a within-subject comparison of two treatment conditions delivered on the same day and also assessed at the same time. A within-subject comparison of conditions in which both conditions occur within the same session and learning differences are evaluated on the same assessment has several advantages, including eliminating the need to control for the effects of time, differences in difficulty or complexity of material across exams, participant characteristics such as mental and physical state, and other instructional variables such as instructor and class meeting time and setting.

Third, and most serious, was the lack of formal reporting of independent variable reliability. While six of the forty studies (Austin, Lee, Thibeault, Carr, & Bailey, 2002;

Baker & Lombardi, 1985; Fleming, 2002; Jenkins & Neisworth, 1973; Mayfield & Chase, 2002; Neef, et al., 2007) implied treatment integrity was observed either directly (Baker & Lombardi, 1985; Fleming, 2002; Neef, et al., 2007) or indirectly (Austin, et al., 2002; Jenkins & Neisworth, 1973; Mayfield & Chase, 2002), there were no reported measures of the correct implementation of the independent variables in any study. Only two studies (Buzhardt & Semb, 2002; Simon, 2005) reported a formal independent variable reliability score of any kind, and in both cases it was for only a single variable, and these were not the only variables being manipulated in either study. This is a concern because the behavior of the teacher and the correct administration of procedures across conditions could have serious implications for the likely effectiveness of teaching procedures.

The issue of independent variable reliability has been raised in the behavior analytic field at large several times (Kazdin, 1977; Peterson, Homer, & Wonderlich, 1982). Potential threats when independent variable integrity is not assessed or reported include: not identifying confounding variables that may be responsible for treatment effects, not identifying inefficient or unnecessary aspects of the treatment package (e.g. inability to identify components and conduct potential component analysis or continue offering an inefficient treatment when a more efficient version may be possible), and making it unlikely the results can be replicated because not all aspects of independent variable implementation are documented or specified (Peterson, et al., 1982). As related to the current review, if future instructors and researchers hope to better integrate the research findings about improving undergraduate learning, it will be important to document better not only exactly what the delivery of intervention components involves

but also how rigorously they must be applied in order to achieve desirable results. The mixed findings about the effectiveness of several interventions discussed above such as response cards could be the result of unidentified differences in the method and rigor of applying the technology within the different classrooms reported in the studies. Knowledge of how the treatment was applied and the consistency with which it was done would allow a more careful analysis of what led to the diverse findings.

In addition to concerns about independent variable reliability, only nine of the forty studies reported dependent variable reliability measures of any kind. Because the majority of the assessments used were multiple-choice in format, however, the likelihood of discrepancy in scoring was low. In the evaluation of essay and short-answer exams, where the scoring of answers is more subjective and based more on interpretation, the risk of discrepancy across graders is higher. As additional studies explore the effects of procedures on exams involving more complex answers, it is important that the reliability of scorers be determined and reported along with independent variable reliability scores.

The purpose of the present investigation was to conduct a component analysis of the variables present in a mock exam study session in order to make the sessions both more effective and efficient. The mock exam study session (described in more detail below) had been used for six semesters prior to the current study, and evaluations had demonstrated that it was effective at improving student performance on application-based, short-essay exams (Dotson, Sheldon, & Sherman, in press). While the sessions were effective, each one lasted two-and-a-half hours and required a large amount of preparation by the course graduate teaching assistant (GTA) who facilitated the sessions.

By investigating the role of individual components of the mock exam on its effectiveness, we hoped to be able to make the sessions both more effective and efficient.

We also designed the studies in the present investigation to address the limitations in the reviewed literature. First, the investigation began the process of evaluating individual components of educational support procedures to determine their effects on student exam performance. Second, the evaluation of specific components was within a course requiring short-essay and essay-based responding on exams. Third, the present investigation presented a further refinement of the within-subject comparison approach to evaluating the effectiveness of interventions to improve exam performance. Fourth, included in the investigation was a potential methodology for evaluating the treatment integrity of teaching interventions to better account for the actual behavior of the teacher and ensure teaching activities are conducted as reported.

This research project attempts to determine what components of the mock exam study session (e.g. requiring students to write answers, providing students with grading keys for questions) were responsible for improvements in student exam performance across three studies. Both within-subject and across-groups comparisons were made between experimental conditions. Also, multiple linear regression analyses were conducted to determine to what degree additional variables (e.g. attendance at class lectures and entering GPA) were also related to student performance in the course.

General Methods

All students (N = 64) enrolled in an undergraduate introduction to behavior modification and therapy course during the spring semester of 2009 participated in the study. Students attended a large state university in the Midwest.

Structure of the Course

The course contained five units. At the beginning of each, an outline of the content of the unit was made available to students online. At the end of each unit, students took an exam over the material covered in the preceding unit. Lectures accounted for 40 total class meetings and exams accounted for an additional 5 class meetings.

Unit exams were worth thirty points each (150 total points in the course) and consisted of essay and short-answer questions. The majority of questions (80-90%) required students to apply behavioral principles and techniques described in the textbook (Martin & Pear, 2007) and discussed in lectures to address novel applied problems that they had not seen before. For example, a question on the principles of reinforcement and shaping might describe a man with disabilities, and students might be asked to explain how they would teach him to brush his teeth. They would be expected to describe how they would use reinforcement, shaping techniques, prompting, and prompt fading to teach the skill. Short-answer questions taken verbatim from the textbook and lectures accounted for 10-20% of the questions on each exam.

Students had opportunities to participate in two forms of optional practice during each unit. One involved writing and submitting answers to online questions from a practice exam. The other was attending a mock exam study session.

Online practice exam questions. For each unit of the course, a practice exam was posted online on the first day of the unit. Questions from the practice exam were assigned throughout each unit (usually several questions after each class period), and students had the option to complete questions for feedback and extra credit. Each

practice exam contained a description of one or more clinical situations and a set of questions. Each question required students to use the information presented in the unit to develop behavioral solutions to the clinical situations described.

The questions asked on the practice exams and the actual exams were similar in structure and format. Both the practice exam and actual exam questions required students to apply the course material to a novel situation. The difference between the practice exams and the actual exams was the situations to which the unit materials were applied. For example, one question on both exams in the third unit was, "Describe how to use extinction to reduce the client's problem behavior." The client described in the practice exam might be a child who tantrumed whenever his parents did not pay attention to him, and a functional assessment might be described that indicated the tantrums were maintained by attention. On the actual exam, the client might be a junior-high student who became aggressive when asked to work, and the functional analysis indicated the aggression was maintained by escape from academic demands. Students had to apply what they had learned about extinction to the different situations. For the first case, a correct answer involved not paying any attention to the child when he tantrumed. For the second, a correct answer involved continuing to present requests to work even if the student became aggressive.

Students had the opportunity to answer questions from the practice exam 4-6 times during each unit (several questions were usually assigned after each class period), and students turned in written answers online on specified dates prior to the unit exam. Students who submitted answers received feedback (a copy of the grading key for those questions with the score for their answers to those questions) online from the course GTA

within a week of submitting answers and earned up to three extra-credit points for each unit of the course for answering the questions. Extra credit points counted toward the final course grade, and the amount of extra credit earned depended on the correctness of the answers submitted. Thus, a student who turned in all of the assignments and was 50% correct would have received 1.5 extra credit points for each unit for a total of 7.5 points for the course. A student who turned in all of the assignments and was 100% correct would have earned 3 extra credit points for each unit for a total of 15 points for the course. Because course grades were assigned based on the percentage of 150 points the students earned, the students could earn up to 15 *additional* points towards their final grade by completing the practice questions accurately and on time.

Mock exam study sessions. During each unit, students could also participate in a mock exam study session led by the course graduate teaching assistant (GTA). The mock exam sessions were held one and two days before the unit exam. For each unit of the course, there were three mock exam sessions. The first occurred in the evening two days before the actual exam. The second occurred during the afternoon the day before the actual exam. The third mock exam session occurred the evening before the actual exam. The sessions took place in a classroom on campus. Students earned no extra credit for participating in mock exam sessions. The format of the mock exam study sessions changed across the three studies as different component variables were manipulated. Overall, the course GTA conducted three types of mock exam study session. The first type of mock exam session involved students writing answers to questions on a mock exam and explored the effects of doing so relative to not writing answers. The second type of session evaluated different types of discussion by comparing the effect of

requiring students to evaluate and correct sample answers to the effect of asking students to listen to the GTA do the same. The third type of session compared requiring students to evaluate and correct sample answers to mock exam questions to only providing them with a grading key to other questions. All three types of session contained a brief introduction and discussion facilitated by the GTA. The format of each type of mock exam study session is discussed in more detail below. Table 6 summarizes the experimental question each type of mock exam was designed to answer.

Independent Variables

The independent variables in the studies below were components of the mock exam session that were manipulated across conditions and studies (e.g., asking students to write answers to mock exam questions or providing students with copies of the grading key to the same questions). The specific components manipulated differed across the studies. The independent variables for each study and analysis are described in more detail below.

Dependent Variables

The primary dependent variable across all three studies was student performance on unit exams. In addition to performance on unit exams, students' final course grades were also recorded.

Reliability

Dependent variable reliability was calculated for student exam performance. Two additional graders (the course professors) independently scored ten percent of the exams in the course. Reliability was calculated by comparing agreement on points earned and not earned on a question-by-question basis. Questions were worth from 1 to 10 points

each. For each question, the total points earned and not earned for which scorers agreed was divided by the total number of points on the question and multiplied by 100. For example, if one grader gave the student 4 of 5 possible points on a question and the second grader gave the student 3 of 5 possible points, then the total number of points on which the graders agreed (3 points earned and one point not earned) would be divided by the total points possible (5) and multiplied by 100 to determine a percentage of agreement (4/5 x 100 = 80%). Rates of reliability were compared between questions in different experimental conditions to determine if there were any grading biases across conditions.

Several measures of independent variable reliability (treatment integrity) were calculated. First, treatment integrity across types of mock exam sessions regarding the presence or absence of the characteristics (e.g. asking students to write answers, providing a brief introduction) of each type of session was calculated. The total number of characteristics present was divided by the total number that should have been present and multiplied by 100 to determine the percentage of characteristics present in each mock exam session. Second, treatment integrity within experimental conditions was measured to determine to what degree the GTA followed the described procedures for each question within each mock exam session (e.g. asking students to offer their evaluation of a sample answer, asking students to correct the sample answer) The total number of characteristics present for each question were divided by the total number that should have been present and multiplied by 100 to determine the percentage of characteristics present for each question. Additionally, the degree of correct implementation of each experimental condition was determined by calculating the average percentage of characteristics present across all of the questions. Tables 7-10 provide examples of the

reliability scoring sheets used to calculate the degree to which the GTA's behavior during the mock exam sessions followed experimental protocol. Finally, the percentage of students completing the grading of sample answers during the third type of mock exam session was calculated by comparing the percentage of sample answers for which students wrote a grade on the grading key to the total number of answers for which they were asked to do so.

Treatment integrity data was calculated by scoring video-tapes of each mock exam study session and by looking at student-completed grading keys for the last three sets of mock exam study sessions. A primary observer scored videos for 33% of the total mock exam sessions held during the semester. A second observer scored 60% of the videos scored by the primary observer. The first observer scored videos of at least one mock exam session for each of the five units of the course, and the second observer scored videos from at least one of each of the three types of mock exam session.

Reliability of the attendance counts at both class lectures and mock exam sessions was determined by having students sign an attendance sheet. The GTA counted the number of students in the room during those sessions and compared the counted number against the number of people who signed the attendance sheets.

Experimental design

The primary comparison across studies was a within-subjects evaluation of exam performance across different questions on the same unit exams for students who attended the mock exam study sessions. Comparisons were made between performances on questions in each of the two experimental conditions to determine if student performance was different across the two conditions. The questions in each experimental condition

were counterbalanced across mock exam sessions within each unit. If a student attended more than one mock exam session, the student's exam scores were not included in the analysis.

Additional analyses were conducted involving across-group comparisons of performance on unit exams. Comparisons looked at exam performance across groups of students who did and did not attend the mock exam study sessions and who did and did not complete practice question assignments. Also, analysis of performance on individual exam questions compared the performance levels of various groups of students. For example, during the third type of mock exam, the performance levels of three groups of students was compared: those who attended and discussed the question in the treatment condition of the mock exam study session, those who attended the mock exam study session but did not discuss the question in the treatment condition, and those not attending the mock exam study session.

The primary method of analyzing student performance was a visual analysis of the individual difference scores calculated for students attending the mock exam sessions, student performances on exams, and the exam score distributions presented on the correlation analysis graphs. To determine if any observed trends were significant, additional statistical analyses were used to evaluate the results.

Study 1

The first study explored the effect of students writing, evaluating and correcting answers on the mock exam on unit exam performance as versus students simply evaluating and correcting provided sample answers. This study was conducted during the first unit of the course.

Procedures. The first type of mock exam session explored the effect on exam scores of requiring students to write answers to mock exam questions versus not requiring them to write answers. The mock exam sessions lasted between 105-120 minutes and contained 3 parts: a brief introduction, writing of answers, and discussion. Table 11 summarizes the structure of this type of mock exam study session. During the brief introduction, that lasted roughly 15 minutes, the GTA discussed the structure of the mock exam session, the general format and content of the actual exam, and provided general study tips to help students prepare for the exam. Table 12 provides an example of the general study tips provided to students. Next, the students were given 20 minutes to write answers to a mock exam. The mock exam contained a sample of the same kinds of questions found on the practice exam and actual exam, but the situations described in the mock exam were different than the situations on either the practice exam or the actual exam. The situation and questions presented on the mock exam were the same across sessions, but across sessions students were asked to write answers to different subsets of questions. The questions to which students were asked to write answers were counterbalanced across sessions. Table 13 summarizes the differences in conditions in this version of the mock exam, and Table 14 summarizes the content covered by the first unit exam and the breakdown of which content was found on both versions of the mock exam for Unit 1. Students who attended the first mock exam study session during Unit 1 wrote answers to Version 1 of the mock exam. Students who attended the second or third mock exam sessions during Unit 1 wrote answers to Version 2 of the mock exam.

