Spring 2012

Formative assessment: Comparing immediate and delayed feedback

Jacob W. Williams
James Madison University

Formative Assessment: Comparing Immediate and Delayed Feedback

Jacob W. Williams

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Arts

Psychological Sciences

May 2012

Acknowledgements

This thesis would not have been possible without the help of a number of people along the way. First, I’d like to thank Whitney Smiley for laying the ground work for my thesis topic. Without her study who knows what my research would have entailed or when I would have completed it. Thank you to Meg Ward, Amanda Devoto, Kristen Kaufman, Alec Bernstein, Matt Meccarriello, and Nicole Roteliuk who met with me each week to grade quizzes and code data. I’d like to also thank my committee members Jessica Irons, for allowing me to do research in her class and Bryan Saville for his constructive criticism on this thesis and over the years. Also, thank you Bryan for introducing me to the philosophy of radical behaviorism. Without it my psychology career would have no doubt been much shorter and much less worthwhile.

Most importantly, I want to thank my advisor Dr. Tracy Zinn for her input, support, and presence throughout this thesis and master’s program. For the past two years and the four prior to that she has inspired me as well as countless others to strive for their goals, whatever those might be. I couldn’t think of a better exemplar of what it means to be an amazing academic, professional, person, and friend.

Table of Contents

List of Figures

Abstract

The current study compared feedback via the Immediate Feedback Assessment Technique to typical scantron with delayed feedback. Following students during the course of one semester in two introductory psychology courses, students alternated between using the IFAT and using a typical scantron with feedback for formative assessment before taking each exam. Summative examination (post-tests) performance on assessments containing questions comparable and identical to questions seen on formative pre-tests were compared in order to assess any differences in performance depending on type of formative assessment and feedback experienced. Results showed mixed findings as to whether the IFAT is more effective at increasing retention compared to more delayed feedback using the typical scantron response form. Limitations and future directions are provided.

I. Introduction

The main goal of any form of education should be learning. Not only should curriculum increase retention, but it should also do so in the most efficient manner by utilizing techniques that improve, increase, and facilitate the acquisition of knowledge. Due to time constraints, course objectives, and other priorities instructors should design courses so that they allow for maximum opportunity for reviewing course material. Instructors should use strategies that could make acquisition of course content more likely.

One important component heavily researched and shown to be beneficial to learning is feedback. Unfortunately in the typical class setting, due to class size and time constraints feedback on course work and assessments is often difficult and sometimes simply not possible (Moreno, 2004). There is also still debate about which characteristics of feedback are most important and what are realistic ways of administering feedback in the classroom (Shute, 2008).

One way to incorporate feedback is during testing. There has been much research suggesting that simply being tested on course material improves learning and long-term retention compared to other strategies (Bjork, 1975). Formative assessment in the form of multiple-choice testing together with efficient feedback methods may improve retention while leaving valuable class time to focus on upcoming course objectives. Recently and more specifically, the Immediate Feedback Assessment Technique, developed by Epstein, Epstein, and Brosvic (2001) provides corrective feedback after each response to a multiple choice question. Research on the IF AT has shown higher test performances on later exams compared to students originally tested without feedback. Although findings typically show feedback in general supports learning, the details and mechanism behind the feedback the IFAT provides during formative assessment is not yet entirely clear. Also, it is possible that other similar formats of teaching by testing are as effective. There still is discrepancy between how immediate feedback has to be to in order to maximize retention in a real classroom setting. Finally, research has not concluded whether the IFAT truly capitalizes on the best characteristics of feedback for educational purposes.

In this paper I plan to discuss (a) types of feedback used in educational settings, (b) general feedback guidelines, (c) feedback through testing, specifically IFAT research, (d) delayed vs. immediate feedback (IFAT) results differing depending on applied vs. lab, (e) and how the present study improves on current literature by comparing formative assessment techniques differing in the immediacy of feedback they provide.

Feedback

In general, feedback is meant to inform someone of where his or her performance stands in comparison to a particular goal and what he or she needs to do in order to reach that goal (Daniels & Daniels, 2006). Although there may be many different interpretations of feedback, this one seemed to be the most parsimonious. There is discrepancy as by which mechanism feedback operates. Daniels and Daniels (2006) argue that feedback serves as a discriminative stimulus whereby in feedback provides information on the possibility of reinforcement. As long as improved performance is the reinforcing consequence in this scenario, feedback may cue the availability of a positive outcome, such as receiving good exam scores. Other interpretations suggest feedback itself is a motivating operation that alters the reinforcing value of a consequence (Agnew & Redman, 1992).

Summative vs. Formative

Particularly in an education context, feedback can be referred to as either summative or formative depending on the type of assessment of performance the provider of feedback has carried out. Feedback from a typical exam that only provides a grade would be considered summative whereas feedback from an exam that provides information for improving later grades or performance would be considered formative. Summative feedback would only capitalize on the first definitional requirement of Daniels and Daniels (2006), that of simply providing the performer with information regarding how he or she performed, not what needs to be done to improve performance. For example, the grade received on an assessment or at the end of the year would be an example of summative feedback because it is information regarding the assessment of whether learning has taken place.

Formative feedback would be information provided during a course that was meant to improve learning. This type of feedback would address both parts of the Daniels and Daniels (2006) definition of feedback. Using the same definition, formative feedback not only tells a learner about performance, but also provides them with information on how to improve that performance. Shute (2008) outlined the features of formative feedback that are the most effective at improving learning and under what conditions this learning is retained. With an extensive review of the literature the author defines formative feedback as “information communicated to the learner that is intended to modify his or her thinking or behavior for the purpose of improving learning” (p. 2).

Also, Kulhavy and Stock (1989) reported that effective feedback provides the learner with verification and elaboration or information on whether a particular response is correct and then subsequently providing information that will guide the learner toward the correct response.

In her review Shute (2008) provides prescriptions for the best practices when applying feedback in a learner based setting depending on a number of variables such as time constraints, difficulty of the material, and learner characteristics.

Some of these guidelines include:

- focusing the feedback on the task as opposed to the learner.
- providing elaborated feedback which has been shown to be more effective than simple verification.
- present feedback in manageable units (Mayer & Moreno, 2002).
- feedback must be clear and specific (Moreno, 2004) and should be linked to specific goals and performance if possible.
- promote a “learning” goal orientation via feedback (Dweck, 1986)
- Only provide feedback after learners have attempted a solution. (Bangert-Drowns et al., 1991)

Shute (2008) also discussed the multiple forms of formative feedback ranging from no feedback at all to highly elaborate examples. For instance, as mentioned previously the least elaboration besides no feedback would simply be verifying whether the answer is correct. Further elaboration might consist of informing the learner of the correct option or ‘try again’ method otherwise known as “repeat-until-correct” feedback where feedback informs the learner that there has been an incorrect response and more attempts are necessary to discover the correct answer.

Shute (2008) provided numerous examples, many of which consist of feedback that might be suitable in separate academic settings and formats. For instance, feedback for an essay question might require more elaboration in larger units than perhaps feedback for a multiple-choice test where verification and slight elaboration are all that is necessary to facilitate acquisition of the correct response.

Multiple-Choice Testing as Formative Assessment

According to McDougall (1997) multiple-choice testing is the most prevalent response medium used in academics settings. With the typical multiple-choice tests, students choose one of several alternatives provided on the exam. They do not construct their own answers. For reasons including efficiency, ease of use, minimum effort and time required for grading, as well as objective evaluation, the multiple choice exam has become the primary means of student assessment (Roediger & Marsh, 2005). In addition, research has shown an added benefit to multiple-choice assessments when used appropriately.

Testing Effect. The main purpose of testing in education as well any other field is a means of assessing what has been learned and whether instruction has been successful. But researchers have also found that testing and more specifically multiple-choice testing may also hold the added benefit of being an effective study tool. Research suggests that when students are tested on material, retention of that material measured by a later test typically improves. This phenomenon is referred to as the testing effect (Bjork, 1975). Previous research suggests that students who take a multiple choice test first perform better on a subsequent test in comparison to those students who did not have a preliminary test (Butler & Roediger, 2007). As part of a larger study Butler, Karpicke, and Roediger (2007) had students read passages and take a preliminary multiple choice test that assessed retention of the information in the prose passages. Learning from the test was measured using a cued-response test given later consisting of fill in the blank and short answer. Participants who were previously given a multiple choice test on the material performed significantly better than those students who were not tested.

Research on the testing effect has resulted in relatively consistent and robust findings when investigating the phenomenon in a lab setting with multiple choice testing. Exposure to the testing format with similar material improves performance on later tests on similar material. The mechanism by which the testing effect operates is debated; however, one hypothesis is that it may be a result of increased incentive to review and critically evaluate course material above and beyond that of equivalent time spent studying. The preliminary assessment could also be more analogous to the actual testing session. If this were the case then it is likely that when later tested on similar material, previous practice better elicits the appropriate response such as in context dependent learning. The testing context could serve as discriminative stimuli that signal the opportunity for reinforcement (correct response to a test question) more adequately than studying alone. In other words, practicing by testing would improve later performance on a similar task because there are similar discriminative stimuli in the environment cueing a response (choosing a certain multiple choice item). But, although the testing effect has robustly shown improvements in retention compared to typical study methods, there is much room for improvement. In many of the testing effect studies, practice by testing did not provide formative feedback regarding whether responses were correct. If practice testing also provided feedback they would become a formative assessment that should increase future retention even more so.

