James Madison University
JMU Scholarly Commons

Masters Theses | The Graduate School

Spring 2011

A systematic evaluation of the immediate feedback assessment technique

Whitney Flemming Smiley
James Madison University

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

Part of the Psychology Commons

Recommended Citation
Smiley, Whitney Flemming, "A systematic evaluation of the immediate feedback assessment technique" (2011). *Masters Theses*. 331.
https://commons.lib.jmu.edu/master201019/331

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

A Systematic Evaluation of the Immediate Feedback Assessment Technique

Whitney Smiley

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 2011

Acknowledgements

This thesis would not have been possible without several individuals whose guidance in one way or another has led to the completion of this thesis. First and foremost, I owe my deepest gratitude to my advisor of 6 years, Dr. Tracy Zinn, for her supervision and patience throughout this thesis. She has imparted such valuable practical knowledge on me over the years that could not be learned in any classroom. I feel extremely lucky to have been assigned to her as an advisee my first year of my undergraduate education. I hope that one day I can affect a student to the extent that she has affected me.

I also want to thank my undergraduate research group, especially Nicole Giacopelli, Meghan Ward, and Kendra Burek for their commitment to data collection and data entry. Without their help, this thesis would have never been completed in time, nor would the quality of the project be as high. This thesis would not have been possible without the help of Florian Buchholz who spent his own free time developing a quality computer program for a person he had barely met. Lastly, I want to think my committee, Bryan Saville and Jessica Irons for the constructive criticism and feedback they have offered me throughout this process.

Table of Contents

Acknowledgments iii

List of Tables v

List of Figures vi

Abstract vii

I. Introduction 1

II. Methods 17

III. Results 23

IV. Discussion 25

V. Appendix A 38

VI. Appendix B 39

VII. Appendix C 40

VIII. Appendix D 42

IX. Appendix E 43

X. References 44

List of Tables

Table 1. Descriptive statistics for Pre test and Post test scores across conditions.

List of Figures

Figure 1. Immediate Feedback Assessment Technique (IF-AT) Form

Figure 2. Number of Items correct for each condition across time

Figure 3. Percent of correct items by item type for each condition

Abstract

Currently, researchers are conducting many studies on n improving teaching methods. However, research is underdeveloped is improving testing formats so that they can promote student achievement. One way of improving testing could be to incorporate feedback into the testing procedure. This study replicated and extended past research done Epstein, Epstein, & Brosvic's (2001) by for a testing technique called the Immediate Feedback Assessment Technique (IF-AT), which allows test takers to receive immediate feedback on each response during testing. I extended Epstein's research by including two new conditions (scantron with feedback and computerized IF-AT). Results showed that adding feedback into the scantron produced significantly higher performance than the scantron, IF-AT, and CIF-AT on both identical and comparable items on the post test

I. Introduction

In universities, the typical teaching technique is the lecture format (Benjamin, 2002). However, over the years, research has generated new and engaging teaching formats that can be more effective than the traditional lecture (Davis, 1993; McKeachie, 2002). What has received less attention is new and engaging testing formats that teach while testing. What is probably the ideal testing situation is an essay test. Not only are essay tests easier to construct, but the format also allows students to construct their own individualized responses to questions, which allows the instructor to be certain that the student is demonstrating knowledge of content and not relying on guessing. However, essay-testing formats have many drawbacks. Some courses involve several hundred students and grading this number of essay tests in a reasonable amount of time is unfeasible. Also, supplying corrective feedback to essay tests involves considerable time because the answers the students give result in greater variation than that of close-ended tests.

Because of the drawbacks of essay tests, instructors frequently use multiple choice testing. With multiple choice testing, tests and quizzes can be scanned through machines and graded much more quickly. Because of this, students typically receive summative feedback about their work much sooner than they would after an essay test (Mislevy, 1991). Although this sounds ideal, scantron forms still have drawbacks. Students receive summative feedback more quickly, however, the scantron form provides little opportunity for elaborate feedback or error correction at the time of test taking. It is possible that including feedback into the test-taking process will improve student learning.