After the writing period, the GTA handed out a grading key containing the grading criteria for every question on both versions of the mock exam and spent 70-80

minutes leading the students in a discussion of the grading criteria and answers for all of the questions on both versions of the mock exam. During this part of the mock exam session, the GTA described the grading criteria and guided students through evaluating and correcting answers to every question on the mock exam. Depending on which question was being discussed, students discussed and corrected sample answers either volunteered by their peers or offered by the GTA. For every question, the GTA first briefly reviewed and explained the grading criteria found in the grading key. If the question was one for which students had written answers, the GTA asked two students from the group to volunteer their answers for the group to hear and asked the other students to evaluate the provided answers according to the grading criteria. If the question was one to which students had not written answers, the GTA showed two sample answers on a projection screen for the group to read and asked the group to evaluate the provided answers according to the grading criteria. Students evaluated answers as a group, and the GTA provided prompts and reinforcement to encourage identification of important details of answers bring discussed. Once the students had evaluated either type of answer, the GTA asked them to offer any corrections necessary to ensure the answer would earn full credit.

By the end of the mock exam study session, the GTA had described the grading criteria for every question on the grading key. Students had also viewed, evaluated, and corrected two examples of answers to each of those questions with guidance from the GTA. Students were allowed to keep both their answers to the mock exam and the grading key. To evaluate the effects of requiring students to write answers to mock exam

questions, both within and across-groups comparisons were made as described in the general methods section.

Results. Figure 1 displays the difference in accuracy between conditions for each student who attended a mock exam session. Each open circle/bar represents one student. If the circle is above the x-axis on the graph, the student performed better on the questions for which they wrote answers. If the circle is below the x-axis, the student performed worse. The distance from the x-axis indicates how much better or worse their performance was in the condition. On average, there was no difference in performance between the two conditions.

An additional comparison was made between groups of students who did and did not attend the mock exam session. Figure 2 represents the average overall performance on exam questions for each of three groups of students. A one-way ANOVA analysis indicated a significant difference between means across the three groups F (2, 125) = 28.37, p = <0.0001, and post-hoc Tukey's tests showed significant differences between those students attending the mock exam sessions and those not attending the mock exam sessions at the p = <.0001 level. There was no significant difference between students writing and not writing answers to mock exam questions.

It may have been that students had already gained the available benefit of writing answers by completing answers to the practice exam questions online before attending the mock exam session. In order to evaluate this possibility, the amount of extra credit earned by each student who attended the mock exam session was determined, and that percentage was compared to their difference score between experimental conditions. If writing correct answers to online practice questions (thus earning extra credit) led to a

reduced effect for writing answers to mock exam questions, then it would be expected that as students earned more extra credit, they would show smaller differences between the writing and not writing conditions during the mock exam. The results suggest that students writing answers to online practice questions did not lead to a reduced effect of requiring students to write answers to mock exam questions. A correlation analysis of the amount of extra credit students earned and their difference scores between the two conditions did not find a significant relationship between the two variables. Figure 3 presents the data used to calculate the correlational analysis. Each open circle represents a single student's difference score between the experimental conditions, and the circle's location on the x-axis represents the amount of extra credit that student earned for the first unit of the course.

Discussion. The purpose of the first experimental condition was to determine the effect of requiring a written response on exam performance. The analysis of the previously published literature showed generally that procedures requiring a written response of some kind led to larger improvements in exam performance than procedures requiring other types of response or no response. We expected that requiring a written response would be effective because writing answers to short-essay questions is more similar to what is done on the unit exam than talking about answers or listening to discussion about answers. Nevertheless, in the present study, no difference was found between questions on which students wrote answers and questions on which students did not write answers. There are several reasons that requiring written answers may not have had an effect.

It may be that other variables associated with improved performance present in both conditions had a larger effect on performance than writing answers would have had alone. For example, in both conditions students saw answers and grading criteria in the same form and content as the actual exam and also heard a discussion of exactly what the performance requirements were for each question. Both similarity in form and content and a clear specification of performance requirements were associated with improvements in exam performance in the reviewed literature, and it may be that the effect of those variables, present in both conditions, improved performance as much as it could be improved in this type of study session. In other words, there may have been a ceiling for the size of the effect the mock exam session could have, and the effects of the similarity and clarity of expectations across the two experimental conditions improved performance to that level, leaving no additional room for the effects of written rehearsal to be seen.

It could also be the case that students were engaging in additional overt and covert practice behaviors for the questions for which they did not write answers. While the literature review suggested requiring written answers produced improvements, the sample of studies was small and the format and content of the exams sufficiently different to make such a conclusion tentative. During the discussion about questions for which students did not write answers, they still had the experience of seeing, evaluating, and correcting sample answers. Practice associated with evaluating and offering a correct version of each sample answer may have been sufficiently equivalent to the writing of answers to produce the same effect on exam performance. Students could also have been

silently thinking about or practicing answering each question as it was discussed as well, and such practice could have also had an effect on their exam performance.

An additional reason that requiring the students to write answers to mock exam questions may not have had an effect involves the timing of when students were asked to write their answers. It may be that because students wrote answers before any discussion, evaluation, or correction of sample answers that the students were not as likely to engage in correct practice. They may have written incorrect or incomplete answers, which would not be expected to produce better unit exam performance (in the same way practicing playing a song on the piano by playing incorrect notes does not improve the successful playing of that song during a concert). Perhaps if the students had been asked to write their answers after the discussion portion of the mock exam session, the practice would have produced positive effects on exam performance.

Study 2

The second study explored the effects on unit exam performance of different types of discussion during the mock exam. The study compared requiring students to evaluate and correct sample answers to asking students to listen to the GTA do the same.

Procedures. The mock exam session in Study Two occurred during the second unit of the course and examined the effect on unit exam performance of requiring students to evaluate and correct sample answers versus listening to and watching the GTA do so. The primary difference between mock exam sessions in the first and second study was that students practiced evaluating sample answers during the second study but were not asked to write answers. The mock exam sessions in study two lasted 105

minutes and consisted of 2 parts. Table 15 summarizes the structure of the mock exam sessions for the second study.

During the brief introduction, which lasted roughly 15 minutes, the GTA discussed the structure of the mock exam session, the general format and content of the actual exam, and provided four general study tips to help students prepare for the actual exam. Following the brief introduction, the GTA handed out a mock exam that was already completed with sample answers. The mock exam contained the same sorts of questions found on the practice exam and actual exam for Unit 2, but the situations described on the mock exam were different than those on the practice or actual exams. The GTA also handed out a grading key to the mock exam. The GTA then led a 90-min discussion of the grading criteria and the sample answers found on the mock exam. The GTA discussed each question on the mock exam. For each question, the GTA described the grading criteria. For some questions, the GTA would then ask the students as a group to evaluate and correct the sample answers for that question. For other questions, the GTA evaluated and corrected the sample answers. Table 16 summarizes the difference in conditions during this mock exam. The questions that students evaluated and corrected and the questions that the GTA evaluated and corrected were counterbalanced across the different mock exam sessions. Table 17 summarizes the content covered during the second unit of the course and how questions about that material were discussed during the mock exam study sessions.

On questions that students evaluated and corrected a sample answer, the GTA asked the students as a group to orally evaluate the provided sample answer according to the grading criteria. Once the students had evaluated the answer, the GTA asked them to

orally offer any corrections necessary to ensure the answer would earn full credit. Once the students offered a correct answer, the GTA moved on to the next question. On questions that the GTA evaluated and corrected the sample answer, he described to students how he would have scored the sample answer according to the grading criteria and explained how the answer could be corrected so it would earn full credit.

By the end of the mock exam session, the GTA had described the grading criteria for every question on the mock exam. Students had also heard sample answers to every question evaluated and corrected, either by other students or the GTA. Students were allowed to take both the sample answers to the mock exam and the grading key home with them for further study at the end of the session. The analysis of student unit exam performance was the same as the first study.

Results. Figure 4 presents the difference in accuracy between experimental conditions for each student who attended a mock exam session during study two. If the circle is above the x-axis on the graph, the student performed better on the questions for which they discussed answers. If the circle is below the x-axis, the student performed worse. The distance from the x-axis indicates how much better or worse their performance was in the condition. On average, there was no difference in performance between the two conditions.

An additional comparison was made between groups of students who did and did not attend the mock exam session. Figure 5 represents the average overall performance on exam questions for each of three groups of students. A one-way ANOVA analysis showed a significant difference in performance between groups [F (2, 445) = 56.70, p = <0.0001], and post-hoc Tukey's tests showed significant differences between those

students attending the mock exam sessions and those not attending the mock exam sessions at the $p = <.0001$ level. There was no significant difference between students evaluating answers to mock exam questions and students listening to the GTA do the same.

Discussion. The purpose of the second type of mock exam session was to explore the effect of requiring students to evaluate and correct sample answers versus watching the course GTA evaluate and correct sample answers –thus evaluating active responding versus passive observation. We hypothesized that requiring students to engage in discussion would produce higher levels of performance than if students were allowed to sit passively and watch the course GTA present the discussion. Overall, however, there was no difference in performance between the two conditions.

As in the failure to find an effect for requiring students to write answers in the first study, several of the same reasons may explain the lack of difference between the two conditions in the second study. Because both experimental conditions in Study Two involved providing students with both clear performance expectations and realistic examples of exam questions and answers, the presence of those two variables may have produced as much improvement as was possible in the mock exam sessions. The potential ceiling effect produced by introducing multiple potentially effective variables into both conditions and only changing one of them possibly obscured any effects a single variable had on exam performance.

Also, students may have engaged in "active" practice even during the review of questions for which only the GTA discussed the sample answers. Anecdotally, the GTA observed that even for questions that the students did not have to discuss, they

nevertheless often interrupted his discussion to offer potential corrections and evaluations of the sample answers on the screen. Also, the students became annoyed when their questions and offered answers were not discussed by the GTA and, in several situations, became insistent that they be allowed to discuss their own answers. Even though the GTA followed the procedural requirements to maintain the integrity of the two experimental conditions, the students did not appear to discriminate the differences as readily and continued to try to discuss all of the questions in the session. Because the behavior of the students could not be easily or accurately scored given the location of the video-camera during the session, it was not possible to provide a quantitative measure of how often they engaged in behaviors intended for the active condition (e.g., offering corrections, describing what was wrong with a question, and asking for feedback on an offered correction) during the GTA's discussion of questions, but the frequency of such behaviors was consistent during the sessions. Given the overlap of student behaviors across conditions, it is not surprising that students did not perform significantly differently in the two experimental conditions.

Study 3

The third study was designed to address one of the procedural limitations found in Study Two by carefully distinguishing and separating the questions in the two experimental conditions. Additionally, a package of variables (as opposed to a single variable) was evaluated in order to maximize the likelihood of producing a large effect on student unit exam performance. The mock exams in study three compared the effects of requiring students to evaluate and correct sample answers to mock exam questions to only providing them with a grading key to other questions.

Procedures. The third type of mock exam study session occurred during the third, fourth, and fifth units of the course, and explored the effects on unit exam performance of requiring students to discuss, evaluate, and correct sample answers versus only providing them with study materials (i.e., mock exam questions and a grading key). Like the first two types of mock exam sessions, the third type of mock exam session contained a brief introduction and a discussion of answers, but during these sessions, students also filled out a grading key for sample answers to a mock exam. Each session lasted approximately 105 minutes. Table 18 summarizes the structure of the sessions.

During the brief introduction, which lasted roughly 15 minutes, the GTA discussed the structure of the mock exam session, the general format and content of the actual exam, and provided four general study tips to help students prepare for the actual exam. Table 19 summarizes the manipulation of the variables during the next phase of the mock exam session. Following the introduction, the GTA handed out a mock exam with sample answers filled in for some questions. The GTA also gave students two copies of the grading key for the mock exam and asked students to grade the sample answers in the mock exam. Students entered a score for each sample answer on both grading keys. They were allowed to keep one copy to take home and handed in the other copy to the GTA at the end of the 20 min grading period. The mock exam contained the same type of questions found on the practice exam and actual exam for each unit, but the situations described on the mock exam were different than those on the practice or actual exams. Only some of the questions on the mock exam had sample answers filled in. Students attending different mock exam study sessions received the same mock exam with different questions answered. Tables 20, 21, and 22 summarize the material covered

in the last three units of the course and which questions about that material were answered on the different versions of the mock exams. After collecting the grading keys from students, the GTA then spent approximately 70 minutes leading the students in a discussion of the questions on the mock exam for which they had graded sample answers. Only questions on which students graded answers were discussed during the discussion portion of the mock exam study session. Questions on the mock exam for which sample answers were not graded were not discussed at any point during the session.

For each graded question, the GTA described the grading criteria for that question. Then the GTA asked the students as a group to evaluate the sample answers according to the grading criteria. Once the students had evaluated the answer, the GTA asked them to offer any corrections necessary to ensure the answer would earn full credit.

By the end of the mock exam session, the GTA had described the grading criteria for each question for which students had graded sample answers. Students had evaluated and corrected the same sample answers, and students were allowed to keep both the sample answers to the mock exam and the grading key. Comparisons of student performance were the same as that conducted in the first two studies.

Results. Figure 6 presents the difference in accuracy between experimental conditions for each student who attended a mock exam session for exams 3-5. Figures 8, 10, and 12 present the same information for the three exams individually. The graphs are the same as those presented for the first two types of mock exam session. Students performed better on questions on which they evaluated and corrected sample answers than on questions for which they were only given the grading key. A one-tailed t-test of the cumulative results from all three exams indicated a significant difference in

performance, t (134) = 6.518, p<.0001 and a large effect, Cohen's d = 0.80. Equivalent results were found when analyses were conducted for each exam individually.

An additional comparison was made between groups of students who did and did not attend the mock exam session. Figure 7 represents the average overall performance on exam questions for each of three groups of students across exams 3-5. Figures 9, 11, and 13 present the same information for the three exams individually. Students evaluating and correcting sample answers on a questions scored the highest, followed by students who received the grading key. Students who did not attend the mock exam session (and thus got nothing) performed the least well on exam questions. A one-way ANOVA analysis of the cumulative results from all exams showed significant differences between groups [F (2, 627) = 51.52, p = <0.0001], and post-hoc Tukey's tests indicated significant differences between all three groups at the p = <0.0001 level. Similar results were found when the same analyses were conducted for each unit exam individually.