Unfortunately there are also drawbacks to using tests in multiple-choice formats. Due to the fact that constructing multiple choice exams is difficult, professors and teachers typically re-use test questions. As a result, instructors may not review the test in class or allow copies of the test to leave the classroom or office thereby providing little feedback except for the overall score (Roediger & Marsh, 2005). This form of feedback is only summative and does not help the learner in the future by providing information on the correct responses to individual test items.

Roediger and Marsh (2005) also studied the possibility of incorrect alternative response options interrupting or hindering learning of correct material. This would certainly be an issue if the main goal of any formative testing is to assess learning and to increase retention of the information. When students are exposed to multiple choice testing formats with typically 4 alternative options, only one being correct, this incorrect alternative could lead to misinformation. If an incorrect alternative is chosen, the test taker may continue to believe they have the correct information, and thereby answer similarly if tested later on the material. Beg, Armour, and Kerr’s (1985) research suggested that even if the correct alternative was chosen, simply reading incorrect statements can increase the likelihood of participants believing that those statements are true. More recently researchers have referred to this phenomenon as the negative suggestion effect (Brown, Schilling, & Hockensmith, 1999). Research has found that if participants choose an incorrect response, when later given a similar question he or she interprets the previously selected response option as more correct than other incorrect responses never seen before (Toppino & Luipersbeck, 1993). Behaviorally, the test question (discriminate stimulus) cues that choosing a particular response will result in some form of reinforcement (i.e. correct answer-positive reinforcement; removal of anxiety of not yet choosing an answer-negative reinforcement) that would increase the likelihood of a similar response being chosen in the future under similar conditions (similar question in a later assessment). So without any corrective or formative feedback after the first test, there is no reason not to choose the same response. After the first exam the test question and the incorrect answer are now associated, and when faced with a difficult question where the correct response is unknown, the student chooses the more familiar incorrect response on association alone.

In Roediger and Marsh's (2005) experiment, four phases consisting of a passage reading, a multiple choice test, a filler task, and a final cued-recall test were carried out. Half the participants read the reading whereas the other half did not. Also, the number of response alternatives varied from zero, two, four, or six and the dependent measures were the proportion of correct answers and the proportion of errors on the final cued-recall test. Results of the study showed that although there was a consistent testing effect (i.e. higher performance on the second test simply as a result of prior testing than those who were not tested prior), those conditions with more incorrect response alternatives saw a much smaller magnitude of the testing effect. The increased amount of misinformation being read increased the likelihood of choosing an incorrect option on a later test, especially in conditions where there was no study. It seems that as the number of possible incorrect responses increased, the more experience the test takers would have with the misinformation, and the more likely they would choose a similar answer on later tests.

In a similar study, Butler and Roediger (2008) also manipulated the number of response alternatives in preliminary multiple choice tests assessing college student's ability to retain information given to them in 12 prose passages covering historical topics. One of 3 variables manipulated were the number of response options available which ranged from two, four, and six alternatives. Results showed a main effect of number of alternative choice alternatives, showing that performance on the preliminary test decreased as the number of alternatives increased. They also found that less prior study and increased response alternatives in a preliminary multiple choice test resulted in a higher proportion of 'intrusions' or incorrect responses on the second test resulting from the incorrect lures from the test taken prior.

Evidence suggests that in order to minimize the possibility of students acquiring misinformation in multiple-choice testing, increase/reinforce correct responding, efficiently capitalize on the testing effect, and use testing as another method of teaching, feedback is necessary. As mentioned earlier by Kulhavy and Stock (1989) and Shute (2008) feedback can verify whether an answer is correct as well as provide information that can lead the learner to the correct response. If multiple choice testing is as prevalent as research suggests and has known drawbacks such as the introduction of misinformation which can lead to incorrect responding on later cumulative tests, then perhaps feedback after testing can minimize the negative effects while increasing the positives of this form of assessment. Also, it is likely that feedback may prevent the negative suggestion effect by reinforcing or punishing responses on the preliminary test depending on whether it is correct or incorrect.

Feedback with Multiple Choice testing

In addition to manipulating the amount of study time participants had to review 12 prose passages (no study, read, read plus key sentences) and the number of alternative answers (2, 4, or 6), Butler and Roediger (2008) also manipulated type of feedback students received after a preliminary multiple choice test (no feedback, immediate feedback, delayed feedback). In the no study condition participants were kept busy with a filler task. In the study condition participants had 30 minutes to read the 12 passages and in the read plus key sentences conditions participants read the material for 30 minutes followed by reviewing key sentences from each passage that were directly related to the test. After a filler task all participants took a 36 question multiple choice test on individual computers. Feedback on the correct answer was provided for 10 seconds in the feedback conditions either directly after each individual question (immediate) or after the test was completed (delayed). Participants in the no feedback conditions received a message informing them to wait patiently for the next question to load in order to keep the time spent on each question equivalent. Feedback presented contained information on whether the response chosen was correct (verification), the correct response, as well as a re-presentation of the question.

One week later participants returned to take a cumulative cued-recall test containing the same questions on the multiple choose test as well as additional ones not previously administered by writing out the answers from memory (as opposed to multiple choice). The first section of the cumulative exam consisted of a forced-response phase where students answered each question then had a free report phase where participants were able to go back through their answers and omit any they may have guessed on. Therefore, any responses that students may have guessed or were not sure on could be removed, creating a separate dependent variable that provided information on how often groups guessed. In accordance with past findings the results of the second cued recall test showed a testing effect in that participants performed significantly better on the questions they encountered on the previous multiple-choice test compared to their performance on novel questions, no matter what study or feedback condition. In addition, when participants were tested and no feedback was provided, performance on the subsequent test was highly dependent on performance on the initial MC test. As expected those who studied more led to a higher performance on the cued recall exam whereas the number of alternatives did not significantly influence performance. Most pertinent to the present research as well as Butler and Roediger's (2008) main hypotheses, feedback on the preliminary multiple choice test resulted in higher performance and retention on the later test. Results showed that participants in the feedback conditions had a higher proportion of correct responses whereas the proportion of intrusions (i.e. number of incorrect responses that matched previous incorrect response) was significantly lower than the no feedback condition. Further, although both feedback conditions led to comparable reduction in the amount of incorrect responses due to misinformation, the delayed feedback condition led to a greater proportion of correct responses than the immediate feedback condition.

As to be expected, with the addition of feedback to formative multiple choice assessments, the positive effect of prior testing on later test performance increased substantially above that of participants who were provided no feedback during prior testing. Additionally, negative effects of being exposed to multiple incorrect alternatives were minimized in conditions who took a practice multiple-choice test that provided feedback. In other words, feedback during practice testing resulted in greater subsequent test performance than participants just being tested prior without out feedback.

IFAT

If feedback does effectively increase retention as well as minimize misinformation due to prior multiple-choice testing, testing formats that incorporate feedback and can be made available for use as a type of formative assessment would be worthwhile. One such format growing in popularity in the academic setting is the Immediate Feedback Assessment Technique (IF AT) developed by Epstein, Epstein, and Brosvic (2001). The IF AT is a commercially sold feedback format similar to a scantron that allows the user to answer until he or she chooses the correct alternative. The IF AT answer form contains a number of rectangular spaces or response options for each question on a multiple choice test. These rectangular spaces are covered with a waxy substance that blocks the test taker from seeing either a blank box or symbol underneath similar to a lottery scratch-off ticket. When the test taker chooses a response alternative on a multiple choice quiz they would scratch off the film on the corresponding box on the answer form revealing a symbol if correct or a blank box if incorrect. If a blank box is revealed the test taker is permitted to continue selecting/scratching other alternatives until the symbol/correct option is revealed. The added benefit of this new format is its utility as a formative assessment tool as opposed to assessment for grading purposes alone.

In their original study, Epstein, Epstein, and Brosvic (2001) tested the IF AT by using the format in one section of a two section Introductory Psychology course. Throughout the semester both sections took 4 exams on course material and differed only in the testing format used. One section used the typical scantron format without feedback when selecting answers for each of the tests and the participants in the IF AT condition were only given credit for their first response. At the end of the semester a cumulative exam containing 3 questions from each of the previous exams plus 38 novel questions were used to measure retention of the course material. The results of the study showed that although sections did not differ on scores for any of the 4 prior examinations, IF AT users scored significantly higher on identical questions previously tested on the cumulative exam. In this study, the scantron condition tests served as summative assessments where only the grades were considered. But for the IFAT condition, although only first responses counted toward their graded performance, the testing format allowed them to continue responding to receive feedback on the correct answer if their original choice was not correct. This assessment format is both summative and formative in that the feedback received should enhance learning and possible future performance on the final cumulative exam.

In later tests of the IF AT's utility in providing feedback in multiple-choice testing, Epstein, Lazarus, Calvano, Matthews, Hendel, Epstein, and Brosvic (2002) found similar results confirming that the IF AT does improve test performance and increases correct answers previously answered incorrectly in comparison to participants taking prior examinations using the typical scantron sheet without feedback. These studies again emphasize the importance of some form of formative feedback following typical testing in academic settings that verifies correct responses, provides information regarding incorrect responses, as well as providing elaboration on what the correct answers are if previously answered incorrectly.