In this paper, I will discuss: (a) different definitions of and explanations for the effectiveness of feedback, (b) feedback best practices, (c) feedback in testing, and (d) how the proposed study builds on the existing literature on feedback in multiple choice testing. Specifically, I will propose an investigation of a testing method called the Immediate Feedback Assessment Technique (IF-AT), which is an answer-until-correct procedure. The proposed study aimed to link different areas of feedback literature in order to understand the utility of the IF-AT method.

Feedback

People have used feedback as an effective intervention for behaviors in industry for decades (e.g., Komaki, Barwick & Scott, 1978; Sulzer-Azaroff, 1978). Beyond industry, research has also shown that feedback is effective for improving many different types of behaviors in sports, education, and energy consumption. For example, Komaki and Barnett (1977) have used feedback to improve target behaviors in little league football, Gaynor (1981) used feedback to feedback to improve retention of computer based math information, and Seligman and Darley (1977) used feedback to reduce energy use in homes.

Although researchers regularly incorporate feedback into applied studies, there is no universal definition of feedback. Daniels and Daniels (2006) defined feedback as information that serves two functions. First, feedback must tell the performer where he or she stands relative to some target or goal (e.g., you answered 80% of the items correctly). Second, feedback should tell the performer what to do to improve (e.g., you missed several questions on WWII, you should review that chapter). However, this definition of feedback is not widely used, and feedback research typically does not address both of

these functions, usually focusing on just the first function. Other definitions of feedback may include just one of these functions, or be specific about the kinds of information that have to be transmitted (e.g., information regarding quality or quantity, [Sulzer-Azaroff & Mayer, 1991]; or information that tells people what and how well they are doing, [Rummler & Brache, 1995]).

Current literature discusses several mechanisms by which feedback may work. According to Daniels and Daniels (2006) and Komaki, Barwick and Scott (1978) feedback is effective due to reinforcement, a consequence of a past behavior that increases the probability of the same behavior occurring in the future. Daniels and Daniels (2006) explain that feedback is only effective if reinforcement is associated with improved performance. Therefore, feedback alone does not improve performance, performance increases because of the consequences or expected consequences associated with feedback. Feedback is most effective when it is a discriminative stimulus (i.e., any stimulus that gives information on the availability of reinforcement for a particular behavior) for positive reinforcement (Daniels & Daniels, 2006)

Other researchers have argued that feedback can be a motivating operation (i.e., any condition that alters the reinforcing value of a particular consequence [Agnew, 1998; Agnew & Redmon, 1992]). If feedback functions as a discriminative stimulus, it would have to be correlated with the presentation of a reinforcer and would evoke behavior immediately. However, this relation is not too common in organizations, yet feedback is shown to improve behavior. Instead, feedback might change the reinforcing effectiveness of performance related behaviors in the future (i.e. a motivating operation). As you can see, both the definition and function of feedback is still in debate.

Feedback Best Practices

Applied research in the area of organizational behavior management (OBM) has focused on feedback for many years. When Balcazar, Shupert, Daniels, Mawhinney, and Hopkins (1989) reviewed studies published in the *Journal of Organizational Behavior Management*, work performance changed following some type of feedback in half of the articles. Since this review, the use of performance feedback as part of a behavioral intervention has only increased (Alvero, Bucklin, & Austin, 2001; Nolan, Jerma, & Austin, 1999).

Alvero, Bucklin, and Austin (2001) reviewed four journals (e.g. *Journal of Organizational Behavior Management*, *Journal of Applied Psychology*) for articles investigating feedback interventions. The aim of this review was to find patterns in how researchers used feedback and what characteristics of feedback yielded consistent effects. The authors examined 68 applications of feedback. From these articles, the authors discerned that the effectiveness of feedback had distinct patterns based on the different sources (e.g., self, supervisor), mediums (e.g., verbal, graphic), frequency (e.g., daily, weekly), participants (e.g., individuals, groups), privacy (e.g., public, private), and content (e.g., comparison with individuals previous performance, comparison to group performance) that was involved in the feedback procedure. Below is a summary of some of Alvero, Austin and Bucklin's results:

1. Source - Feedback is most effective when delivered by a supervisor and researcher; although very few articles included feedback given through mechanical means.