Discussion. The third study investigated the effects of a package of some of the active variables (i.e., discussion of performance expectations, written and verbal evaluation of sample answers, and correction of sample answers) used in mock exam sessions versus only providing the study materials associated with mock exam sessions (i.e. being given a blank mock exam and the grading key for each question). The sessions were designed to eliminate the potential for students to engage in similar practice behaviors across conditions seen in the first and second types of mock exam sessions (e.g. students engaging in "practice" verbally or covertly) by only talking about one set of experimental questions during the mock exam sessions. Thus, students were not given time during the mock exam session to look at, discuss, ask questions about, or

offer answers to any of the questions in the non-discussion condition. Students performed significantly better on questions for which discussion, evaluation, and correction of sample answers occurred than on questions for which they only received the question and the grading criteria for that question in the absence of discussion. These results were replicated on two additional exams, with statistically significant differences seen each time.

The results of study three indicated that active discussion of exam expectations and sample answers to potential exam questions produced larger improvements in exam performance than simply providing study materials. What they did not do is identify a single variable most responsible for the improvements seen in student exam performance. Conclusions can only be made regarding the effectiveness of the package of treatment components on exam performance, and future research is needed if we are to further isolate the effects of the individual components.

The results of study three were not completely consistent with results in the previously reviewed literature. In the prior studies in which the largest effects were found, the interventions required no responses from students, but provided question pools and clear statements of performance expectations about content in a similar form to the actual exam (Jenkins & Neisworth, 1973; Miles, et al., 1967; Semb, et al., 1973). The study materials provided in the non-active condition in the third study appear on the surface to be identical to those materials in the studies reporting the largest effects (i.e. provided the questions that would appear on the exam and also a clear statement of expectations in the form of the grading key for each question). While providing the grading key did produce improvements in student performance, they were not of the size

of those reported in the earlier studies. A key difference, however, was the type of exam used to evaluate student learning. In the studies cited above, the exams used multiple-choice questions and the students could perform well on those exams by simply memorizing the provided study materials and the correct answers to the provided questions. In the present course, however, the exams required students to apply their knowledge to novel situations by writing short-essay answers to exam questions. Thus, simply writing answers consisting of memorized study materials would not earn credit for the student. It is possible that improving performance on short-essay exams requires supports of greater complexity or intensity than those used for multiple-choice exams.

Analysis of additional variables

All three studies in the current investigation reported a significant effect of attending mock exam sessions on unit exam performance. Performance on the five unit exams determined overall performance in the course, so it is important to explore the degree to which attendance at mock exam sessions was related to both exam performance and overall course performance when controlling for the presence of other variables that have also been associated with improved performance in the current course and the reviewed literature.

Student attendance at class lectures and entering GPA have sometimes been associated with overall performance in college courses in studies reported in the literature (Balch, 1998; Clump, Bauer, & Whiteleather, 2003; Gunn, 1993; Hancock, 1994; Lamdin, 1996; Wilder, Flood, & Stromsnes, 2001). Additionally, research in previous semesters of the current course (Dotson, et al., In press) indicated that the amount of extra credit students earned and their attendance at mock exam sessions were correlated with

final performance in the course. In the present study, I measured each of those variables throughout the semester so that analyses could be done to determine their relationship with exam performance and with final grade in the course. I also wanted to examine the relationship between the variables to determine if there was a tendency for only the stronger students (e.g., those with higher entering GPAs and who earned more extra credit) to attend mock exam sessions. If such a relationship existed it may suggest that the effects of the mock exam seen in the three studies were only a result of the stronger students attending.

Analyses. First, six multiple linear regression analyses were conducted using SPSS software. For the first five analyses, the dependent variables were each student's five unit exam scores. The independent variables (predictor variables) were mock exam attendance during the particular unit, entering GPA, the number of class lectures attended during that unit, and the percentage of extra credit earned by the student during the same unit. The sixth analysis input final course grade as indicated by the percentage of total points earned as the dependent variable. The independent variables were mock exam attendance, entering GPA, the number of class lectures attended, and the percentage of extra credit earned by the student across the entire semester. Second, correlation analyses were conducted between the predictor variables to see if there were any relationships between the variables.

Results. The first five multiple linear regression analyses looked at student performance on the five unit exams and the effect of the predictor variables on that performance. For the first exam, when exam score was predicted it was found that GPA (Beta = 7.328, $p < 0.01$) and mock exam attendance (Beta = 13.883, $p < 0.0001$) were

significant predictors. Amount of extra credit earned (Beta = 0.064, n.s.) and attendance at lectures (Beta = 0.078, n.s.) were not significant predictors. The overall model fit was $R^2 = 0.385$. Stepwise analyses produced equivalent results even when lectures attended and extra credit earned were entered into the model first. Table 23 presents the results of this analysis. The results of the analysis for second, third, and fifth exam scores replicated those of the first exam, with the independent variables predicting exam performance at a significant level. Entering GPA and mock exam attendance predicted exam performance at a significant level and extra credit earned and lectures attended did not predict exam performance at a significant level. Additionally, stepwise analyses produced equivalent results even when lectures attended and extra credit earned were entered into the models first. Tables 24, 25, and 27 present the results of those analyses. For the fourth exam, when exam score was predicted it was found that GPA (Beta = 9.407, $p < 0.05$) and lecture attendance (Beta = 4.034, $p < 0.0001$) were significant predictors. Amount of extra credit earned (Beta = -0.046, n.s.) and mock exam attendance (Beta = 5.101, n.s.) were not significant predictors. The overall model fit was $R^2 = 0.468$. Stepwise analyses produced equivalent results, even when mock exams attended was entered into the model first. Table 26 presents the results of that analysis.

An analysis of the final grade earned by students in the course was also conducted. The independent variables of interest predicted final course grade at a significant level (overall model fit was $R^2 = 0.673$). Entering GPA was a significant predictor of final course grade (Beta = 11.099, $p = <0.001$). A one point increase in entering GPA (e.g., from 2.5 to 3.5) was related to an eleven percent predicted increase in final course grade when controlling for lectures attended, percent of extra credit earned,

and number of mock exams attended. Mock exams attended also predicted final course grade at a significant level (Beta = 2.922, p=<0.0001) when controlling for the other variables. Each mock exam attended predicted an almost three percent increase in final course grade. Percent of extra credit earned also significantly predicted final course grade (Beta = 0.104, p=.029); a one percent increase in extra credit earned predicted a tenth-of-a-percent increase in final course grade when controlling for the other variables in the model. The number of lectures attended, however, was not a significant predictor of final course grade (Beta = .221, p=.195). When entering GPA, mock exams attended, and percentage of extra credit earned were included in the model, lecture attendance did not account for a significant percentage of additional variance. Stepwise analyses produced equivalent results, even when lectures attended was entered into the model first. Table 28 presents the results of this analysis.

Table 29 presents data about the relationships between the different variables measured in the study, including: entering GPA, lectures attended, extra credit earned, and attendance at mock exam sessions. The table presents the correlation coefficients for each comparison, along with an indication of whether or not the results were statistically significant. Student entering GPA, attendance at lectures, and extra credit earned were all positively correlated with each other at a significant level. Students with a higher entering GPA also tended to earn more extra credit and attend more lectures. There was not a significant correlation found between class attendance, entering GPA, or extra credit earned and attendance at mock exam sessions.

Discussion. The multiple linear regression analyses of exam performance indicated that entering GPA predicted exam performance for all five unit exams.

Attendance at mock exams also predicted exam score for four of the five unit exams. Attendance at lectures only predicted exam performance for the fourth exam, and the amount of extra credit earned did not predict exam performance on any of the unit exams once the other independent variables were accounted for. When looking at final course grade, all of the predictor variables except attendance at lectures predicted final course grade.

Two aspects of these results were surprising. First, data from prior semesters of the course indicated a strong correlation between extra credit earned and performance on course exams. We assumed the correlation was the result of the practice students got from completing the practice questions. Second, given the amount of material covered in lectures that was not available anywhere else, we assumed that attendance at lectures would be predictive of performance in the course. That appeared to be the case only for the fourth exam, which covered more material than any other and was the first unit exam given by the second instructor in the course. The results of the analyses showing that neither extra credit earned nor lecture attendance were associated with large improvements in course performance (and that writing answers did not improve performance in the first study) suggest that students' entering GPAs were responsible for not just performance in the course but also attendance at lectures and completion of practice questions, and that writing answers as practice may not have played the functional role we thought it did in improving student exam performance in this course. These findings also highlight the importance of careful and sophisticated analysis of the interaction between variables in any course and may explain why earlier studies examining the effects of attendance and extra credit earned on performance report mixed

results (Clump, et al., 2003; Dotson, et al., In press; Gunn, 1993; Hancock, 1994; Lamdin, 1996; Padilla-Walker, 2006; Wilder, et al., 2001).

Another interesting result of the analyses involved the relationship of attendance at mock exam sessions to other behaviors and characteristics of the students. Because attendance at mock exam sessions was optional, the positive effects seen as a result of attending the mock exam sessions may have been a result of only the better students (e.g., higher entering GPAs, more often attending class, earning more extra credit) attending the sessions. The results of the correlation analyses do not support that conclusion for the present study. There was no significant positive correlation found between attendance at mock exam sessions and any of those three factors. The lack of a significant correlation suggests that the positive effects seen for the mock exam session were not the result of only the stronger students attending mock exam sessions. The lack of positive correlation also suggests that the mock exams were helpful to students of all ability levels. Such a finding would be exciting, since earlier studies have reported that review activities are often most helpful only for the better students (Aamodt, 1982a, 1982b; Balch, 1998; Padilla-Walker, 2006). Future research should explore the degree to which procedures like the mock exam improve performance for students of all ability levels.

Overall, the results of the multiple linear regression and correlation analyses suggest that performance in the present course was influenced by a number of variables. The results also suggest the degree to which each component of the course contributed to how well students performed, sometimes in surprising ways. Figure 14 represents a preliminary model of the interaction of the variables in the course based on those results. The direction and strength of the relationships between variables is indicated by the

arrows. Thicker arrows represent stronger influences while thinner arrows represent weaker influences. If no arrows connect two boxes, no functional relationship is assumed to exist. While the model is only a suggestion, it does provide a framework within which to conduct future research evaluating the role of the individual course components relative to each other. The model also complements the further analysis of the specific components of the mock exam sessions that the first three studies suggest should be continued.

Reliability

Table 30 presents the reliability measures for both the dependent and independent variables in the study. Dependent variable reliability was calculated for both student exam performance (overall 96%, range 94-97%) and attendance at class lectures and mock exam review sessions (100% for both). Several measures of independent variable reliability were calculated, including treatment integrity across types of mock exam sessions (overall 99%, range 94-100%), treatment integrity within experimental conditions (overall 94%, range 91-97%), and treatment integrity regarding student completion of the grading key during the third, fourth, and fifth exams (100% for all sessions).

General Discussion

The purpose of the present investigation was to conduct a component analysis of the variables present in a mock exam study session in order to make the sessions both more effective and efficient. Specifically, the goal was to identify which components of the mock exam study session were responsible for improvements in student exam performance. An additional series of analyses explored the effect of the mock exam

sessions when other variables such as entering GPA and attendance at lectures were controlled for. Requiring students to write answers to mock exam questions did not produce any differences in exam performance versus not asking students to write answers to questions. Requiring students to evaluate and correct sample answers versus watching and listening to the course GTA evaluate and correct sample answers also did not produce any differences in exam performance. There was a statistically significant difference in exam performance between questions for which students evaluated and corrected sample answers and questions for which students only received the grading criteria in the absence of any discussion. Students performed better, on average, on questions for which sample answers were evaluated and corrected than they did on questions for which they were only given the grading criteria. Across all units of the course students who attended mock exam sessions scored higher on exams than students who did not. Multiple linear regression analyses indicated that both entering GPA and attendance at mock exams regularly predicted exam performance while attendance at lectures only predicted performance on the fourth exam and amount of extra credit earned did not predict exam performance at all. Multiple linear regression analysis also indicated that entering GPA, mock exam attendance, and extra credit earned predicted final course grade, while attendance at lectures did not.

This research contributes to the literature on improving college student exam performance in several ways. First, most of the previous literature on improving exam performance used multiple choice assessments. This study demonstrated that supports can be designed and implemented for exams that required more complex and sophisticated short-essay answers. Thus, the present study extends the literature on

improving exam performance and provides a replication of the positive effect seen by Rust and colleagues in their 2003 study (Rust et al., 2003) using a similar review session. Second, the present study begins the evaluation of the individual components of educational supports to identify which are important and effective at improving student performance, and suggests that the effectiveness of a component may be determined, at least in part, by the type of assessment being used. Third, the study provides potential methodological models for evaluating educational supports by utilizing both within-subject and across-groups comparisons and providing comprehensive measures of independent and dependent variable reliability. An additional methodological contribution was the use of more sophisticated measurement and analysis of variables such as student entering GPA and attendance at lectures and their interaction with the effects of the mock exam on student performance in the course.

The existing empirical literature on improving student exam performance suggested that several variables were most often associated with improvements: similarity of the review activities to the actual exam, provision of a clear description of performance expectations or a question pool, requiring students to write answers to questions, and providing feedback or sample answers. The results of the present study generally supported those relationships, but also suggested that the effectiveness of those variables changed across different types of assessments.

In previous studies, researchers found the largest amounts of improvement in student exam and quiz performance when review materials were similar to the actual exam in both form and content. Across all of the experimental conditions in the present study, during mock exams, students were provided realistic examples of potential exam

questions in the same format and about the same content as the questions on the actual exam. Regardless of the experimental condition, students attending the mock exam scored higher on questions than students who did not attend the mock exam sessions. This finding generally supports the conclusion that providing students review materials similar in form and content to the actual exam improves performance on that exam. The differences seen across conditions in the last study, and especially the failure to find differences across conditions in the first two studies, however, suggest that for exams requiring students to engage in more complex behaviors such as application of knowledge to new situations, there is more to the most effective interventions than just providing a set of questions, a key, and asking students to answer the exact same questions that will be on the exam. The interactive nature of the mock exam sessions improved student exam performance, independent of the type of response required (i.e., whether students wrote answers, wrote grades for sample answers, or verbally discussed sample answers). Only when students were not allowed to discuss questions at all, as was done in the materials only condition of the last three mock exam sessions, did a difference appear between student performances on mock exam questions. This suggests that the sessions evoked behavior of students (both their overt verbal behavior during the sessions and perhaps their covert, private thoughts about the material being discussed) that was not measured, but which positively influenced their ability to more successfully apply their knowledge to questions on the unit exams. The mock exam sessions, then, appeared to set the occasion for students to engage in behaviors which helped them learn most successfully, and those behaviors occurred across all of the experimental conditions that involved any type of discussion within the session. Future research should attempt to

identify more exactly the form and function of those behaviors. For example, the behaviors could involve rehearsal of application skills (e.g. saying answers to questions either out loud or silently) or the learning and practice of discrimination skills (e.g. when to use an application skill or whether or not a particular answer is sufficient or requires more details to earn full credit); both of which are important to doing well on complex essay exams. The ability to discriminate correct from incorrect or sufficient from insufficient answers in particular may be an important aspect of what the mock exam taught students, and is related to another component explored in the present study: providing clear expectations.