Also using an IFAT condition, DiBattista, Gosse, Sinnige-Egger, Candale, and Sargeson (2009) were interested in how testing scheme influenced performance on a later test. The different testing schemes included: a) feedback on number correct (NC), b) correction-for-guessing (CG), and partial credit (PC). Researchers also manipulated test difficulty (low, medium, or high difficulty) influenced test performance on a later test. The number correct scheme only scores correct responses so students were not penalized for guessing or omitting answers. The CG scheme uses an equation to minimize the inflation of correct responses that may have resulted from guessing. Generally speaking, depending on how many questions scored correct and incorrect on a multiple-choice test and the number of alternative response options for each question, it can be assumed that a portion of the correct answers were due to guessing. The partial credit test scheme provides points for correct responses that were not found on the initial attempt. Participants at each level of test difficulty were randomly assigned to a testing scheme where approximately equal numbers of participants separated into 9 groups took the IFAT for the first test. As a control a tenth group took the test in a regular scantron format for both tests. Like other research testing the influence of prior testing and feedback, subsequent scores on a later test showed that IF AT users test scores increased significantly on the second test, while users that used the typical testing form without feedback did not have significant gains in learning. Also, they found out of those conditions using the IF AT, the magnitude of increase in scores increased with the difficulty of the test in both the NC and PC conditions whereas the CG scheme seemed to interfere with later test performance. Judging from the results of studies comparing testing followed by feedback to testing without feedback, it is apparent that in order to enhance and improve overall retention, feedback must be included if testing is to be used as formative assessment technique and that the IFAT is a viable method of doing so.

Immediacy of Feedback

Amongst Shute's (2008) many prescriptions presented above for proper formative feedback such as the necessity for clear, concise, relatively elaborate feedback that is presented in manageable units, guidelines presented related to the timing of that feedback were not yet discussed in the present literature review. Some considerations mentioned are that for difficult tasks and retention of procedural or conceptual knowledge, immediate feedback should be used. But for relatively simple tasks or learning that requires a transfer of task performance, delayed feedback has been shown to be more effective. Daniels and Daniels (2006) suggest that feedback should be provided as soon as possible after a response in order to increase the likelihood of acquisition. But this immediacy, typically referring to behaviors in performance management and OBM settings, refers to immediacy in hourly, daily, weekly, monthly units where the sooner the feedback, the more opportunities for future feedback. Also, based on theories of reinforcement, feedback is suggested to be given as soon as possible after an incorrect response has occurred. This serves to increase contiguity and association of the feedback with the response in order to decrease the likelihood of the same mistake in the future. On the other hand, many researchers suggest that delayed feedback might be more effective because it allows for original incorrect responses to be forgotten or dissipate

(Kulhavy & Stock, 1989). The interference-perseveration theory suggests that immediate feedback produces response competition when the correct response is immediately presented after the incorrect response, resulting in a possible confusion when associating the test question with the correct response option. The amount of delay required for the incorrect response to dissipate is not immediately obvious; what is delayed in one case may be considered immediate in others. But perhaps the immediate presentation of the correct stimuli after an incorrect response creates competition between the two stimuli during the formative test which would then interfere with correct recall when presented with similar scenario during the second test. In the Butler & Roediger (2008) study they found that delayed feedback led to a higher proportion of correct responses on subsequent tests (M-.56) than immediate (M=.45). A similar theory posits that the delayed spacing of time might allow for better retention of the feedback, synonymous with distributed study over mass study on retention (this is typically only found when first answer is correct).

Classroom vs. Lab findings

Several studies have found that the relationship between retention and the type of feedback received in prior multiple choice testing depends on whether the study was done in a classroom or in a more lab based setting. There is likely an interaction between the experimental setting and type of feedback participants/students receive during testing. In a meta-analysis of feedback timing Kulik and Kulik (1988) as well as Butler and Karpicke (2008) stated that benefits of delayed feedback are typically found in more lab based studies whereas findings in actual classroom settings typically favor immediate feedback. One theory on this interaction is that retention rates may differ as a function of the amount of time students spend reviewing the feedback. In a typical classroom setting students may breeze over or ignore any feedback received due to any number of factors including busy schedule, discounting of seeing material again, or social distractions as opposed to a lab based setting where they are instructed and allotted time to do so and the previously mentioned factors are not available. This explanation would not be a surprise considering the purpose of having lab based studies is to avoid extraneous variables and have better control over variables of interest. In the classroom setting, multiple contingencies are competing for a student's attention that might control behavior, thereby making feedback less of a priority and diluting the effectiveness of delayed feedback in applied settings.

In Butler, Karpicke and Roediger's (2007) first study they used a two (immediate vs. delayed feedback) x two (type- standard first response vs. answer-till-correct) design where participants had to read 12 passages then take a 40 question multiple choice test; eight questions with no feedback, 16 with standard, 16 with answer till correct, eight were a part of the no test condition that would be later found in the later test. Immediate feedback was provided after the incorrect or correct answer was chosen unless it was an answer until correct question. Delayed feedback was given after a 10 minute time filler then feedback was given in similar fashion. The following day participants returned to complete a cued-recall test. Results showed a significant increase in retention in both feedback conditions above the no feedback condition but no effect found for type of feedback.

The methods of Butler et al.'s (2007) second experiment were similar but the delay between the first and second testing session was extended from 48 hours to a week in order to test whether the effects of the delayed feedback weren't noticeable until an extended amount of time has passed. Again there was no effect of type of feedback, suggesting that the standard feedback was just as effective as the answer-till-correct method (similar to the IFAT). The authors suggested that this may be a result of how controlled the experiment was, not having much generalizability to the typical classroom. They suggested that perhaps in a more applied setting, the answer-until-correct format will show its merit because it forces students to review correct and incorrect responses, something students may not do with regular feedback. Finally, another possibility discussed for why the type of feedback wasn't significant is that the increased chances to respond may hinder performance by increasing the amount of incorrect lures. As discussed before, the interference-perseveration theory would argue that responding multiple times to a single question immediately after each other may cause confusion when later presented with similar questions. Answer-until-correct (along with IFAT format) allows you to choose multiple answers that could later be confused, resulting in continued errors or less variation in findings between groups.

However, Butler et al. (2007) did find that participants who received feedback 10 minutes after a formative exam retained more information than immediate feedback participants who received feedback following every question. It was also found that the longer the time between the first multiple-choice test and the second the better the delayed feedback conditions performed on the cued- recall test, suggesting the benefits of delayed feedback might depend on how much time has elapsed between feedback and the following assessment. The positive effects of delayed feedback may not be fully known until a certain amount of time has passed. These results also support the interference-perseverance theory and replicate the general findings that in lab-based settings, delayed feedback seems to be more beneficial in learning and retention.

Brosvic, Dihoff, Epstein, and Cook (2006) compared immediate and delayed forms of feedback in a study geared toward assisting elementary school students categorized as normally achieving or those with mathematics learning disabilities in acquiring math operations such as addition, division, and multiplication. Participants were assigned to one of three feedback conditions, (delayed feedback, IF AT, educator) or the control condition for one of the arithmetic operations. The experimental phase consisted of 20 sessions of one presentation of the fact series (0 to 9 consisting of addition, division, multiplication, and subtraction) provided using multiple choice questions shown on flash cards and recorded by participants with either a scantron or IF AT form. Following the 20 sessions the maintenance phase consisted of 5 sessions where all participants recorded answers to the math problems on a scantron format without any form of feedback. During the experimental phase of the delayed feedback condition after problems were completed participants were given 30 minutes to review their responses and the correct answers. In the experimental phase of the educator feedback condition, answers were recorded on scantron form and verbal feedback was provided by the educator. During this condition the educator provided reinforcement for correct responses and if incorrect gave a verbal prompt to try another response. The maximum number of attempts allowed was comparable to the amount of opportunities available on the IFAT. Like other IFAT conditions reinforcement was provided if a star was revealed and were allowed to continue answering until the correct response was found. The results of this preliminary study showed that both immediate feedback from the IF AT and the educator resulted in comparable acquisition of the fact series above that of the control and delayed feedback conditions during the maintenance phase.

In study 2, only those participants originally in the control and delayed feedback conditions were participants. In this study half of the remaining participants were assigned to the aforementioned conditions while the remaining half completed the calculations in one of the two immediate feedback conditions (IF AT or educator). During the second intervention, the performance of participants who had been switched to the immediate feedback conditions improved significantly and did not significantly differ in performance from those participants originally under the immediate feedback conditions.

Overall, the Brosvic et al. (2006) results demonstrated that immediate and corrective feedback assists in the learning of the four arithmetic operations above that of the control and delayed feedback conditions. Also, the modality of the feedback, whether it was through the IF AT or an educator, did not differentially influence acquisition. Another interesting finding was that in studies 1 and 2, delayed feedback received at the end of each session resulted in retention rates that did not significantly differ from the control (no feedback). These findings suggest a clear efficacy in utilizing immediate feedback when teaching mathematics to elementary school students with special needs as well as again the utility of the IF AT as a response medium. These findings support more behavioral theories of feedback as being a reinforcer in that the presentation of immediate feedback shortly after a correct response, whether it be with a star or praise increases the likelihood of that response during later testing. The contiguity or closeness between choosing an answer and receiving consequences contingent upon that answer increases their association, thereby improving test performance and retention. On the other hand, since the settings were not in a typical classroom, these findings also go against the relatively ongoing trend in research suggesting delayed feedback typically outperforms immediate feedback in lab based settings.