2. Medium – Feedback given in the form of graphs and written feedback were the most helpful.
3. Frequency – More frequent feedback seems to result in better performance especially when behaviors are new.
4. Content – most effective when comparisons are made with current performance and previous individual performance combined with standard individual performance.

From the Alvero et al. (2001) study, it is clear that feedback has an effect on target behaviors. In the majority of the studies, feedback had consistently positive or mixed effects (no negative effects). However, feedback involves an array of characteristics that can be combined in many ways. Thus, the authors concluded that despite the fact that people have used feedback for 30 years, we still do not know how to use feedback to its greatest potential. Based on the literature in OBM, feedback that is from a supervisor, given relatively frequently, and is specific regarding individual performance will be most beneficial

Feedback involved with testing

Instructors use several different feedback methods of testing in education. Item-by item knowledge of correct answers, answer-until correct procedures, and test corrections are a few of the different feedback methods teachers can use. Shute (2008) described different methods of feedback, organized along the continuum of how elaborate the feedback is. The list (in order of least elaborate to most elaborate) is as follows: no feedback, verification, correct response, try again, error flagging, elaborated, attribute isolation, topic contingent, response contingent, hints/cues/prompts,

bugs/misconceptions, and informative tutoring (see Shute for a complete explanation for each type of feedback). Research has found that the more specific the feedback used, the better the outcome (Phye & Sanders, 1994; Bangert-Drowns et al., 1991; Pridemore & Klein, 1995). However, other researchers use different terms and possibly different types of feedback. For example, Bitchener, Young, and Cameron (2006) wanted to compare different types of feedback for adult English as a second language (ESL) learners using no corrective feedback, explicit written feedback only, and explicit written feedback with a 5-min student-researcher conference. Researchers found that the participants' accuracy was significantly higher after receiving the combination of written and conference feedback than the accuracy of those in the other conditions. Notice that these terms are similar to the terms used by Shute, yet different terminology, which makes comparisons across studies difficult.

While there are several different ways that feedback can be involved in testing. The type of feedback analyzed in the current study is the answer-until-correct type of feedback, specifically the IF-AT. Although this procedure does not fit neatly into any of Shute's (2008) categories, it consists of different combinations from Shute's list. For example, the answer-until-correct procedure is a combination of correct response, try again, and hints/cues/prompts. The usefulness of this type of feedback in testing is gaining in popularity, which has led to the production of testing format such as the IF-AT.

Epstein, Epstein, and Brosvic (2001) developed an example of a commercially sold feedback testing procedure that utilizes the answer-until-correct style of examination. The IF-AT is an answer form, similar to a scantron, with rows of

rectangular answer spaces for each question on a multiple-choice test/quiz. In order to answer the question, the student must scratch off a waxy coating, where he or she will find either a blank box or a symbol (e.g., a star; See Figure 1). If the symbol appears, the student has answered the question correctly, if the student instead sees a blank box, he or she has not answered the question correctly on the first attempt and is free to try other answers until he or she answers the question correctly.

In Epstein, Epstein and Brosvic's (2001) original study, researchers gave 75 students enrolled in two sections of an introductory psychology course the same examinations. The only difference between the two sections was the form on which the students recorded their answers. One group received a scantron form where they bubbled in their answers with a pencil, whereas the other section received the IF-AT form described above. IF-AT users were allowed to continue answering until the right answer was revealed; however, they could only earn points for answering correctly on the first try. Neither the scantron group nor the IF-AT group reviewed the test after the testing session was over. The students completed four unit tests in the same manner throughout the semester. On the final examination, both sections used a scantron form. The final examination consisted of 38 new questions and 3 questions from each of the four previous unit tests.

Results from Epstein, Epstein, and Brosvic (2001) showed the two groups did not differ in initial accuracy on any of the unit tests. However, on the final examination, students using the IF-AT all semester were twice as likely to respond correctly as students in the scantron group were. Although the mean for the repeated questions were

higher than for new questions for both groups, IF-AT users were significantly more likely to answer the repeated items correctly than the scantron group.