In prior research, the variable associated with the next highest average amount of improvement was the provision of clear performance expectations and/or a question pool to students. As with the degree of similarity, students across all experimental conditions who attended the mock exam were provided information about which questions were likely to appear on the exam and the grading criteria for those questions. Students attending mock exams (where those things were provided) scored higher on questions than students who did not. This supports the conclusion that providing clear performance goals and/or a question pool improves performance. The third study, however, demonstrated that provision, while helpful, was not as effective as practice in applying both the material and the expectations. In one condition, students were simply given copies of potential exam questions and the grading keys for those questions. In the second condition, students were led through a discussion of the specific grading criteria for each question within the context of evaluating and correcting sample answers provided by the course GTA. Both conditions led to statistically significant

improvements in exam performance relative to students who did not attend the mock exam sessions, but the second, more active condition, produced significantly larger improvements in performance. Students scored the highest when they not only received clear expectations, but practiced applying those expectations with guidance from the GTA. This supports the assertion above that the mock exam sessions evoked practice behavior beyond just rehearsal of application skills, and that in order to perform at higher levels students had to do more than just memorize potential questions and the grading criteria (something they could do for all questions). The large effects seen for providing learning goals and question pools seen in the reviewed literature (Jenkins & Neisworth, 1973; Miles, et al., 1967; Semb, et al., 1973) occurred in the context of courses utilizing multiple-choice assessments where students could memorize the provided questions and their answers in order to improve performance. The current study required students to apply knowledge to a new situation and write answers in an essay format to a question similar to the one seen on the mock exam, a more challenging task, and one for which the results of the present study suggest more intensive supports are needed. Future research might further explore parametric manipulations of procedural intensity (e.g. providing different numbers of study questions or holding review sessions of different lengths and breadths) in an effort to identify critical levels of support required to produce optimal learning while not being overly demanding of time. For example, instructors for a course may decide that the additional improvements associated with the more intensive mock exam sessions are not worth the effort relative to the more modest gains associated with the much easier provision of materials to students without the scheduling of an outside of class review session. Likewise, additional research may identify degrees of support

associated with maximal improvements. For example, it may be that students only need exposure to discussion of five or six questions and sample answers to gain the benefit of the active discussion, and any additional, and more time consuming, discussions do not produce proportional increases in performance. Additionally, if practice applying grading criteria and learning expectations proves to be an important to achieving complex application to exam situations, the same investigations into the amount of practice required (e.g., how many questions for which grading criteria need to be explained and applied in detail to achieve desired outcomes) should be conducted. By identifying the point at which additional supports do not produce additional benefits, instructors can tailor review activities to optimize both their time and the benefits to students.

The effect of requiring students to write answers and of providing feedback or access to sample answers was not as clear either in the reviewed literature or the present study. Because the effects of these variables produced smaller average improvements than the first two variables discussed above and because they were rarely presented in the absence of additional supports, a clear demonstration of their effect was difficult to establish. For example, while the study by Malanga and Sweeney (2008) appeared to demonstrate the superiority of requiring written responses versus asking student to engage in verbal practice and hold up response cards, Simon (2005) found that requiring students to write answers to study questions in class in conjunction with providing a study guide did not produce any additional benefits relative to providing the study guide alone. These results, apparently contradictory, point to a larger issue when dealing with the evaluation of components of educational supports, including the current study: the presence of multiple, potentially confounding or contradictory variables in many

educational procedures. In the Malanga and Sweeney (2008) study comparing response cards to written study questions, the procedures used were fairly simple, not comprehensive (they did not review all of the material covered in each unit of the course), and the comparison was between two procedures that included few steps. With such simple procedures, the effect of requiring a written response in the experimental condition was more pronounced and the procedures allowed a discrimination of the effect of that single variable. In the present investigation and the Simon study (2005), rather than evaluate the effect of writing alone, the effect of writing was evaluated within the context of a more comprehensive review, and the failure to find an effect may have been a result of the ceiling effect discussed earlier.

The presentation of multiple variables as part of a treatment package while only manipulating a single variable, as done in the first study, was a procedural limitation to the current study. A related limitation was the failure to measure the behavior of the students during the mock exam sessions. An important step for future research in this area will be the isolation and evaluation of individual components of educational supports utilizing experimental methodologies that minimize the likelihood of confounds and ceiling effects. Future research will have to account for not only the instructor's behavior (e.g., how material is discussed in class, the form of materials provided by the instructor), but also the behavior of the students in the course (e.g., their verbal behavior during review sessions, their accuracy in answering questions during reviews).

An additional goal for future research involves identifying the components and mechanisms by which procedures produce improvements in learning. For example, two components may have similar effects on exam performance, but for different reasons. A

procedure such as frequent, short, in-class quizzes may produce improvements in learning due to the provision of timely feedback on student practice answers –an important dimension in improving acquisition of behaviors. The provision of study guides or learning objectives may lead to a similar increase in performance because such materials act as a prompt for students to guide their study behaviors more efficiently. Careful component analyses could reveal both the effect of each component on performance, and also the functional mechanism responsible for those effects.

Conclusion

Overall, the present study extends the literature on the effectiveness of review activities at improving student exam and quiz performance in several important ways. First, the demonstration of the effectiveness of the mock exam sessions presents evidence that educational supports containing components that have been effective at improving multiple-choice exam performance can also be effective at improving performance on short essay exams requiring more complex, application-based answers. The demonstration of effectiveness also highlights several areas where supports may vary in effectiveness across types of exams (e.g. the higher overall improvement seen on multiple-choice exams when providing students only with study guides versus the smaller effect seen for doing so in the third type of mock exam session) and suggests the need for continued investigation into the circumstances under which particular support procedures are most appropriate for different types of assessments. Future research should continue to evaluate and develop procedures designed to improve learning of more complex skills and for assessments of other formats.

Second, the present study presented the results of preliminary analyses of individual components of a mock exam study session. While no single component was identified as responsible for changes in exam performance, it was demonstrated that different packages of components (i.e., provision of questions and grading key versus discussion, evaluation, and correction of sample answers) led to differential effects on exam performance. The present study also identified additional components that should be explored, particularly those related to behaviors evoked by discussion of performance expectations and sample answers. The active nature of the mock exam sessions highlighted the complex nature of the interactions and the behaviors those interactions evoked. It will be important in analyzing the impact of such active support procedures to identify the type of behavior it produces and targets (e.g. practice describing a procedure versus practice evaluating a description of a procedure), and to explore what role each type of practice plays in improving exam performance.

Third, the present study provides potential methodological models for evaluating educational supports by utilizing both within-subject and across-groups comparisons and providing comprehensive measures of independent and dependent variable reliability. The within-subject comparisons made in the present study were unique because they involved comparing performances by the students on the same unit exams and because the students experienced both experimental conditions during the same mock exam session. By presenting both experimental conditions within the same mock exam session, the design controlled for variables that may have influenced results in earlier studies, such as the motivation levels of the students (e.g., how close in time to the actual exam review materials were presented), the difficulty of the particular unit exam (e.g., different

units of a course may be more difficult than others), and aspects of the mock exam session such as GTA mood or pace of the session (e.g., moving faster, covering more materials, or hearing different questions asked and answered across different sessions).

Additionally, because none of the previously reviewed studies reported comprehensive measures of either independent or dependent variable reliability, it may be that the differential outcomes for the same variables seen across studies were a result of differences in implementation of educational supports or in students' behaviors within sessions that were not captured in the reports. By demonstrating that reliability measures can be collected regarding the integrity of the independent variables (and some difficulties associated with a comprehensive collection of treatment integrity variables, i.e., the inability to document students' verbal behaviors in the second study) and the reliable scoring of dependent measures, the present study improves the literature by providing data (and a method for gathering it) that allow a more careful analysis of the factors responsible for the results seen in the study.

Finally, the present study identified ways to more efficiently conduct the existing mock exam study session without losing effectiveness. By eliminating the time spent writing answers under exam-like conditions from the session and focusing discussion more carefully on evaluation and correction of sample answers, the sessions were shortened from two-and-a-half hours to less than ninety minutes. The time savings were significant and by shortening the session, it is hoped other instructors will be more willing to adopt supports similar to the mock exam when using essay and short-essay-based exams.

In summary, the present study presented the successful demonstration of the effectiveness of a mock exam study session to improve the performance of students on application-based, short-essay exams while also exploring the components responsible for the improvements within an experimental design utilizing both within-subject and across-groups analyses of performance and comprehensive reporting of both independent and dependent variable reliability. The results of the study generally confirmed trends seen in the empirical literature that educational supports that resemble the actual exam in both form and content and that provide clear educational objects can produce significant improvements in student exam performance, but also suggested that more variables are involved when the exam requires more complex skills. In particular, while providing realistic practice materials and clear expectations in the form of a grading key led to improvements in learning, additional aspects of the mock exam sessions appeared to set the occasion for students to engage with the course material in ways that led to larger improvements in their learning than that seen when not discussing the material in the session. While no single component of the mock exam session was identified as having an effect on student exam performance, a package of components involving discussion, evaluation, and correction of a sample answer to a mock exam question produced superior performance over simply providing students with copies of study materials. Additionally, multiple linear regression analyses indicated that attendance at mock exam sessions was significantly predictive of both individual exam scores and final grade earned in the course even when attendance at lectures, extra credit earned, and student entering GPA were controlled for. Also, since attendance at mock exam sessions was not correlated with entering GPA, lecture attendance, or extra credit earned it is unlikely that

the positive effects of the mock exam session were the result of only the stronger students attending the sessions. Thus, the sessions appeared to be effective for students of all ability levels.

References

Aamodt, M. G. (1982a). A closer look at the study session. *Teaching of Psychology, 9*(4), 234-235.

Aamodt, M. G. (1982b). The effect of the study session on test performance. *Teaching of Psychology, 9*(2), 118-120.

Austin, J. L., Lee, M. G., Thibeault, M. D., Carr, J. E., & Bailey, J. S. (2002). Effects of Guided Notes on University Students' Responding and Recall of Information. *Journal of Behavioral Education, 11*(4), 243-254.

Baker, L., & Lombardi, B.-R. (1985). Students' lecture notes and their relation to test performance. *Teaching of Psychology, 12*(1), 28-32.

Balch, W. R. (1998). Practice versus review exams and final exam performance. *Teaching of Psychology, 25*(3), 181-185.

Bol, L., & Hacker, D.-J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *Journal of Experimental Education, 69*(2), 133-151.

Brosvic, G. M., & Epstein, M. L. (2007). Enhancing learning in the introductory course. *Psychological Record, 57*(3), 391-408.

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

Buzhardt, J., & Semb, G.-B. (2002). Item-by-item versus end-of-test feedback in a computer-based PSI course. *Journal of Behavioral Education, 11*(2), 89-104.

Clayton, M. C., & Woodard, C. (2007). The effect of response cards on participation and weekly quiz scores of university students enrolled introductory psychology courses. *Journal of Behavioral Education, 16*(3), 250-258.

Clump, M.-A., Bauer, H., & Whiteleather, A. (2003). To attend or not to attend: Is that a good question? *Journal of Instructional Psychology, 30*(3), 220-224.

Cornelius, T. L., & Owen-DeSchryver, J. (2008). Differential effects of full and partial notes on learning outcomes and attendance. *Teaching of Psychology, 35*(1), 6-12.

Dickson, K. L., & Bauer, J. J. (2008). Do students learn course material during crib sheet construction? *Teaching of Psychology, 35*(2), 117-120.

Dickson, K. L., Devoley, M.-S., & Miller, M.-D. (2006). Effect of study guide exercises on multiple-choice exam performance in introductory psychology. *Teaching of Psychology, 33*(1), 40-42.

Dickson, K. L., Miller, M.-D., & Devoley, M.-S. (2005). Effect of Textbook Study Guides on Student Performance in Introductory Psychology. *Teaching of Psychology, 32*(1), 34-39.

Dickson, K. L., & Miller, M. D. (2005). Authorized Crib Cards Do Not Improve Exam Performance. *Teaching of Psychology, 32*(4), 230-233.

Dickson, K. L., & Miller, M. D. (2006). Effect of crib card construction and use on exam performance. *Teaching of Psychology, 33*(1), 39-40.

Dorow, L. G., & Boyle, M. E. (1998). Instructor feedback for college writing assignments in introductory classes. *Journal of Behavioral Education, 8*(1), 115-129.

Dotson, W. H., Sheldon, J. B., & Sherman, J. A. (In press). Supporting Student Learning: Improving Performance on Short-Essay Exams Using Realistic Practice Opportunities. *Journal of the Scholarship of Teaching and Learning*.

Drabick, D. A. G., Weisberg, R., Paul, L., & Bubier, J. L. (2007). Keeping it short and sweet: Brief, ungraded writing assignments facilitate learning. *Teaching of Psychology, 34*(3), 172-176.

Fleming, V.-M. (2002). Improving students' exam performance by introducing study strategies and goal setting. *Teaching of Psychology, 29*(2), 115-119.

Grabe, M., & Christopherson, K. (2008). Optional student use of online lecture resources: Resource preferences, performance and lecture attendance. *Journal of Computer Assisted Learning, 24*(1), 1-10.

Gunn, K.-P. (1993). A correlation between attendance and grades in a first-year psychology class. *Canadian Psychology, 34*(2), 201-202.