In a similar study, Brosvic, Epstein, Dihoff and Cook (2006) attempted a lab based study that closely mirrored a typical classroom by manipulating the number of response options on 5 laboratory examinations as well as the timing of the feedback (no feedback control, an end-of-test delayed feedback, a 24 hour delayed feedback, immediate feedback/assistant, and immediate feedback using the IF AT response medium) in order to test the acquisition of Esperanto words. The procedure consisted of 7 one- hour sessions per examination that were a mixture of lectures on vocabulary words, individual, and group learning activities followed by an overall review of content and a 50 item examination. After the five examinations, participants completed a cumulative exam that included questions from each of the previous exams plus 50 new items. For the assistant facilitated feedback condition, after the participant made answered a question the assistant would hold up a flash card signaling if correct. If the original answer was incorrect the assistant would hold up a flashcard signaling the incorrect alternatives already answered and would allow for more responding until the correct answer was discovered. In the IF AT condition, participants were permitted to continue responding until the symbol signifying a correct response was revealed on the form. In the 24 hr delayed condition, participants wrote down the answers they selected and on the following day were able to review the correct answers, the exam, and their corrected answer sheets for 30 minutes. In the end-of-test delayed feedback condition participants were given 30 minutes at the end of the test to review the correct answers.

Participants took the final exam 1 week after the last laboratory examination, as well as 3 and 6 months later using a regular scantron form. The results of the study showed that in each of the three cumulative test conditions, immediate feedback groups retained significantly more information than the delayed and control conditions. These results go against previous research and perhaps common 'belief' that delayed feedback is more effective in controlled laboratory conditions while immediate feedback is more useful in more applied classroom settings. It is clear that the general findings of these researchers tend to conflict with the lab based findings of other researchers (Butler, Karpicke, & Roediger (2007); Butler & Roediger (2008)) in which the slightly more delayed form of feedback in formative testing is most beneficial to retention.

In a recent study, Smiley's (2011, still in press) research evaluated the IF AT by testing students randomly selected from a university participant pool on retention of an introductory psychology textbook chapter. Participants were separated into 4 feedback conditions that consisted of a no feedback, scantron with feedback, IF AT, or Computerized IF AT conditions. In the Scantron with feedback condition participants took a test with a typical scantron form and students were provided with a sheet of paper with the correct answers to be reviewed after the quiz (delayed feedback). The IF AT condition was similar to other studies where students could answer until correct but only the original response counted. Finally the Computerized IF AT condition was similar to a typical IF AT but students used a mouse to uncover or 'scratch' the correct alternative on the computer screen. One week later participants in each condition took a test containing identical and comparable questions using a scantron form.

The results of the study, like similar studies, showed participants of the three feedback conditions outperforming the scantron without feedback condition. Also there was no significant difference in retention rates between the IF AT and CIF AT. Finally, the scantron with feedback condition resulted in the highest retention of the chapter material of all the conditions. The delayed feedback in the lab setting resulted in higher retention of the material for both the comparable and identical multiple choice questions than either of the immediate (IF AT) feedback conditions. Also, performances of participants in the scantron with feedback condition improved significantly during the post- test while the two IFAT conditions only improved significantly on identical questions. The no feedback group did not improve significantly on comparable or identical questions. The main results of Smiley (2001) also align more with the typical finding that delayed feedback has been shown to be more effective in more controlled lab settings, possibly due to what has been called the interference-perseverance theory, where due to the contiguity of the incorrect and correct feedback found in the IFAT, a failure to discriminate between similar response options on a later assessment results in an inability to choose the reinforcing response option, thereby resulting in poorer performance in long-term retention.

In the above literature review I have cited that multiple choice testing is the most prevalent form of testing in today’s academic environment. I have shown that the testing effect is a highly reliable finding in the literature. I have also presented studies suggesting the utility of providing feedback during testing in order to capitalize and maximize the positive effects of the testing effect as well as minimize the negative effects of multiple-choice testing such as the presentation of misinformation. Finally, I have reviewed findings showing the effectiveness of the IFAT at providing feedback, as well as a testing format that can facilitate learning and increase retention rates above that of the typical scantron forms without feedback. But researchers disagree as to whether the IF AT is any more effective of a formative feedback tool than simply using a typical scantron format in conjunction with delayed feedback and if this effectiveness only pertains to identical material. Past research has touched on a possible discrepancy in the feedback literature regarding the immediacy of feedback most influential at increasing later retention of class material. One rationale for why this discrepancy may exist is that the effectiveness of immediate versus delayed feedback after testing may depend on the setting of the research and whether the testing environment is done in a lab or an actual classroom.

Current Study

The present study attempted to measure how differences in pre-test format, either taken with scantron with feedback or the IFAT, influenced later test performance on comparable and identical questions in a true classroom setting by alternating the type of pre-test format used on each exam throughout the semester.

II. Method

Participants

Participants were sophomore and junior undergraduate psychology students enrolled in one of two separate psychology courses at a large, southern public university. The first class was an introductory measurement and statistics (psyc 210) course consisting of 43 undergraduates. The second course consisted of 33 people enrolled in a research methods (psyc 211) class.

Introductory measurement and statistics (psyc 210). Psyc 210 is meant to provide new psychology students with an overview of descriptive and inferential statistics used in the social sciences as well as provide students with the ability to demonstrate basic research skills, use computer software package SPSS, interpret research findings, and think critically. The course consisted of 5 exams, one approximately every 3 weeks, each worth 90 points (out of a total 1000) that included manly multiple-choice and short answer questions. At the end of the course there was an in-class cumulative exam.

Research methods (psyc 211). This course, which is designed to be taken following completion of the Introductory Measurement and Statistics course, is meant to provide students with an understanding of the scientific methods used in psychology as well as acquaint students with conducting and presenting research in APA format. There were three exams approximately every four weeks worth 100 points each (out of a total 1000) that included multiple-choice, fill-in-the-blank, and short-essay questions. There was also a cumulative final exam at the end of the semester.

Materials

For both courses I introduced two different formative practice test formats: a) scantron form with feedback and b) the IFAT form.

Scantron with feedback. The first format was the typical scantron form with delayed feedback presented after completion of the quiz. The students filled in a circle that corresponded to the answer they selected as correct. Scantron sheets had five options from which to choose from, A through E; however, the pretests only included four response-options to match the IF-AT form. Students completed a 16-item quiz, where they specified the answer on the scantron as well as on the quiz itself in order to keep track of the answer they selected. At the completion of the quiz, the instructor distributed the answer key to the students.

IF-AT. The IF-AT response format was similar to the scantron such that it required the students to select the appropriate letter that corresponded to the correct answer on the quiz (i.e., choices A through D/E). However, the IF-AT required participants to scratch off a waxy covering, similar to a scratch off ticket (the IF-AT form has four options), in order to select an answer. If the selected option was correct, scratching off the waxy covering revealed a star, immediately indicating that the correct option was selected. If no star was revealed, the participants continued to scratch off the next best option until the correct answer and star were selected (see Appendix A).

Procedure

General Procedure. Prior to the beginning of the semester participants were randomly assigned to one of two groups for each class using an identification number and remained in those groups for each pre-test throughout the semester. In the class period prior to each test, both classes were given a 16 question pre-test based on class material.

The formative pre-test, either IF-AT or scantron with feedback, was alternated between groups (i.e. Group 1 took quiz one using the IF-AT whereas Group 2 took the same quiz using the scantron with feedback format: for the second quiz the groups switched response formats, and so forth) for the tests throughout the semester (see Appendix S). This alternating format was the same in each course. There were five pre-/post-tests for the introductory statistics course and three pre/post-tests for the research methods course.

For both courses, eight identical and eight comparable questions testing the same material from the pre-tests were found on the corresponding post-test the following class period. Identical questions were the same exact multiple-choice question. Comparable questions covered the same material but the format of the question was altered to be either another multiple-choice question using a different example or a short answer question (see Appendix C).

At the end of the course students filled out an informed consent asking for permission to analyze their aggregate data as well as a demographics questionnaire.

Instructions. On the day of the formative pre-test students were separated according to which group they were randomly assigned. A test packet containing 16 questions was provided to both groups. One group was provided a scantron answer form whereas the other group received the IF-AT form. An undergraduate or graduate teaching assistant gave instructions for using each format and told participants to begin testing. Participants in the scantron condition were informed to write their answers on the test packet in order to later compare their answers to the correct ones provided on a key distributed upon completion of the pre-test. When finished answering all 16 questions students using the scantron form traded their scantron in for an answer key whereas those using the IFAT form were told to review their answers and turn in their materials when finished.

Dependent Variable

Mean performance on formative pre-tests and summative post-tests was the primary dependent variable and was defined by the percentage of questions answered correctly. More specifically, the percentage of correct first responses on identical and comparable questions for both the pre-tests and the post-tests was calculated. Also, for the introductory statistics course, mean performance on the portion of the cumulative final identical and comparable to the four prior formative tests was also calculated.

III. Results

Demographic analyses

Prior to running the primary analyses, differences between groups 1 and 2 on demographics for each psychology course were compared. Groups did not significantly differ on GPA, number of psychology courses completed and currently taking, any of these variables (*p* > .05).