Hancock, T.-M. (1994). Effects of mandatory attendance on student performance. *College Student Journal, 28*(3), 326-329.

Hautau, B., Turner, H. C., Carroll, E., Jaspers, K., Krohn, K., Parker, M., et al. (2006). Differential Daily Writing Conditions and Performance on Major Multiple-Choice Exams. *Journal of Behavioral Education, 15*(3), 171-181.

Hove, M. C., & Corcoran, K. J. (2008). If you post it, will they come? Lecture availability in introductory psychology. *Teaching of Psychology, 35*(2), 91-95.

Huxham, M. (2005). Learning in lectures: Do 'interactive windows' help? *Active Learning in Higher Education, 6*(1), 17-31.

Jenkins, J.-R., & Neisworth, J.-T. (1973). The facilitative influence of instructional objectives. *Journal of Educational Research Vol, 66*(6), 254-256.

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: the ABCs of reliability. *Journal of Applied Behavior Analysis, 10*, 141-150.

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

Kellum, K. K., Carr, J. E., & Dozier, C. L. (2001). Response-card instruction and student learning in a college classroom. *Teaching of Psychology, 28*(2), 101-104.

Kulik, J. A., Jaksa, P., & Kulik, C.-l. C. (1978). Research on component features of Keller's Personalized System of Instruction. *Journal of Personalized Instruction, 3*(1), 2-14.

Lamdin, D.-J. (1996). Evidence of student attendance as an independent variable in education production functions. *Journal of Educational Research, 89*(3), 155-162.

Landrum, R. E. (2007). Introductory psychology student performance: Weekly quizzes followed by a cumulative final exam. *Teaching of Psychology, 34*(3), 177-180.

Lloyd, M. E., & Lloyd, K. E. (1986). Has lightening struck twice? *Teaching of Psychology, 13*, 149-151.

Malanga, P. R., & Sweeney, W. J. (2008). Increasing active student responding in a university applied behavior analysis course: The effect of daily assessment and response cards on end of week quiz scores. *Journal of Behavioral Education, 17*(2), 187-199.

Martin, G., & Pear, J. (2007). *Behavior modification: What it is and how to do it (8th ed.)*. Upper Saddle River: Pearson Prentice Hall.

Mayfield, K.-H., & Chase, P.-N. (2002). The effects of cumulative practice on mathematics problem solving. *Journal of Applied Behavior Analysis, 35*(2), 105-123.

Miles, D.-T., Kibler, R.-J., & Pettigrew, L. E. (1967). The Effects of Study Questions on College Students' Test Performances. *Psychology in the Schools*(1), 25-26.

Miller, M. L., & Malott, R. W. (1997). The importance of overt responding in programmed instruction even with added incentives for learning. *Journal of Behavioral Education, 7*(4), 497-503.

Miller, M. L., & Malott, R. W. (2006). Programmed Instruction: Construction Responding, Discrimination Responding, and Highlighted Keywords. *Journal of Behavioral Education, 15*(2), 111-119.

Morling, B., McAuliffe, M., Cohen, L., & DiLorenzo, T. M. (2008). Efficacy of personal response systems ("clickers") in large, introductory psychology classes. *Teaching of Psychology, 35*(1), 45-50.

NCES. (2002). *Profile of undergraduates in U.S. postsecondary education institutions: 1999-2000*.

Neef, N. A., Cihon, T., Kettering, T., Guld, A., Axe, J. B., Itoi, M., et al. (2007). A comparison of study session formats on attendance and quiz performance in a college course. *Journal of Behavioral Education, 16*(3), 235-249.

Nevid, J. S., & Mahon, K. (2009). Mastery quizzing as a signalling device to cue attention to lecture material. *Teaching of Psychology, 36*(1), 29-32.

Oliver, R., & Williams, R.-L. (2005). Direct and Indirect Effects of Completion Versus Accuracy Contingencies on Practice-Exam and Actual-Exam Performance. *Journal of Behavioral Education, 14*(2), 141-152.

Padilla-Walker, L. M. (2006). The Impact of Daily Extra Credit Quizzes on Exam Performance. *Teaching of Psychology, 33*(4), 236-239.

Peterson, L., Homer, A. L., & Wonderlich, S. A. (1982). The Integrity of Independent Variables in Behavior Analysis. *Journal of Applied Behavior Analysis, 15*(4), 477-492.

Poirier, C. R., & Feldman, R. S. (2007). Promoting active learning using individual response technology in large introductory psychology classes. *Teaching of Psychology, 34*(3), 194-196.

Price, M., O'Donovan, B., & Rust, C. (2007). Putting a social-constructivist assessment process model into practice: Building the feedback loop into the assessment process through peer review. *Innovations in Education and Teaching International, 44*(2), 143-152.

Rust, C., Price, M., & O'Donovan, B. (2003). Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment & Evaluation in Higher Education, 28*(2), 147-164.

Semb, G., Hopkins, B. L., & Hursh, D.-E. (1973). The effects of study questions and grades on student test performance in a college course. *Journal of Applied Behavior Analysis Vol, 6*(4), 631-642.

Shabani, D. B., & Carr, J. E. (2004). An Evaluation of Response Cards as an Adjunct to Standard Instruction in University Classrooms: A Systematic Replication and Extension. *North American Journal of Psychology, 6*(1), 85-100.

Simon, J. L. (2005). Effects of a Study Guide and Written Rehearsal on Undergraduate Test Performance. Unpublished doctoral dissertation, University of Kansas, Lawrence.

Taveggia, T. C. (1976). Personalized instruction: A summary of comparative research, 1967-1974. *American Journal of Physics, 44*, 1028-1033.

Wilder, D.-A., Flood, W.-A., & Stromsnes, W. (2001). The use of random extra credit quizzes to increase student attendance. *Journal of Instructional Psychology, 28*(2), 117-120.

Table 1

Summary of articles included in component analysis

Article	*Similarity With Assessment*	*Criteria Provided*	*Response Type During Review*	*Answers and/or Feedback Provided During Review*	*% Improvement*
Aamodt, 1982a	SC/NSF	YES	No response	NO	8.9%
Aamodt, 1982b Cond. 1	SC/NSF	YES	No response	NO	6.5%
Cond. 2	NSC/NSF	NO	No response	NO	0%
Austin, et al, 2002	SC/NSF	YES	No response	NO	5.8%
Clayton & Woodard, 2007	NSC/SF	NO	Other response	Not clear	4%
Dickson, et al, 2005	NSC/NSF	NO	Written response	YES	2.8%
Drabick, et al, 2007	NSC/NSF	NO	Written response	Not clear	3%, 6%
Fleming, 2002	NSC/NSF	NO	No response	NO	0%
Flora & Logan, 2006	SC/NSF	YES	Written response	Feedback only	2%
Jenkins & Neisworth, 1973	SC/SF	YES	No response	NO	26.7%
Malanga & Sweeney, 2008 Cond. 1	SC/NSF	YES	Written response	YES	12.5%
Cond. 2	SC/NSF	NO	Other response	Not clear	0%
Miles, et al, 1967	SC/SF	YES	No response	NO	14.7%
Miller & Malott, 1997 Cond. 1	SC/SF	NO	Written response	Not clear	16.1%
Cond. 2	SC/SF	NO	Written response	Not clear	11.1%
Miller & Malott, 2006	SC/SF	NO	Written response	Feedback only	11%
Morling, et al, 2008	NSC/SF	NO	Other response	Feedback only	1.5%
Neef, et al, 20007 Exp. 1 Cond. 1	NSC/NSF	NO	Other response	Feedback only	9.5%
Cond. 2	NSC/NSF	NO	No response	NO	0%
Exp. 2 Cond. 1	NSC/NSF	NO	Other response	Feedback only	2%
Cond. 2	NSC/NSF	NO	No response	NO	0%
Poirer & Feldman, 2007	NSC/Not clear	NO	Other response	Answers only	1.31%

Price, et al, 2007			SC/SF	YES	Written response	Feedback only	1.6%
Rust, et al, 2003			SC/SF	YES	Written response	YES	12%, 14%
Semb, et al, 1973	Cond. 1		SC/SF	YES	Written response	YES	33.5%
	Cond. 2		SC/SF	YES	No response	NO	28%
	Cond. 3		SC/SF	YES	Written response	YES	19.25%
Shabani & Carr, 2004	Exp. 1	Cond. 1	NSC/SF	NO	Other response	YES	0%
	Exp. 2	Cond. 1	NSC/SF	NO	Other response	YES	3.6%
		Cond. 2	NSC/SF	NO	Other response	YES	2.6%

Table 2. *Analysis of percentage improvement in student performance by how similar the review procedure is to the actual assessment*

Similar Content/Not Similar Format	**Similar Content/Similar Format**
Avg. Imp: 5.95%	Avg. Imp: 17.09%
5 Studies	7 Studies:
Aamodt, 1982a 8.9% Aamodt, 1982b Cond 1 6.5% Austin, et al, 2002 5.8% Flora & Logan, 2006 2% Malanga & Sweeney, 2008 Cond 1 12.5% Cond 2 0%	Jenkins & Neisworth, 1973 26.7% Miles, et al, 1967 14.7% Miller & Malott, 1997 Cond 1 16.1% Cond 2 11.1% Miller & Malott, 2006 11% Price, et al, 2007 1.6% Rust, et al, 2003 12% 14% Semb, et al, 1973 Cond 1 33.5% Cond 2 28% Cond 3 19.25%
Not Similar Content/Not Similar Format	**Not Similar Content/Similar Format**
Avg. Imp: 2.59%	Avg. Imp: 2.34%
5 Studies	3 Studies
Aamodt, 1982b Cond 2 0% Dickson, et al, 2005 2.8% Drabick, et al, 2007 3% 6% Fleming, 2002 0% Neef, et al, 20007 Exp 1 Cond 1 9.5% Cond 2 0% Exp 2 Cond 1 2% Cond 2 0%	Clayton & Woodard, 2007 4% Morling, et al, 2008 1.5% Shabani & Carr, 2004 Exp 1 Cond 1 0% Exp 2 Cond 1 3.6% Cond 2 2.6%

Table 3. *Analysis of percentage improvement in student performance when providing and not providing specific competency requirements and/or a question pool to students*

No Specific Competency Requirements/Question Pool Provided		**Provided Specific Competency Requirements/Question Pool**	
Avg. Imp: 3.92%		Avg. Imp: 14.27%	
12 Studies		10 Studies	
Aamodt, 1982b Cond 2	0%	Aamodt, 1982a	8.9%
Clayton & Woodard, 2007	4%	Aamodt, 1982b Cond 1	6.5%
Dickson, et al, 2005	2.8%	Austin, et al, 2002	5.8%
Drabick, et al, 2007	3%	Flora & Logan, 2006	2%
	6%	Jenkins & Neisworth, 1973	26.7%
Fleming, 2002	0%	Malanga & Sweeney, 2008 Cond 1	12.5%
Malanga & Sweeney, 2008 Cond 2	0%	Miles, et al, 1967	14.7%
Miller & Malott, 1997 Cond 1	16.1%	Price, et al, 2007	1.6%
Cond 2	11.1%	Rust, et al, 2003	12%
Miller & Malott, 2006	11%		14%
Morling, et al, 2008	1.5%	Semb, et al, 1973 Cond 1	33.5%
Neef, et al, 20007 Exp 1 Cond 1	9.5%	Cond 2	28%
Cond 2	0%	Cond 3	19.25%
Exp 2 Cond 1	2%		
Cond 2	0%		
Poirer & Feldman, 2007	1.31%		
Shabani & Carr, 2004 Exp 1 Cond 1	0%		
Exp 2 Cond 1	3.6%		
Cond 2	2.6%		

Table 3a. *Analysis of percentage improvement in student performance when providing specific competency requirements or a question pool to students*

Provided Specific Competency Requirements/Question Pool

Avg. Imp: 14.27%

10 Studies

Aamodt, 1982a		8.9%
Aamodt, 1982b Cond 1		6.5%
Austin, et al, 2002		5.8%
Flora & Logan, 2006*		2%
Jenkins & Neisworth, 1973		26.7%
Malanga & Sweeney, 2008 Cond 1		12.5%
Miles, et al, 1967		14.7%
Price, et al, 2007		1.6%
Rust, et al, 2003		12%
		14%
Semb, et al, 1973	Cond 1	33.5%
	Cond 2	28%
	Cond 3	19.25%

Specific Competency Requirements Provided

Avg. Imp: 12.08%

5 Studies

Aamodt, 1982a	8.9%
Aamodt, 1982b Cond 1	6.5%
Austin, et al, 2002	5.8%
Jenkins & Neisworth, 1973	26.7%
Malanga & Sweeney, 2008 Cond 1	12.5%

Question Pool Provided

Avg. Imp: 17.58%

4 Studies

Miles, et al, 1967		14.7%
Price, et al, 2007		1.6%
Rust, et al, 2003		12%
		14%
Semb, et al, 1973	Cond 1	33.5%
	Cond 2	28%
	Cond 3	19.25%

*Flora and Logan, 2006 not included in lower analysis –insufficient information to determine if questions from study guide were re-used on exam.

Table 3b. *Analysis of percentage improvement in student performance when providing a question pool to students for either a multiple-choice or essay-based exam*

Question Pool Provided

Avg. Imp: 17.58%

4 Studies

Miles, et al, 1967		14.7%
Price, et al, 2007		1.6%
Rust, et al, 2003		12%
		14%
Semb, et al, 1973	Cond 1	33.5%
	Cond 2	28%
	Cond 3	19.25%

Question Pool Provided for Multiple Choice Exams

Avg. Imp: 23.86%

2 Studies

Miles, et al, 1967		14.7%
Semb, et al, 1973	Cond 1	33.5%
	Cond 2	28%
	Cond 3	19.25%

Question Pool Provided for Essay Exams

Avg. Imp: 9.2%

2 Studies

Price, et al, 2007	1.6%
Rust, et al, 2003	12%
	14%

Table 4. *Analysis of percentage improvement in student performance across different response requirements for students*

Required Other Response that was Not Written	**Required Written Response of Some Kind**
Avg. Imp: 2.72%	Avg. Imp: 11.14%
6 Studies	9 Studies
Clayton & Woodard, 2007 4%	Dickson, et al, 2005 2.8%
Malanga & Sweeney, 2008 Cond 2 0%	Drabick, et al, 2007 3%
Morling, et al, 2008 1.5%	6%
Neef, et al, 20007 Exp 1 Cond 1 9.5%	Flora & Logan, 2006 2%
Exp 2 Cond 1 2%	Malanga & Sweeney, 2008 Cond 1 12.5%
Poirer & Feldman, 2007 1.31%	Miller & Malott, 1997 Cond 1 16.1%
Shabani & Carr, 2004 Exp 1 Cond 1 0%	Cond 2 11.1%
Exp 2 Cond 1 3.6%	Miller & Malott, 2006 11%
Cond 2 2.6%	Price, et al, 2007 1.6%
	Rust, et al, 2003 12%
	14%
	Semb, et al, 1973 Cond 1 33.5%
	Cond 3 19.25%
No Response of Any Kind Required	
Avg. Imp: 9.06%	
8 Studies	
Aamodt, 1982a 8.9%	
Aamodt, 1982b Cond 1 6.5%	
Cond 2 0%	
Austin, et al, 2002 5.8%	
Fleming, 2002 0%	
Jenkins & Neisworth, 1973 26.7%	
Miles, et al, 1967 14.7%	
Neef, et al, 20007 Exp 1 Cond 2 0%	
Exp 2 Cond 2 0%	
Semb, et al, 1973 Cond 2 28%	

Table 5. *Analysis of percentage improvement in student performance when providing or not providing sample/correct answers and/or performance feedback to students*

Sample or Correct Answer Given without Performance Feedback		**Sample or Correct Answer Given with Performance Feedback**	
Avg. Imp: 1.31%		Avg. Imp: 11.1%	
1 Study		5 Studies	
Poirer & Feldman, 2007	1.31%	Dickson, et al, 2005	2.8%
		Malanga & Sweeney, 2008 Cond 1	12.5%
		Rust, et al, 2003	12%
			14%
		Semb, et al, 1973 Cond 1	33.5%
		Cond 3	19.25%
		Shabani & Carr, 2004 Exp 1 Cond 1	0%
		Exp 2 Cond 1	3.6%
		Cond 2	2.6%
Neither Sample or Correct Answer Nor Performance Feedback Given		**No Sample or Correct Answer, but Performance Feedback Given**	
Avg. Imp: 9.06%		Avg. Imp: 4.6%	
8 Studies		5 Studies	
Aamodt, 1982a	8.9%	Flora & Logan, 2006	2%
Aamodt, 1982b Cond 1	6.5%	Miller & Malott, 2006	11%
Cond 2	0%	Morling, et al, 2008	1.5%
Austin, et al, 2002	5.8%	Neef, et al, 20007 Exp 1 Cond 1	9.5%
Fleming, 2002	0%	Exp 2 Cond 1	2%
Jenkins & Neisworth, 1973	26.7%	Price, et al, 2007	1.6%
Miles, et al, 1967	14.7%		
Neef, et al, 20007 Exp 1 Cond 2	0%		
Exp 2 Cond 2	0%		
Semb, et al, 1973 Cond 2	28%		

Table 6

Questions to be answered by mock exam study sessions

Mock Exam Study Session Type 1:

Research Question: How effective is writing answers to short-answer questions and evaluating and discussing student-provided answers relative to seeing, evaluating, and discussing sample answers at improving student performance exam questions?

Brief Methods: The primary comparison was a within-subjects comparison between accuracy of answering exam questions that students had practiced writing answers to and discussed during the mock exam and accuracy of answering questions that students had only discussed during the mock exam. An additional comparison was made between students attending and not attending the mock exam sessions.

Mock Exam Study Session Type 2:

Research Question: How effective is having students evaluate and correct sample answers to short-answer questions relative to seeing the course TA evaluating and discussing sample answers at improving student performance on short-essay exam questions?

Brief Methods: The primary comparison was a within-subjects comparison like that done for the first exam. An additional comparison was made between students attending and not attending the mock exam sessions.

Mock Exam Study Session Type 3:

Research Question: Do students write better answers on questions for which they have heard the TA describe the grading criteria and they have evaluated and improved a sample answer versus questions for which they received the grading criteria but did not hear the grading criteria discussed and did not evaluate or improve a sample answer?

Brief Methods: The primary comparison was a within-subjects comparison like that done for the first two exams. Additional comparisons were made across groups on a question by question basis and between students attending and not attending the mock exam sessions.

Table 7

Reliability scoring items for session integrity

For the mock exam session in the video, did the GTA:

Discuss the general material to be included on the exam:

YES / NO

Describe the structure of the mock exam session:

YES / NO

Provide students with keys to doing well on the exam:

YES / NO

Hand out a mock exam:

YES / NO

Hand out a mock exam grading key:

YES / NO

Ask students to write answers to mock exam questions:

YES / NO

Provide students time (20 min) to write answers to mock exam questions:

YES / NO

Ask students to fill out a grading key evaluating sample answers to mock exam questions:

YES / NO

Provide students time (15 min) to fill out the grading key for sample answers to mock exam questions:

YES / NO

Discuss specific grading criteria for questions on the mock exam:

YES / NO

Ask students to volunteer their answers to mock exam questions:

YES / NO

Ask students to volunteer their evaluation of answers to mock exam questions:

YES / NO

Show students a sample answer to a mock exam question:

YES / NO

Explain how he would grade a sample answer (only one he provided) to a mock exam question:

YES / NO

Ask students to correct a sample answer to a mock exam question:

YES / NO

Explain how he would correct a sample answer (only one he provided) to a mock exam question:

YES / NO

Table 8

Reliability scoring items for experimental condition integrity: Mock exam session type 1

For the question above, did the GTA:
Describe the things the students must do to correctly answer the question for each part of the question:
YES / NO
Describe the way he would grade that question on the exam in reference to the points/credit earned or not earned:
YES / NO
Show a sample answer to students on the overhead:
YES / NO
Did students volunteer their answer to the question:
YES / NO
Require the students to evaluate an answer (either student-volunteered answer or GTA-provided sample answer):
YES / NO
Require the students to successfully correct an answer (either student-volunteered answer or GTA-provided sample answer):
YES / NO

Table 9

Reliability scoring items for experimental condition integrity: Mock exam session type 2

For the question above, did the GTA:
Describe the things the students must do to correctly answer the question for each part of the question:
YES / NO
Describe the way he would grade that question on the exam in reference to the points/credit earned or not earned:
YES / NO
Show a sample answer to students on the overhead:
YES / NO
Ask the students to evaluate a sample answer provided by the GTA:
YES / NO
Did the students successfully correct a sample answer provided by the GTA:
YES / NO
Describe how he would evaluate a sample answer provided by the GTA:
YES / NO
Did the GTA successfully correct a sample answer provided by the GTA:
YES / NO

Table 10

Reliability scoring items for experimental condition integrity: Mock exam session type 3

For the question above, did the GTA:

Describe the things the students must do to correctly answer the question for each part of the question:

YES / NO

Describe the way he would grade that question on the exam in reference to the points/credit earned or not earned:

YES / NO

Show a sample answer to students on the overhead:

YES / NO

Ask the students to evaluate a sample answer provided by the GTA:

YES / NO

Did the students successfully correct a sample answer provided by the GTA:

YES / NO

Table 11

Structure of First Type of Mock Exam Study Session

Activity	*Time Spent*
Brief Introduction	15m
GTA passed out blank mock exam	2-3m
Students wrote answers on mock exam	20m
GTA passed out grading key	2-3m
GTA-led question-by-question discussion: Grading criteria Evaluation of answers Correction of answer	70-80m

Table 12

General study tips provided to students for exam 1

- Know the details
- Be specific in your descriptions
- Tell me what it looks like in THIS situation
- Read the chapters and answer the study questions

Table 13

Comparison of conditions in first type of mock exam study session

Student Writing of Answers to Mock Exam Questions	*VS*	*No Student Writing of Answers to Mock Exam Questions*
Students receive grading criteria for question		Students receive grading criteria for question
Students write their own answers to mock exam question		GTA provides sample answer to mock exam question
GTA reviews grading criteria for question		GTA reviews grading criteria for question
Students evaluate volunteered student answers to question		Students evaluate provided sample answer to question
Students correct volunteered student answers to questions		Students correct provided sample answer to question

Note: The mock exam questions that students wrote answers to were counterbalanced across mock exam sessions.

Table 14

Material covered in Unit 1 and how covered in mock exam study session

Topics of Questions	Wrote Answer	Sample Answer Provided
Measurement Systems	Ver. 1	Ver. 2
Reliability	Ver. 2	Ver. 1
Experimental Designs:		
Reversal Design	Ver. 1	Ver. 2
Multiple-Baseline Design	Ver. 2	Ver. 1
Social Validity:		
Normative Measures	Ver. 1	Ver. 2
Consumer Satisfaction	Ver. 2	Ver. 1

Topics not counterbalanced across sessions (and not included in analysis) included: Behavioral Definitions

Table 15

Structure of second type of mock exam study session

Activity	*Time Spent*
Brief Introduction	15m
GTA passed out completed mock exam with sample answers filled in and mock exam grading key	2-3m
GTA-led question-by-question discussion of: Grading criteria Evaluation of sample answer Correction of sample answer	90m

Table 16

Comparison of conditions in second type of mock exam study session

Student Evaluating	*GTA Evaluating*
Students receive grading criteria for question	Students receive grading criteria for question
GTA provides sample answer to mock exam question	GTA provides sample answer to mock exam question
GTA reviews grading criteria for question	GTA reviews grading criteria for question
Students evaluate provided sample answer to question	GTA evaluates provided sample answer to question
Students correct provided sample answer to questions	GTA corrects provided sample answer to question

Table 17

Material covered in Unit 2 and how covered in mock exam study session

Topics of Questions	Students Evaluated Answer	Watched GTA Evaluate Answer
Teaching a non-verbal behavior:		
Positive reinforcer	Ver. 1	Ver. 2
Shaping	Ver. 1	Ver. 2
Prompts and prompt fading	Ver. 2	Ver. 1
When/where behavior occur	Ver. 2	Ver. 1
How to maintain behavior	Ver. 1	Ver. 2
Teaching verbal imitation:		
Positive reinforcer	Ver. 1	Ver. 2
Reinforcing all verbalizations	Ver. 2	Ver. 1
Modeling first sound	Ver. 1	Ver. 2
Shaping latency	Ver. 1	Ver. 2
Shaping topography	Ver. 1	Ver. 2
Second sound and discrimination	Ver. 1	Ver. 2
Chaining sounds to form words	Ver. 2	Ver. 1
Labeling:		
Model	Ver. 1	Ver. 2
Fading prompts	Ver. 2	Ver. 1
Second label and discrimination	Ver. 1	Ver. 2

Table 18

Structure of third type of mock exam study session

Activity	*Time Spent*
Brief Introduction	15m
GTA passed out mock exam with some sample answers provided and a mock exam grading key	2-3m
Students graded sample answers on mock exam	20m
GTA collected completed grading keys and passed out blank grading key	2-3m
GTA-led question-by-question discussion: Grading criteria Evaluation of sample answers Correction of sample answer	70m

Table 19

Comparison of conditions in third type of mock exam study session

Evaluation of Mock Exam Questions	*VS*	*No Evaluation of Mock Exam Questions*
Students receive grading criteria for question		Students receive grading criteria for question
GTA provides sample answer to mock exam question		
GTA reviews grading criteria for question		
Students evaluate provided sample answer to question		
Students correct provided sample answer to questions		

Table 20

Material covered in Unit 3 and how covered in mock exam study session

Topics of Questions	Students Discussed Question	Students Did Not Discuss Question
6 Procedures to Reduce Problem Behavior		
DRO	Ver. 1	Ver. 2
DRL	Ver. 2	Ver. 1
Extinction	Ver. 1	Ver. 2
Response Cost	Ver. 1	Ver. 2
Time-Out	Ver. 2	Ver. 1
Aversive Stimulus	Ver. 2	Ver. 1

Topics not counterbalanced across sessions (and not included in analysis) included:
Functional Assessment
Functional Alternative Behaviors
Preventitive Strategies
Choosing Procedures
Considerations when selecting procedures
Informed Consent

Table 21

Material covered in Unit 4 and how covered in mock exam study session

Topics of Questions	Students Discussed Question	Students Did Not Discuss Question
Reciprocity Counseling	Ver. 1	Ver. 2
Client-Therapist Contract	Ver. 2	Ver. 1
Aversion Therapy and Self-Management	Ver. 1	Ver. 2
Systematic Desensitization	Ver. 2	Ver. 1

Topics not counterbalanced across sessions (and not included in analysis) included:
Parent-Child Contract
Self-Control Contract

Table 22

Material covered in Unit 5 and how covered in mock exam study session

Topics of Questions	Students Discussed Question	Students Did Not Discuss Question
Token Economy:		
Backup Reinforcers	Ver. 1	Ver. 2
Form of Token	Ver. 2	Ver. 1
Method of Exchange	Ver. 2	Ver. 1
Delay to exchange	Ver. 2	Ver. 1
Considerations for token value of jobs/tasks	Ver. 1	Ver. 2
Considerations for token value of backup reinforcers	Ver. 2	Ver. 1
Fading tokens	Ver. 2	Ver. 1
Legal and Ethical Issues:		
Basic Rights	Ver. 1	Ver. 2
Institutional Labor	Ver. 2	Ver. 1
Aversive Techniques	Ver. 2	Ver. 1

Topics not counterbalanced across sessions (and not included in analysis) included:
Token Economy: Target behaviors, Baseline, and Reducing Problem Behaviors

Table 23

Results of linear regression analysis of predictor variables for Exam 1 grade (percentage of exam credit earned)

Variable	Betaweight	t	p-value
Constant (Intercept)	42.73	4.917	<0.0001***
Entering GPA	7.33	2.718	0.009**
Mock Exams Attended	13.88	3.801	<0.0001***
% Extra Credit Earned	0.03	0.507	0.614
Lectures Attended	0.73	0.6363	0.528

Analysis conducted using SPSS Software

Table 24

Results of linear regression analysis of predictor variables for Exam 2 grade (percentage of exam credit earned)