Assumptions

The data for nine test sections (statistics course: five summative exams and one cumulative final; research methods: three summative exams) were analyzed using a 2 (test format: scantron with feedback and IFAT) x 2 (question type: identical and comparable) x 2 (time: pre- and post-test) mixed ANOVA. Prior to running the ANOVAs analysis of assumptions were tested. Levene's tests assessing the homogeneity of variances between groups were not significant (*p* > .05). After exclusion of outliers skewness was within acceptable bounds suggesting the data were approximately normal.

Main Effects

The results of the nine mixed ANOVAs showed one consistent main effect of time. Seven of the overall nine ANOVAs showed a significant increase in mean test percentage from pre- to post-test (*ps* < .05 level), whereas an eighth test approached significance (*p* = .053). For both the Research methods and Introductory Statistics course, no main effects of quiz format were significant (*ps* > .05 level). See Figures 1 and 2.

Interactions

Results showed 2-way interactions between time and question type across four of the test sections as well as interactions between time and quiz format in two of the test sections (described below). None of the test format x question type x time 3-way interactions were significant.

Question Type by Time Interactions. Significant 2-way interactions between time and question type were found for the first (*F*[1,34] = 14.79, *p* = .001, partial η^2 = .301), second (*F*[1,33] = 35.75, *p* < .001, partial η^2 = .52) and cumulative final exam (*F*[1, 256] = 68.33, *p* < .001, partial η^2 = .209) of the statistics course as well as for the first exam (*F*[1,26] = 16.46, *p* < .001, partial η^2 = .039) of the research methods course, suggesting the increase in test performance from time one to time two depended on whether the questions on the post-test were identical or comparable to questions on the pre-test. More specifically, across all four of the question type by time interactions, mean percentage score on identical questions increased significantly more between time 1 and 2 than comparable questions. For the first statistics testing section, performance on identical items significantly increased from pre- (*M = 58%, SD = 32%)* to post-test (*M* = 80%, *SD* = 29%), *t*(34) = -4.5, *p* < .001, *d* = -.72, while there was no significant change in performances on comparable items. For test 2 material, although performance on comparable items did increase between formative pre- (*M* =65%, *24*%) to summative post-test (*M* = 73%, *SD* = 19), performance on identical questions from practice (*M* = 56%, *SD* = 16%) increased to a much higher degree (*M* = 85%, *SD* = 16.6%), *t*(33) = 6.99, *p* < .001, *d* = -1.78. For the statistics cumulative final exam, not only did performance on identical questions increase from pre-test to post-test, but performance on identical questions (*M* = 62%, *SD* = 25.5%) originally significantly lower than performance on comparable questions (*M* = 71.35%, *SD* = 23%) at time 1 increased to be significantly higher (*M* = 86.7%, *SD* = 23) than performance on comparable questions (*M*

=76%, *SD* = 23.4%) at time 2, *t*(258) = 4.14, *p* < .001, *d* = .43. Finally, for the research methods test 1 material, performance on identical items at pre-test (*M* = 85%, *SD* = 13.9%) increased significantly at post-test (*M* = 94.9%, *SD* = 8.7%), *t*(26) = -3.2, *p* < .003, *d* = -85, where as there was no significant change in performance for comparable items. Figure 1 and 2 depict percentage of correct items by item type across time for both the IFAT and scantron with feedback conditions.

Further, contrasts showed that performance on identical questions for 3 of the 4 testing sections (Statistics test 1: *M*=58%, *SD* = 32%; test 2: *M* =56%, *SD* = 16%; Cum. Final: *M* = 62%, *SD* = 26%; Research methods test 1: *M* = 85%, *SD* = 14%) were significantly lower than performance on comparable questions (Statistics test 1: *M*=78%, *SD* = 16%; test 2: *M* =65%, *SD* = 24%; Cum. Final: *M* = 71%, *SD* = 23%; Research methods test 1: *M* = 95%, *SD* = 9%) at time 1, indicating that questions selected to be identical may have been more difficult. For three out of the four interactions, performance at time 1 on identical questions were significantly lower than performance on comparable questions at time 1 (*p* < .05). At time 2, average performance on the identical questions (Statistics test 1: *M*=80%, *SD* = 29%; test 2: *M* =85%, *SD* = 16.6%; Cum. Final: *M* = 86%, *SD* = 18%; Research methods test 1: *M* = 95%, *SD* = 9%) increased to levels comparable or greater than comparable items (Statistics test 1: *M* = 77%, *SD* = 19%; test 2: *M* = 73%, *SD* = 19%; Cum. Final: *M* = 76%, *SD* = 23%; Research methods test 1: *M* = 88%, *SD* = 11%). All significant interactions of this kind show the same pattern as results displayed in Figure 3.

Quiz Format by Time Interactions. There were significant 2-way interactions between time and quiz format for statistics test 5 material (*F*[1, 29] = 6.73, *p* < .015, partial η^2 = .2) and the statistics cumulative final exam (*F*[1,256] = 7.03, *p* < .009, partial η^2 = .027). Figure 3 shows a separate graph comparing differences in mean percent of correct pre- and post-items by testing formats. For statistics test five material, mean percent correct for those students using the IFAT at time 1 (*M* = 72%, *SD* = 17%) was significantly less than those students using the scantron with feedback (*M* = 85%, *SD* = 16%), *t (27)* = -2.26, *p* = .032, *d* = -.79. At time 2, performance for those participants who took the pre-test with the IFAT increased significantly (*M* = 88%, *SD* = 13%) resulting in mean post-test performance comparable to the scantron with feedback group (*M* = 89%, *SD* = 11%), *t*(14) = -4.896, *p* < .001, *d* = -1.06. A similar trend occurred for the final exam, where mean correct performance on IFAT and scantron with feedback were cumulated across groups. At pre-test (time 1), mean performance for participants using the IFAT (*M* = 63%, *SD* = 25%) were significantly less than mean performance on items taken using the scantron with feedback (*M* = 70%, *SD* = 24%), *t*(258) = -2.275, *p* = .024, *d* = -28.6. But, at time 2, although both performances on items previously taken using the IFAT (*M* = 83%, *SD* = 22%) and scantron with feedback (*M* = 80%, *SD* = 22%) significantly increased, performance on questions previously seen on the IFAT increased by twice as many percentage points, *t*(129) = 9.62, *p* < .001, *d* = -.85 (See Figure 4). This increase resulted in mean performance on items previously seen using the IFAT to be 3 percentage points higher than performance in items previously experienced with the scantron with feedback condition, although this difference was not statistically significant.

IV. Discussion

The results of the present study showed only one consistent finding across the majority of the analyses: Test performance on the summative post-tests was significantly higher than performance on the formative pre-tests. These findings suggest that in general, feedback provided by both the IFAT and scantron with feedback may have contributed to increased performance on post-tests. This finding is partially incongruent with results of multiple researchers comparing the IFAT and scantron with feedback test formats (Brosvic et al., 2006; Epstein, et al., 2002; Butler & Karpicke, 2008; Kulik & Kulik, 1988), arguing immediate feedback, especially in applied settings, results in higher retention of test material. Although, for two of the analyses differences in later test performance depending on the type of formative assessment participants used did occur, these findings were not consistent across the majority of tests and may be a function of the particular test material. These specific analyses are discussed below. Also, the majority of analyses (8 of 9) showed main effects of time, but only 4 of these analyses did not show interactions between time and other variables requiring further interpretation.

Question Type and Time Interaction

Although less consistent across the nine different testing sections analyzed, for four of the test sections, the increase in test performance from formative pre-tests to summative post-tests depended on the type of question asked. More specifically, even though mean performance on identical questions was lower than comparable questions at pre-test, performance on items identical to what students had already seen increased significantly on the corresponding post-test whereas performance on comparable questions did not significantly increase. These findings partially replicate those of Smiley et al.’s (2011) research showing that participants who took the formative pre-test with the IFAT only significantly increased their performance on identical items. But unlike Smiley et al. the current study did not find that the effect of item type on test performance depended on which type of formative assessment they used previously. No matter what type of formative assessment used, on these four tests, students performed better on the items they had already seen.

This result is most likely due to what common sense would tell us: identical questions are simply that, identical. But what is the mechanism that causes identical questions to better occasion the appropriate response after the passage of time? The exact test question probably still has strong stimulus control over the behavior of choosing the response previously reinforced (correlated with reinforcement) considering the characteristics of the two stimuli are the exact same. In the presence of identical stimuli, the response previously paired with positive consequences such as ‘being correct’ or receiving a ‘star’ is much more likely.

On the other hand, with comparable questions, because they were similar but not identical in form, a student would need to generalize across questions that vary slightly in form, but still be able to differentiate between stimuli ‘meant’ to occasion completely different responses (so separate stimulus classes). To do this, participants would have to receive reinforcement for responding in the presence of approximations of the stimulus (question), which would require much more practice or trials. A student would need to have a more general understanding of that question as a result of experience with multiple similar questions in the presence of which a particular response was reinforced.

Compared to an identical question where in the presence of the same stimulus, the same response will result in reinforcement. Because of this extra experience required to generalize across questions, the questions have not been equated, resulting in the questions setting the occasion for different or incorrect responses. So byy definition, if two separate stimuli (questions) occasion separate responses, they have not entered into the same stimulus class. Another way of looking at it is that the lower performance on comparable questions reflects a low level of information processing in that students might only be attending to structural aspects of the question as opposed to a deeper understanding that requires a more thorough semantic analysis (Chew, 2004). Through this more rigorous analysis the student would notice the similarities with comparable questions experienced before and by responding similarly in the presence of the new question, would be showing stimulus generalization (Cooper, Heron, & Heward, 2007).