Variable	Betaweight	t	p-value
Constant (Intercept)	35.66	4.717	<0.0001***
Entering GPA	10.01	3.791	<0.0001***
Mock Exams Attended	15.81	4.442	<0.0001***
% Extra Credit Earned	-0.04	-0.023	0.982
Lectures Attended	0.93	1.043	0.302

Analysis conducted using SPSS Software

Table 25

Results of linear regression analysis of predictor variables for Exam 3 grade (percentage of exam credit earned)

Variable	Betaweight	t	p-value
Constant (Intercept)	18.11	2.301	<0.025*
Entering GPA	14.37	5.052	<0.0001***
Mock Exams Attended	13.62	3.995	<0.0001***
% Extra Credit Earned	-0.03	-0.425	0.673
Lectures Attended	0.34	0.056	0.627

Analysis conducted using SPSS Software

Table 26

Results of linear regression analysis of predictor variables for Exam 4 grade (percentage of exam credit earned)

Variable	Betaweight	t	p-value
Constant (Intercept)	13.75	1.328	0.190
Entering GPA	9.41	2.553	0.014*
Mock Exams Attended	5.10	0.833	0.409
% Extra Credit Earned	-0.46	-0.639	0.526
Lectures Attended	4.034	3.759	<0.0001***

Analysis conducted using SPSS Software

Table 27

Results of linear regression analysis of predictor variables for Exam 5 grade (percentage of exam credit earned)

Variable	Betaweight	t	p-value
Constant (Intercept)	28.609	3.453	0.001**
Entering GPA	10.677	3.596	0.001**
Mock Exams Attended	8.773	2.099	0.040*
% Extra Credit Earned	0.053	1.064	0.292
Lectures Attended	1.266	1.339	0.186

Analysis conducted using SPSS Software

Table 28

Results of linear regression analysis of predictor variables for final course grade (percentage of course credit earned)

Variable	Betaweight	t	p-value
Constant (Intercept)	26.94	4.375	<0.0001***
Entering GPA	10.91	5.236	<0.0001***
Mock Exams Attended	2.651	3.743	<0.0001***
% Extra Credit Earned	0.101	2.183	0.033*
Lectures Attended	0.238	1.411	0.164

Analysis conducted using SPSS Software

Table 29

Correlation analysis results (All tests report Pearson-r scores)

Variables	Correlation Coefficient
Entering GPA and Extra Credit Earned	0.45***
# Lectures Attended and Extra Credit Earned	0.61***
# Lectures Attended and Entering GPA	0.46***
Entering GPA and # Mock Exams Attended	0.13
Extra Credit Earned and # Mock Exams Attended	0.20
# Lectures Attended and # Mock Exams Attended	0.25

*** = $p < 0.0001$

Table 30

Reliability results

Test	% Agreement
Dependent Variable Reliability Measures	
Exam Scores	
Overall	96%
Experimental Conditions	90%
Control Conditions	92%
Did not attend mock exam	89%
Independent variable reliability	
Session Integrity	99%
IOA	98%
Condition Integrity	94%
Additional Variables	
Attendance at class lectures	100%
Attendance at mock exam sessions	100%

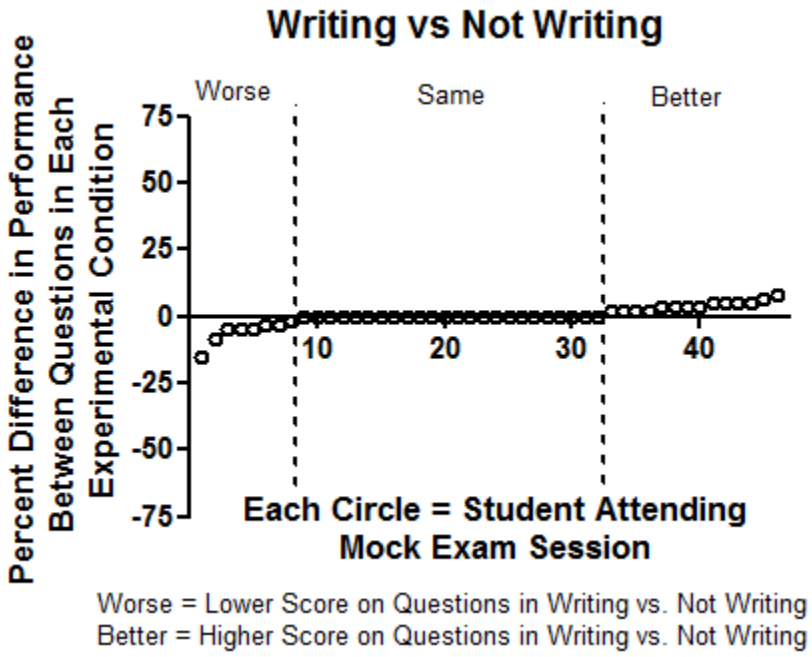

Figure 1. Writing answers versus not writing answers comparison

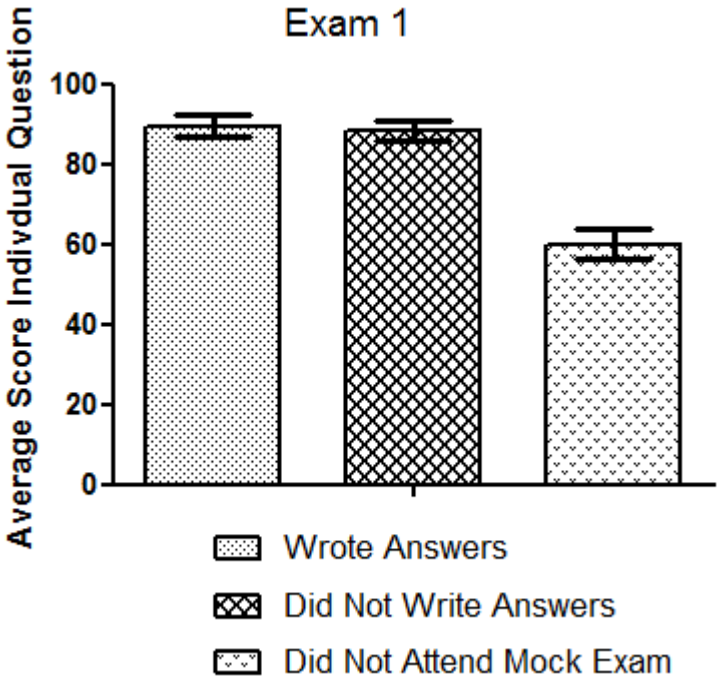

Figure 2. Exam 1 Across-groups comparison

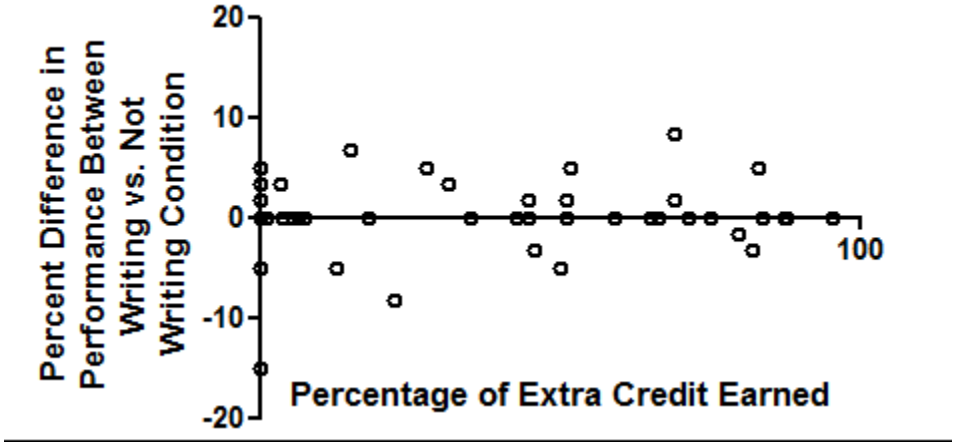

Figure 3. Difference score between writing and non-writing conditions by amount of extra credit earned (more extra credit = more practice writing correct answers).

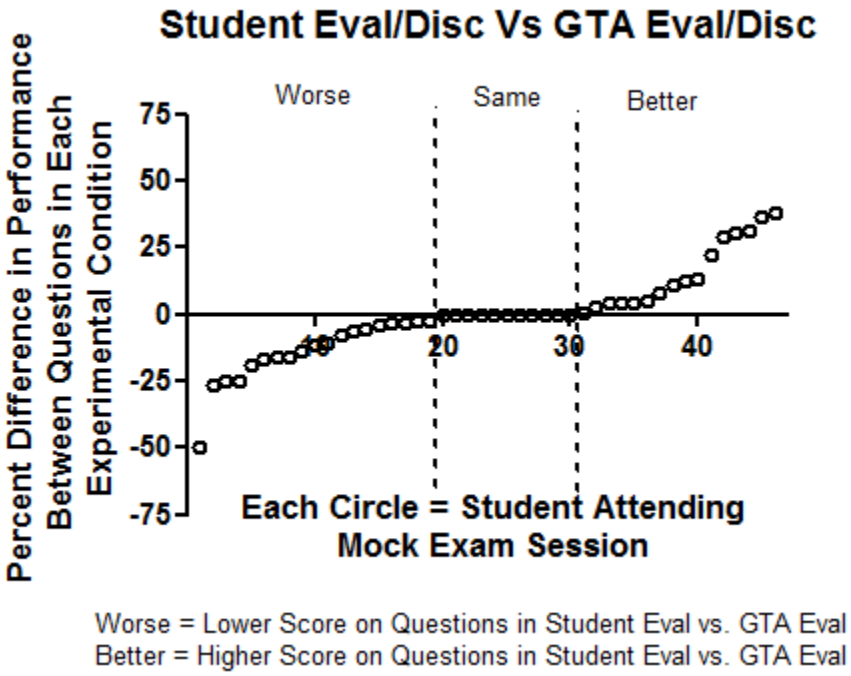

Figure 4. Student evaluation and discussion versus GTA evaluation and discussion comparison

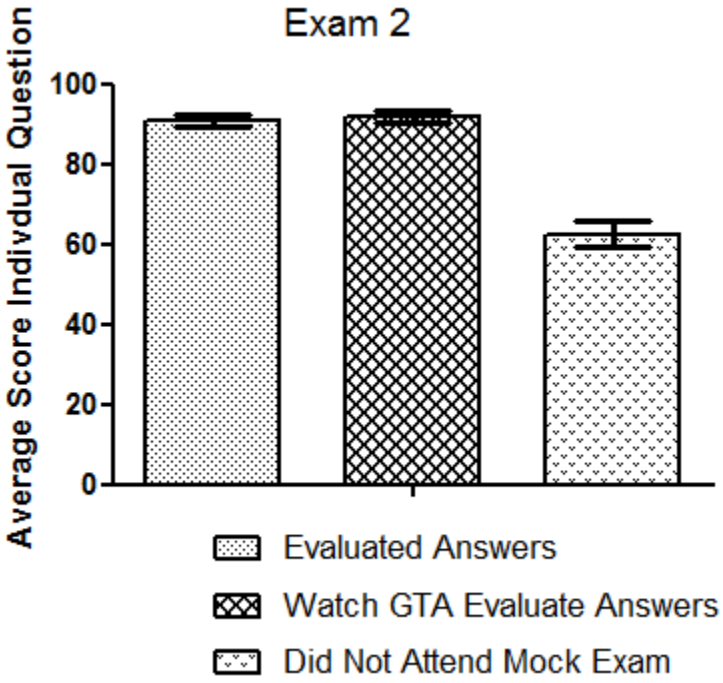

Figure 5. Exam 2 Across-groups comparison

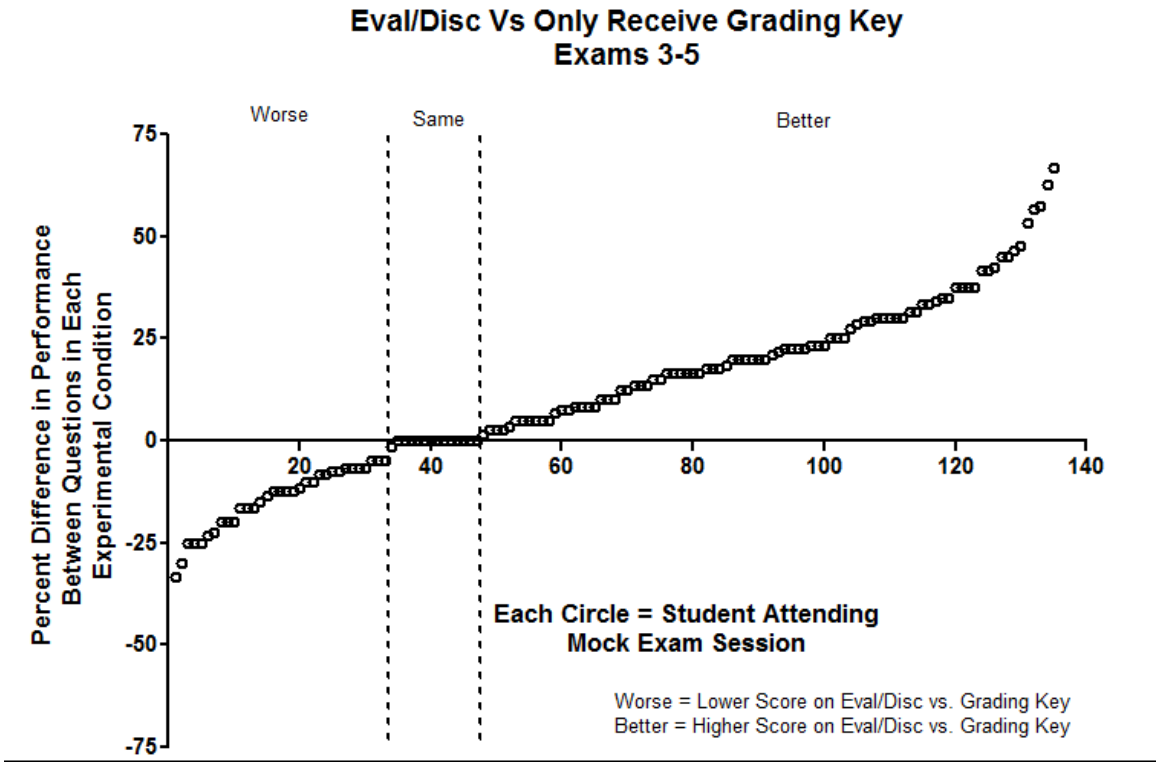

Figure 6. Evaluation and discussion of sample answers versus only receiving grading key **(**Exams 3-5) Within-subject comparison

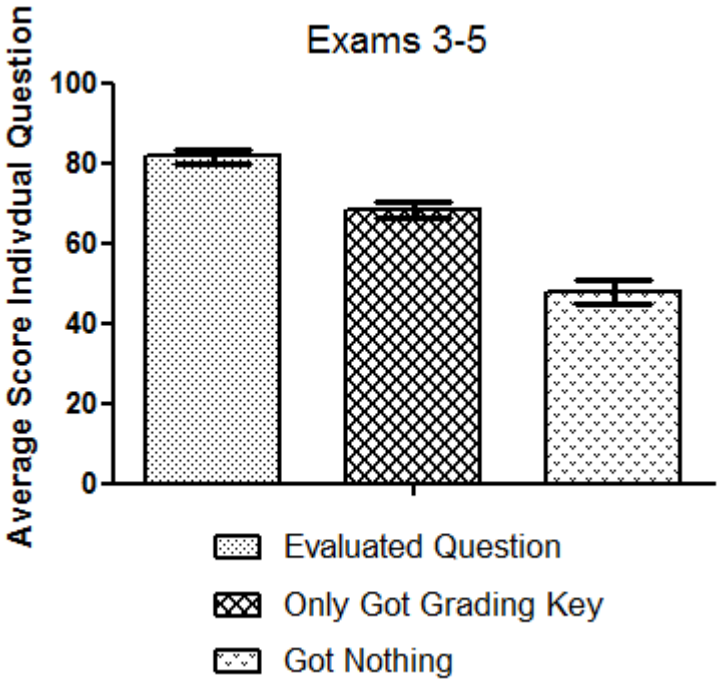

Figure 7. Exam 3-5 Across-groups comparison

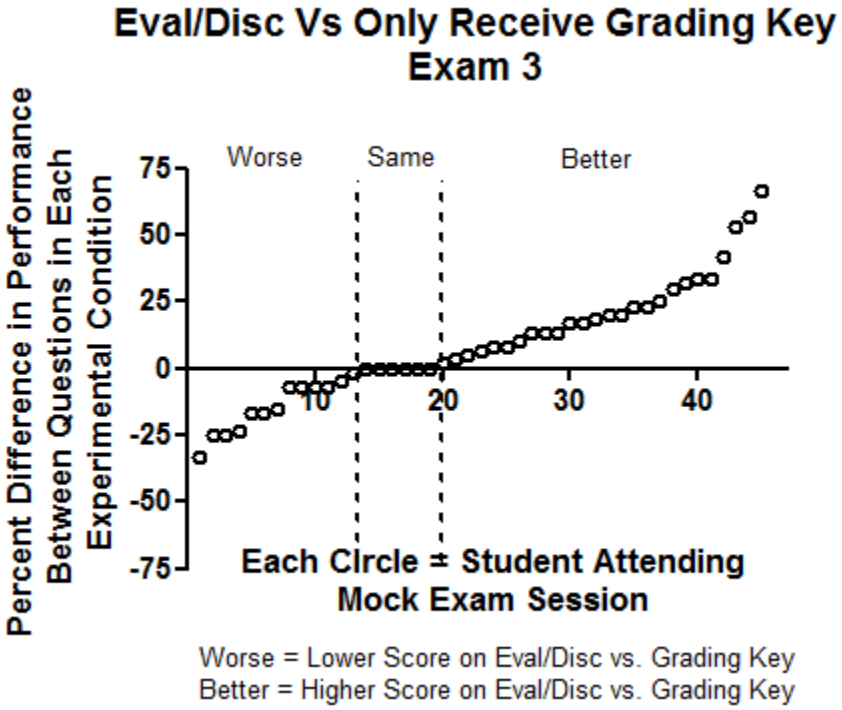

Figure 8. Evaluation and discussion of sample answers versus only receiving grading key **(**Exam 3) Within-subject comparison

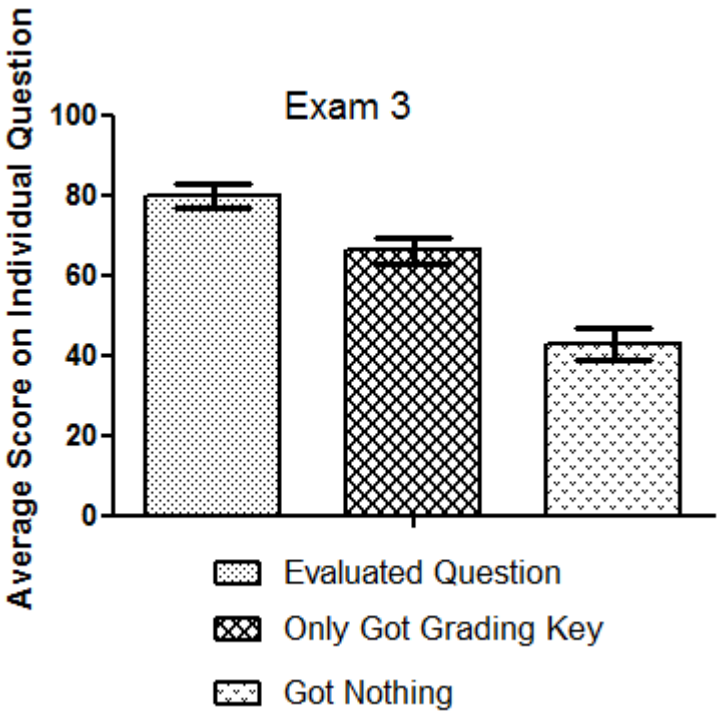

Figure 9. Exam 3 Across-groups comparison

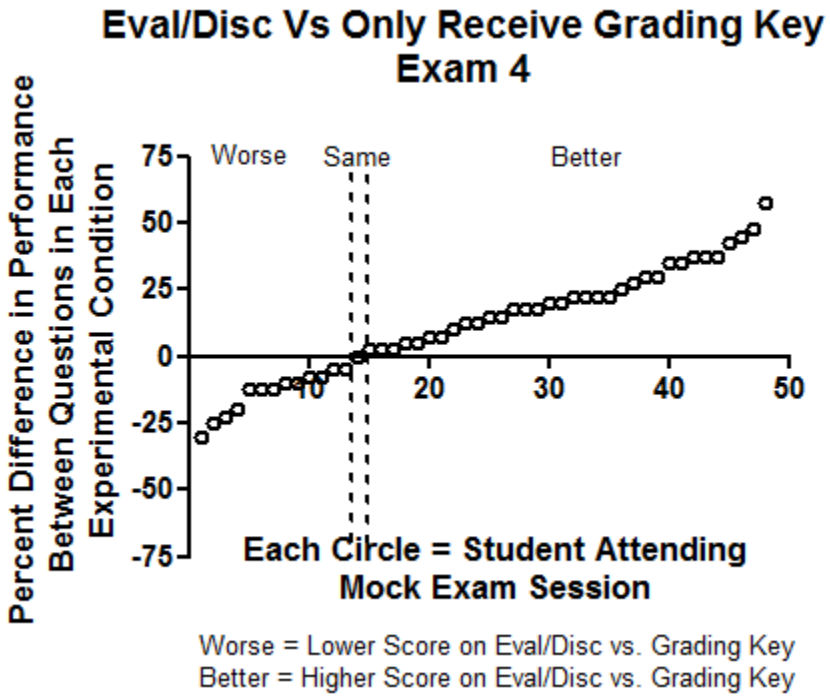

Figure 10. Evaluation and discussion of sample answers versus only receiving grading key **(**Exam 4) Within-subject comparison

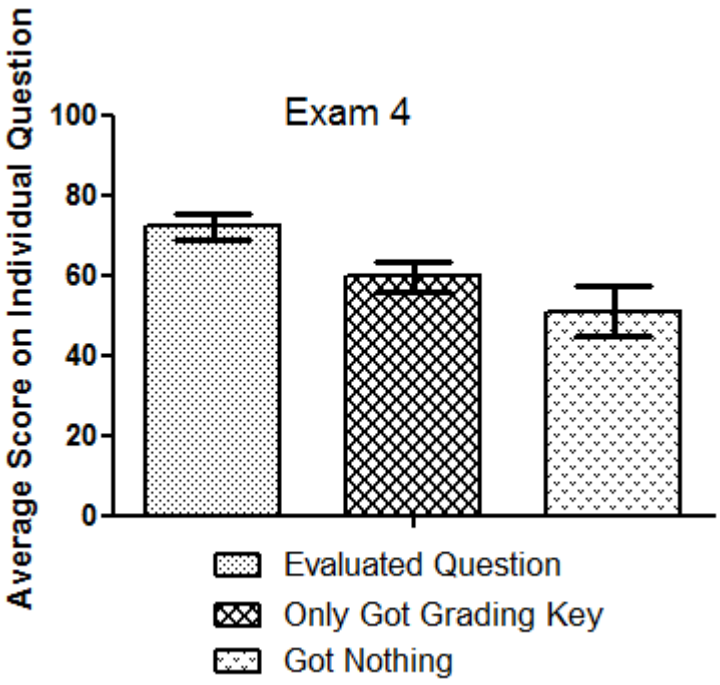

Figure 11. Exam 4 Across-groups comparison

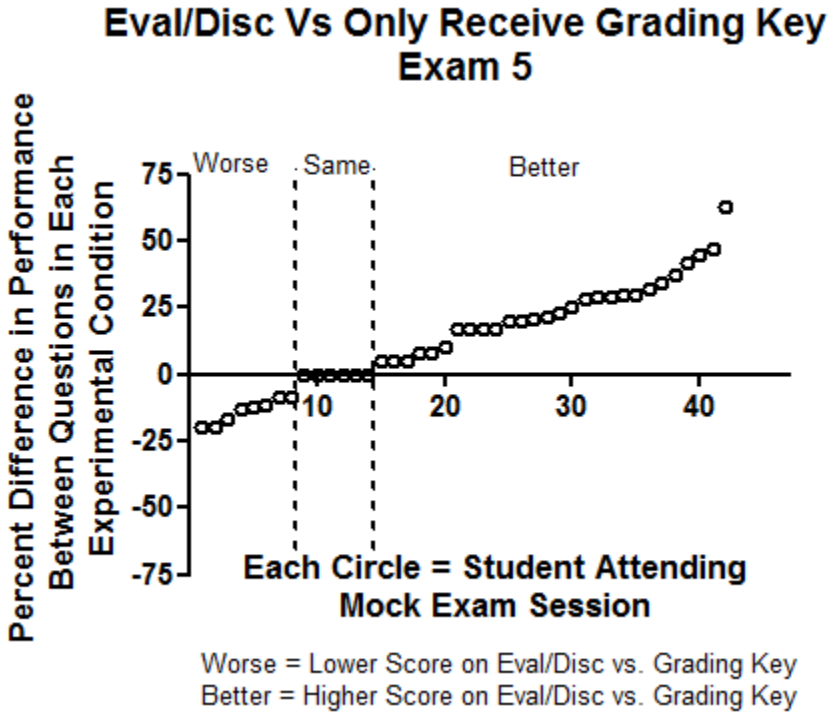

Figure 12. Evaluation and discussion of sample answers versus only receiving grading key **(**Exam 5) Within-subject comparison

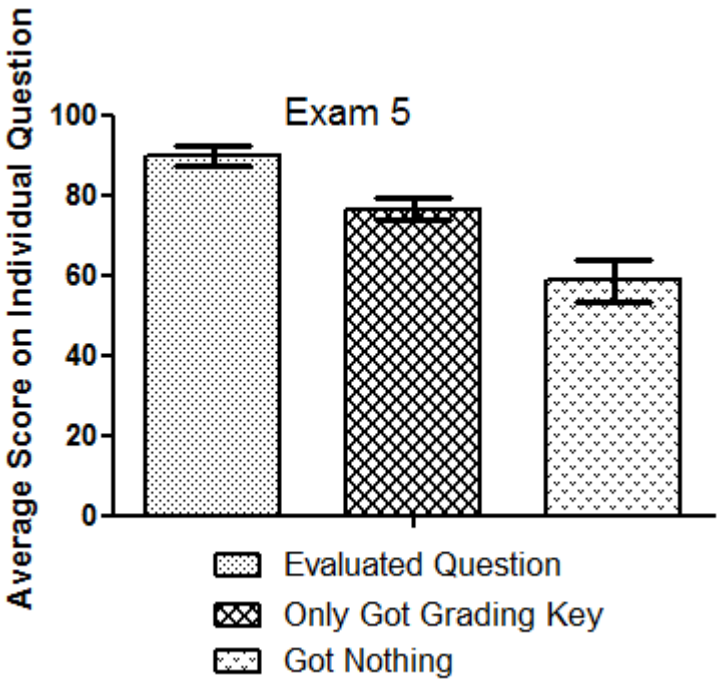

Figure 13. Exam 5 Across-groups comparison

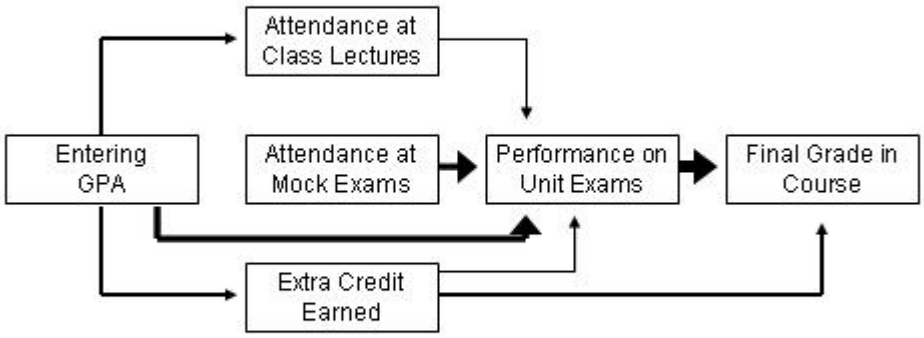

Figure 14. Interaction model for variables associated with performance in course. Thickness of line represents strength of variable influence.