For the statistics course cumulative final, the significant increase in only identical questions was even more pronounced. This is most likely due to the issues of stimulus generalization discussed above, as well as the additional experience with the identical items at both the original formative and summative exams. With the additional experience with the item, participants essentially had twice as many training trials with identical questions than they had with the comparable items.

Test Format and Time Interaction

Even though not consistent across all of the analyses, results for the statistics test five and the cumulative final exam both showed an interaction between formative quiz format and time. For the statistics test five material students who used the IFAT format for the formative pre-test significantly increased their scores on the summative post-test whereas students who used the scantron with feedback did not see the same significant improvement. Similarly, for the cumulative final, exam performance on items previously taken using both the IFAT and scantron with feedback was significantly higher than formative pre-test scores, but the magnitude of that increase was much larger for items taken using the IFAT. Interestingly enough, for both testing sections, pre-test performances for participants using the IFAT were significantly lower than participants using the scantron with feedback.

Formative Performance. At first glance results for the statistics test five material and the cumulative exam for the introductory statistics course show some support for the utility of the IFAT as a formative assessment tool over other methods like the scantron with feedback. But first, the issue of why pre-test performance using the IFAT was significantly worse than those using the scantron with feedback should be addressed. For the cumulative final material, when aggregating the pre-test performance of both groups depending on what type of format students used a clear difference in average performance emerged. When all participants used the IFAT on the formative exam they scored 63 percent of the questions correct whereas when they used the scantron with feedback they averaged 70 percent of the questions correctly. This difference is meaningful in that participants alternated test formats, so that each group used each format two times. There should be no other reason why there is a consistent trend to perform worse at pre-test while using the IFAT other than something specific to the format and the testing behavior it occasions.

One possible interpretation for differences in pre-test performance is that there was a lack of consequences for performing well on the formative pre-test along with the immediacy of feedback while using the IFAT. Even though there were also no real consequences for performing badly on the scantron with feedback condition either, because the feedback was presented after a slight delay, students chose not to rush through to answer the question. Also, due to a long history of testing with the scantron, students may realize the utility of taking their time. When using the IFAT, students may not choose to consider the material for long due to the fact that the answer is only a scratch away, perhaps showing impulsive behavior due to the immediacy and certainty of receiving corrective feedback. By the fifth test, students are probably realizing that there are no real negative consequences for performing badly on the preliminary practice quiz. No matter if their first or second response is incorrect, feedback will be provided promptly if they simply pick a best guess.

Another difference between the IFAT and scantron with feedback response formats is that the scantron format allows users to erase first responses. The IFAT on the other hand, once an answer has been ‘scratched,’ cannot be erased and is scored as incorrect. The overall trend across four out of the five formative quizzes for lower performance for which ever group used the IFAT could be a result of the inability to correct their first response. The efficacy of answer changing has been questioned for many years but consistent empirical evidence has shown that multiple-choice test takers are more likely to improve test scores by changing answers instead of going with their original response (Benjamin, Cavell, & Shallenberger, 1984; Waddel & Blankenship, 1994). It is possible that during the pre-test performances were systematically higher in the scantron condition simply because those students changed their first incorrect response to the correct one after considering the question further.

But the average performance on items previously seen using the IFAT did increase 20 percentage points whereas performance on items previously seen using the scantron with feedback condition increased by 10 percentage points. By comparing the change in performance as a result of which formative test format was used, one could argue that the IFAT lead to twice as much retention as the scantron with feedback. When considering the results in this fashion, the evidence would support past research suggesting that Immediate Feedback Assessment Technique increases retention of course material in applied settings better than assessment tools with more delayed forms of feedback (Epstein et al., 2002; Brosvic et al., 2006; Kulik & Kulik, 1988). Unfortunately, we do not know how much of the 20 percent increase in test performance for the IFAT condition is attributable to other factors like the lack of effort or inability to change answers during the formative assessment or interactions with specific test material. It is reasonable to ask that if performances on the formative pre-tests for each condition were equal, would performance on items using the IFAT still increase by twice as much?

Summative performance. For the statistics course test five material, there was no significant difference on the summative exam between participants who previously received feedback through the IFAT or scantron with feedback. Even though there was only a significant increase in test performance for those who used the IFAT, whether that increase was due to the immediate feedback is questionable. For the introductory statistics cumulative final exam material the average performance on items previously seen using the IFAT did not significantly differ from average performance on items experienced using the scantron with feedback. Mean percent correct on questions previously seen using the IFAT (83%) was only three percentage points higher than that of performance on questions previously seen using the scantron with feedback (80%). Because there was no difference on summative post-test performances between groups, we have to question whether there is any real practical difference between receiving formative feedback via the IFAT or the scantron with feedback. If students ultimately end up with the same cumulative final exam grades regardless of what formative test they studied with, the difference in increases in performance between pre- and post-tests might not matter.

Limitations and Future Directions

Although for this particular study the results showed overall significant increases in test performance across the nine sections of test material for two undergraduate courses, without a control condition that did not experience any formative practice test, we cannot fully rule out the possibility that the increase in performance was due solely to typical study behavior. As a result of the applied setting, limited class size, and ethical issues related to withholding study tools from a portion of the class, it was decided to only test for differences between the IFAT and scantron with feedback conditions. Having addressed those issues, it has been well established in the testing effect literature that being tested on course material leads to higher retention than comparable amounts of pure study (Carrier & Pashler, 1992; Bjork & Bjork, 1992). Considering this, I am relatively confident that if there were a control condition, summative test performance for those in that condition would be significantly less so than those receiving formative assessment. None-the-less, for the sake of increasing internal validity, future research in classroom settings should include a control if possible.

Another limitation of the present study is the discrepancy found in pre-test performance between IFAT and scantron with feedback conditions. Preferably, there would be no differences in test performance across test format at time one in order to better attribute any differences in test performance at time two to the differences in immediacy and feedback format. But from the results of the present study we may have stumbled upon possible issues regarding the efficacy of student performance using the IFAT when contingencies increasing the probability of sincere effort are not in place. The discrepancy of pre-test performance has also drawn attention to how the IFAT does not allow for a student to change answers after their first response and how this might influence performance on the IFAT if only first responses are being considered for grading purposes. Future research replicating this study should both apply appropriate contingencies for effortful performance on formative assessments as well as consider collecting data on how often people change answers while using the scantron in order to better assess what forms of formative assessment and feedback maximize student retention and academic performance.

Figure 1.

Percent of correct items by quiz format across statistic exam material

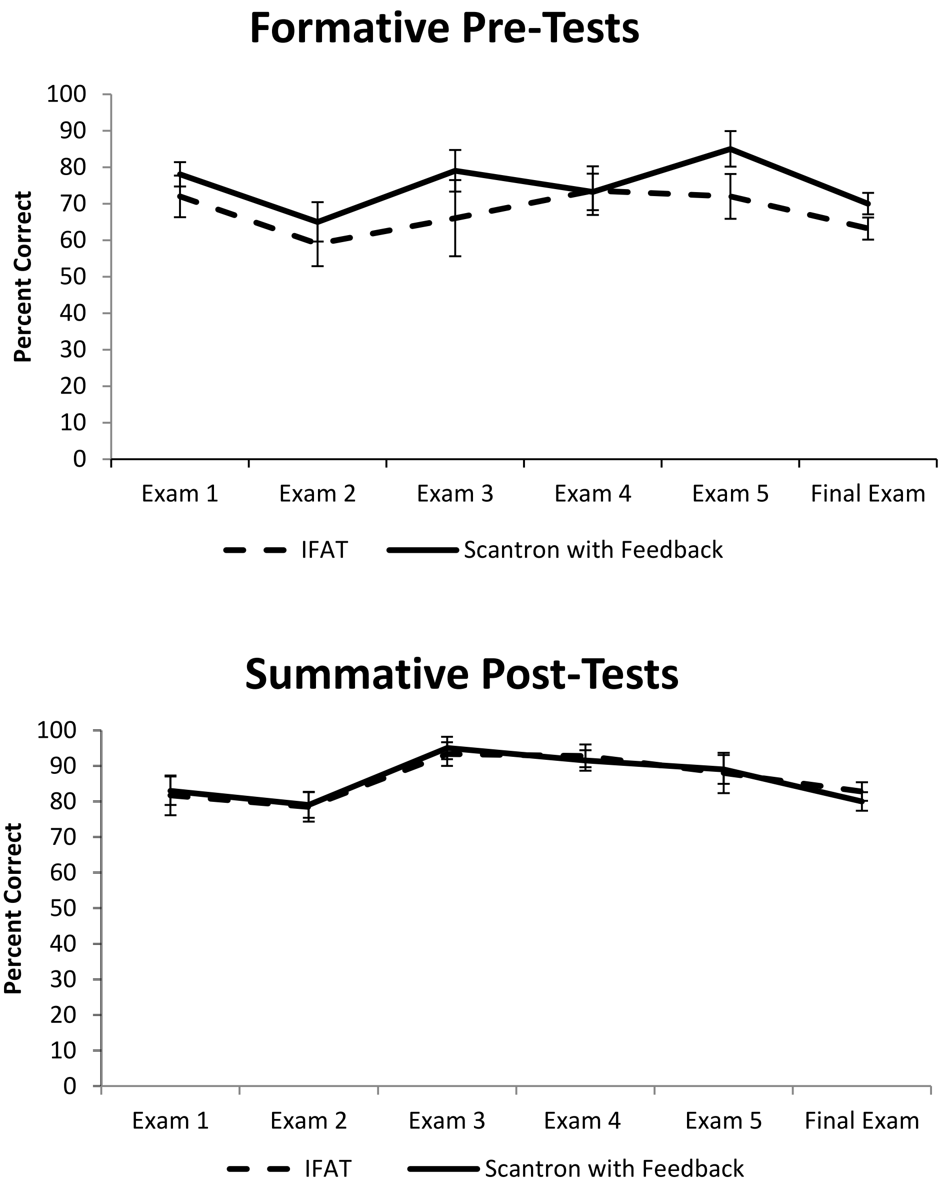
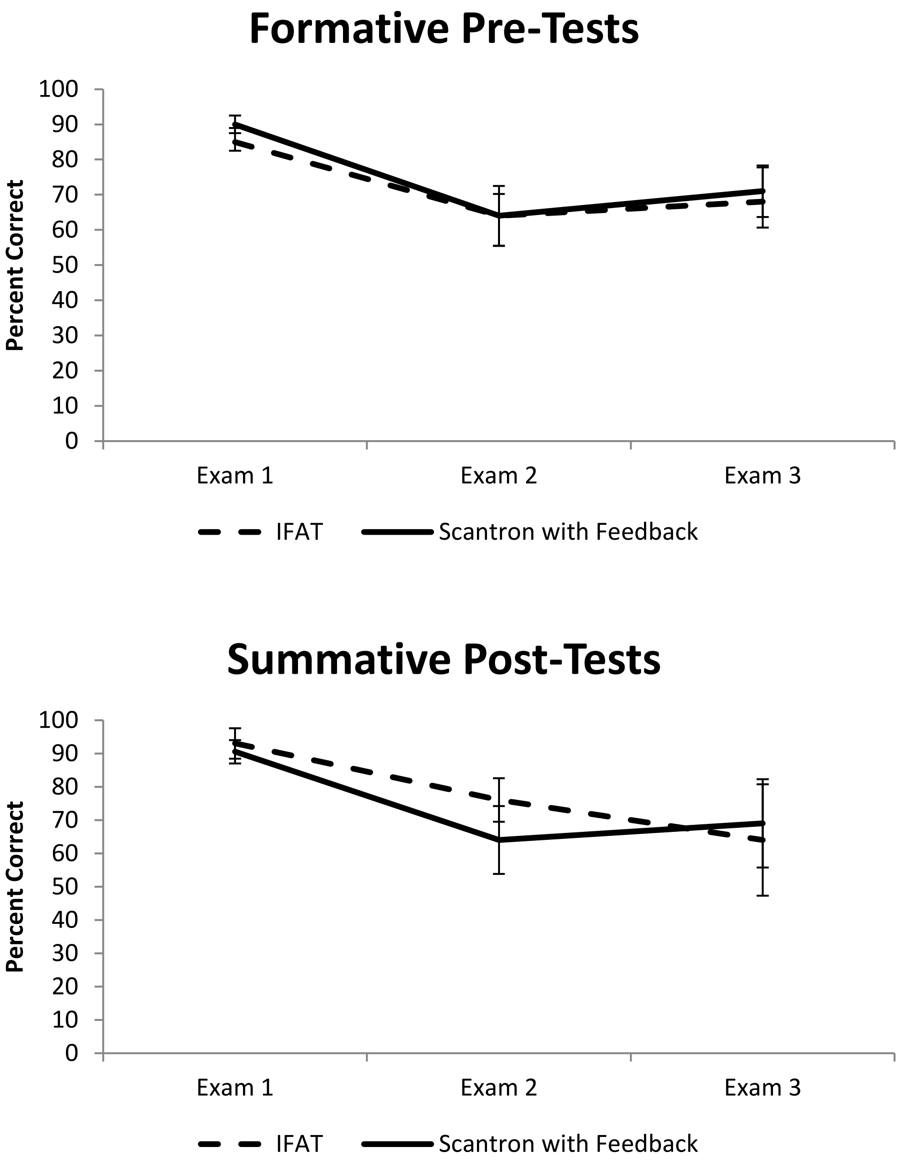


Figure 2.

Percent of correct items by quiz format across research methods exam material



Formative Pre-Tests

Summative Post-Tests

Figure 3.

Percent of correct items by item type on pre- and post-tests for statistics final exam material

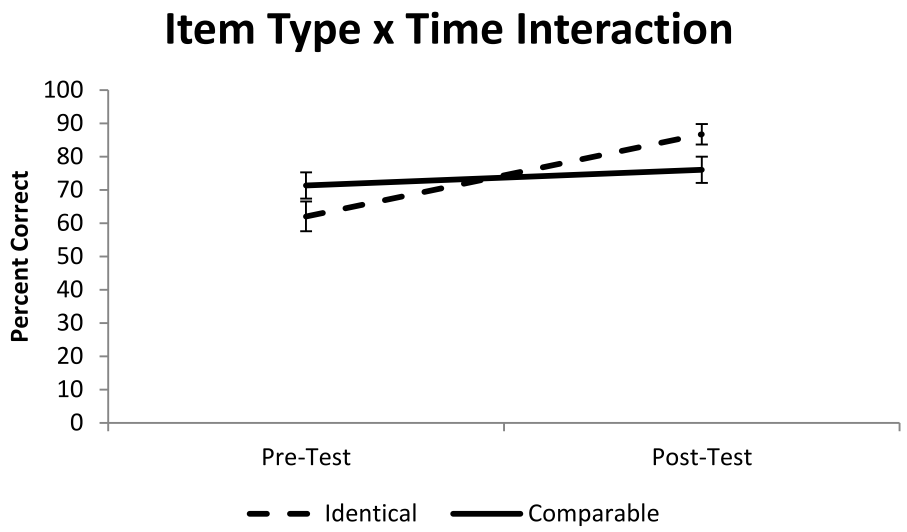
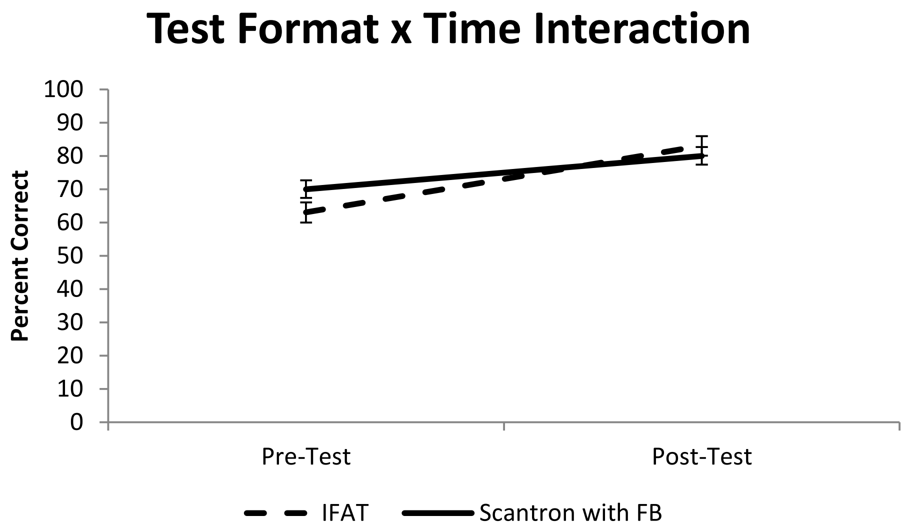


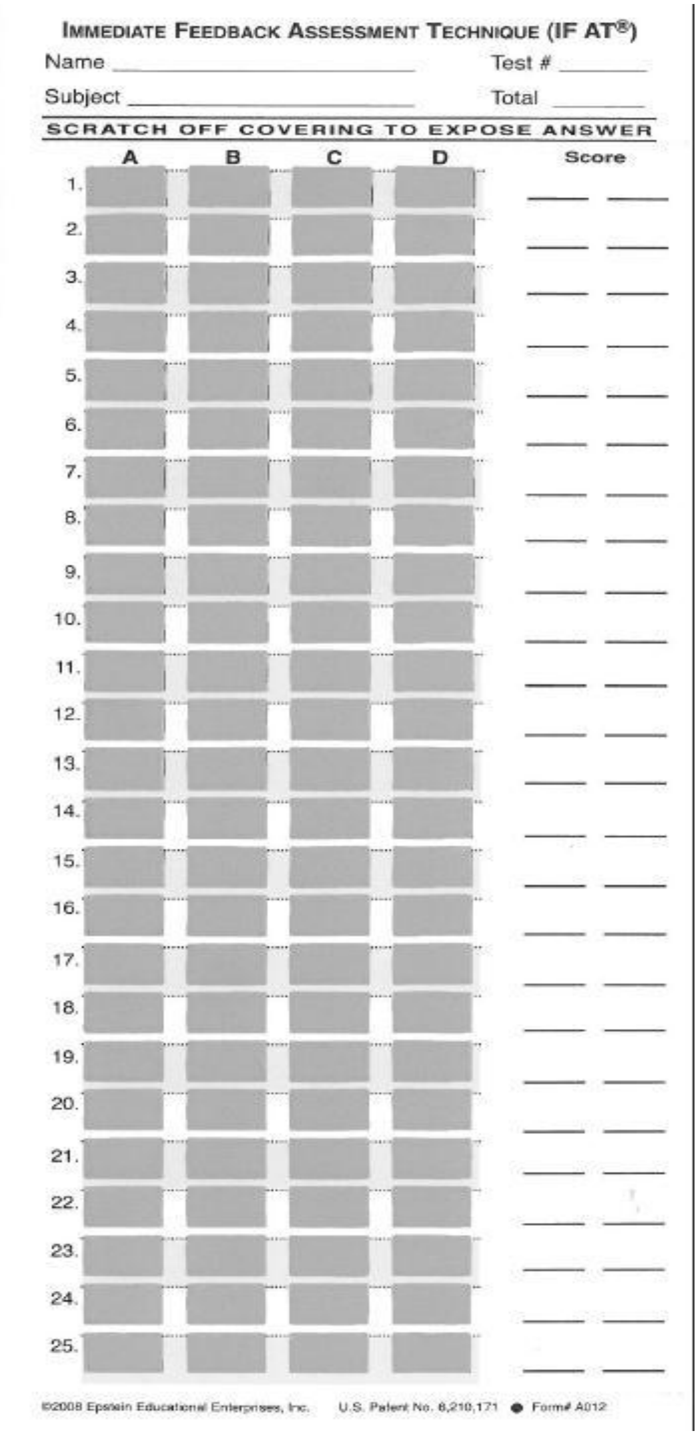
Figure 4.

Percent of correct items by quiz format for statistics final exam material



V. Appendix A

Immediate Feedback Assessment Technique (IFAT) Form



Immediate Feedback Assessment Technique (IF AT®)

Name ____________________ Test # ________

Subject ____________________ Total ________

SCRATCH OFF COVERING TO EXPOSE ANSWER

	A	B	C	D	Score
1.	▨	▨	▨	▨	____ ____
2.	▨	▨	▨	▨	____ ____
3.	▨	▨	▨	▨	____ ____
4.	▨	▨	▨	▨	____ ____
5.	▨	▨	▨	▨	____ ____
6.	▨	▨	▨	▨	____ ____
7.	▨	▨	▨	▨	____ ____
8.	▨	▨	▨	▨	____ ____
9.	▨	▨	▨	▨	____ ____
10.	▨	▨	▨	▨	____ ____
11.	▨	▨	▨	▨	____ ____
12.	▨	▨	▨	▨	____ ____
13.	▨	▨	▨	▨	____ ____
14.	▨	▨	▨	▨	____ ____
15.	▨	▨	▨	▨	____ ____
16.	▨	▨	▨	▨	____ ____
17.	▨	▨	▨	▨	____ ____
18.	▨	▨	▨	▨	____ ____
19.	▨	▨	▨	▨	____ ____
20.	▨	▨	▨	▨	____ ____
21.	▨	▨	▨	▨	____ ____
22.	▨	▨	▨	▨	____ ____
23.	▨	▨	▨	▨	____ ____
24.	▨	▨	▨	▨	____ ____
25.	▨	▨	▨	▨	____ ____

VI. Appendix B

Alternating quiz response format across tests

	Test 1 Material		Test 2 Material		Test 3 Material		Test 4 Material		Test 5 Material		Cumulative Final
Groups	Pre-Test	Post-Test	Pre-Test	Post-Test	Pre-Test	Post-Test	Pre-Test	Post-Test	Pre-Test	Post-Test	Post-Test
A	IFAT	8 Ident 8 Comp	Scantron with FB	8 Ident 8 Comp	IFAT	8 Ident 8 Comp	Scantron with FB	8 Ident 8 Comp	IFAT	8 Ident 8 Comp	16 Identical 16 Comparable To Pre-tests (1-4)
B	Scantron with FB		IFAT		Scantron with FB		IFAT		Scantron with FB		

VII. Appendix C

Example of comparable item

Pretest Question

Time Spent on Math Problems

SCENARIO A

	Male	Female	
US	4	4	4
Foreign	10	10	10
	7	7	

SCENARIO B

	Male	Female	
US	7	7	7
Foreign	7	11	9
	7	9	

11. Which of the following is true for scenario A?

a. **Time spent on math problems was higher for foreign exchange students regardless of gender**
b. There is a moderate interaction
c. Time spent on math problems was higher for foreign exchange students depending on gender
d. Time spent on math problems was higher for males, regardless of where students were raised.

Comparable Post-Test Question

Importance of Religion

SCENARIO A

	Rural	Urban	
Poor	50	50	50
Rich	80	80	80
	65	65	

SCENARIO B

	Rural	Urban	
Poor	50	80	65
Rich	50	50	50
	50	65	

Which statement is true for scenario A?

a.) There is a moderate interaction effect.
b.) Religion is more important to people who live in an urban areas, regardless of their wealth.
c.) Religion is consistently more important for rich people than for poor people, regardless of where they live.
d.) Religion is particularly important to people who are both poor and live in rural areas.

VIII. References

Agnew, J. L., & Redmon, W. K. (1992). Contingency specifying stimuli: The role of "rules" in organizational behavior management. *Journal of Organizational Behavior Management*, 12, 67-76.

Begg, I., Armour, V., Kerr, T. (1985). On believing what we remember. *Canadian Journal of Behavioral Science, 17,* 199-214.

Benjamin, L. T., Cavell, T. A., & Shallenberger, W. R. (1984). Staying with initial answers on objective tests: Is it a myth? Teaching of Psychology, *11*, 133-141.

Bjork, R. A. (1975). Retrieval as a memory modifier. An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition* (p. 123-144). New York: Wiley.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), From learning processes to cognitive processes: Essays in honor of William K. Estes (Vol. 2, pp. 35-67). Hillsdale, NJ: Erlbaum. Based upon research done by: IZAWA, C. Reinforcement-test sequences in pairedassociate learning. Psychological Reports, 1966, 18, 879-919.

Brosvic, G. M., Dihoff, R. E., Epstein, M. L., & Cook, M. J. (2006). Feedback facilitates the acquisition and retention of numerical fact series by elementary school student with mathematics learning disabilities. *The Psychological Record, 56*(1), 35-54.

Brosvic, G. M., Epstein, M. L., Dihoff, R. E., & Cook, M. J. (2006). Acquisition and retention of esperanto: the case for error correction and immediate feedback. *The Psychological Record,56,* 205-218.

Brown, A. S., Schilling, H. E., & Hockensmith, M. L. (1999). The negative suggestion effect: pondering incorrect alternative may be hazardous to your knowledge. *Journal of Educational Psychology, 91,* 756- 764.

Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19,* 514-527.

Butler, A. C., & Roediger, H. L., III. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36* (3), 604– 616.

Butler, A. C., & Karpicke, J. D., & Roediger, H. L., III. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13,* 273- 281.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory and Cognition*, 20, 632-642.

Chew, S. L. (2004). Using conceptests for formative assessment. *Psychology Teacher Network,* 14(1), 10-12.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River, NJ: Pearson.

Daniels, C. D., & Daniels, J. E. (2006). *Performance management.* Atlanta: Performance Management Publications.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41,* 1040 – 1048.

DiBattista, D., Gosse, L., Sinnige-Egger, J., Candale, B., & Sargeson, K. (2009). Grading scheme, test difficulty, and the Immediate Feedback Assessment Technique. *Journal of Experimental Education, 77,* 311-336.

Epstein, M. L., Epstein, B. B., & Brosvic, G. M. (2001). Immediate feedback during academic testing. *Psychological Reports, 88*, 889-894.

Epstein M. L., Lazarus, A. D., Calvano, T. B., Matthews, K. A., Hendel, R. A., Epstein, B. B., & Brosvic, G. M. (2002). Immediate feedback assessment technique promotes learning and corrects inaccurate first responses. *The Psychological Record, 54,* 187-201.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: the place of response certitude. *Educational Psychology Review, 1,* 279 – 308.

Kulik, J. A., & Kulik, C. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58*(1), 79-97.

Mayer, R. E. & Moreno, R. (2002). Aids to computer based multimedia learning. *Learning and Instruction, 12*(1), 32-50.

McDougall, D. (1997). College faculty’s use of objective tests: State-of-the-practice versus state-of-the-art. *Journal of Research and Development in Education*, *30*, 183-193.

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback on discovery-based multimedia. *Instructional Science: Special Issue on Cognitive Load Theory* , 32, 99-113.

Roediger, H. L. III, & Marsh, E. J. (2005). The positive and negative consequences of multiple choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1155-1159.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78, 153-189. DOI: 10.3102/0034654307313795.

Smiley, W. F. (2011). *A systematic evaluation of the immediate feedback assessment technique* (Master’s thesis). James Madison University, Harrisonburg, VA.

Toppino, T. C., & Luipersbeck, S. M. (1993). Generality of the negative suggestion effect in objective tests. *Journal of Educational Research, 86,* 357-362.

Waddel, D. L., & Blankenship, J. C. (1994). Answer changing: A meta-analysis of the prevalence and patterns. *The Journal of Continuing Education in Nursing, 25*(4), 155-158.