In 2002, Epstein, Lazarus, Calvano, Matthews, Hendel, Epstein, and Brosvic continued research on the IF-AT with three separate studies. In the first study, students enrolled in an introductory psychology course worked in groups of five and completed a multiple-choice trivia test. Researchers randomly assigned groups to use the scantron or the IF-AT and the groups received the same instructions on how to use the answer forms (see Epstein, Epstein, and Brosvic, 2001). After completing the initial test, researchers randomly assigned students to return in either 1 day or 1 week. When the participants returned, they took the same test again with the questions in a different order. Researchers scored the IF-AT form based on students' initial accuracy (i.e., if students answered correctly on their first attempts). Results showed that in the initial testing, IF-AT and Scantron users did not differ from one another. However, mean scores at the 1-day and 1-week delays were significantly higher for the IF-AT users than the scantron users. In fact, IF-AT users significantly outperformed their initial testing session. In addition, scantron users' results did not change significantly over time.

In the second study, Epstein et al. (2002) asked individual participants to read a three-page article on extrasensory perception. At the conclusion of the initial session, participants answered 15 multiple-choice questions about the article. Half of the participants answered these questions on an IF-AT whereas the other half used a typical scantron. Neither the IF-AT nor the scantron users were able to review the test after the initial testing session. Participants then returned either one day or one week later and responded to a post test with different but comparable questions. On this post test, all

participants responded on a regular scantron form. Results revealed that no significant differences existed during the pre test of the material (as measured by first response). However, IF-AT users yielded a significantly higher number of correct responses on the post test at both the one-day and the one-week delay. Furthermore, following the scantron pre test, quiz scores actually decreased from one to ten days after the initial testing session, whereas following the IF-AT pre test, users demonstrated enhanced quiz scores at both time points.

Epstein et al. (2002) also compared the IF-AT to a computerized version of the IF-AT in the third study. The computerized IF-AT directly imitated the paper format (although specific details about the computer program were unavailable at the time of this writing). Researchers randomly assigned 45 participants enrolled in an introductory psychology course to one of two conditions. The first group of students completed a quiz using the paper IF-AT, whereas the second group used the computerized IF-AT. Participants then took a quiz either 1 day or 1 week later on a paper scantron. Results showed that the computerized IF-AT did not improve retention, and the paper IF-AT resulted in significantly higher post test scores at both delay periods. The mean percent of change in the IF-AT was higher for the 1-day delay than the 1-week delay. The mean percent change in the computerized IF-AT was lower for the 1-day delay than for the 1-week delay. The IF-AT conditions significantly outscored the computerized IF-AT on the post test. Thus, students actually did worse on quizzes for both of the delay periods following the computerized IF-AT despite receiving feedback on their responses. Researchers concluded that students who receive the first test with the paper IF-AT performed better than those who used the computerized version of the IF-AT.

In addition to the effectiveness of the IF-AT in research, participants reported enjoying the format. Epstein and Brosvic (2002) gave students a multiple-choice questionnaire after completing an exam with either a scantron or IF-AT. Students responded to 15 questions on a 1 (*much less*) to 5 (*much more*) scale regarding their perceptions of the testing process. Results showed that test anxiety, time requirements, and satisfaction with response format did not differ by response format. Clarity of response requirements, desirability of the response format, and the benefits of testing were significantly higher for the participants who used the IF-AT. These results were consistent with a study conducted by DiBattista, Mitterer, and Gosse (2004) who surveyed 234 upper level undergraduate students enrolled in the same course. All students reported that in previous courses, they used scantrons to answer multiple-choice questions. The course in which they were enrolled during the study used the IF-AT procedure for a 3-hour test. Along with the test, the students received a 22-item questionnaire about their opinions of the IF-AT procedure. The majority of the participants (83.6%) said they would like to use the IF-AT in all of their courses that have multiple-choice testing. Five percent of the participants reported not liking the IF-AT and fewer than 10% preferred the Scantron to the IF-AT.

The IF-AT does capitalize on some of the recommendations from the feedback literature. According to Shute's list (2008), the IF-AT contains verification as well as "try again" feedback. It also capitalizes on the value of the immediacy of feedback due to its answer-until-correct presentation. The IF-AT also conforms to several recommendations from the OBM literature, in that it is specific and tells the student what he needs to do to answer correctly (i.e., keep reading the options); also frequent feedback is given (after

each individual question) and feedback is given in written form. Not only does the participant instantly know whether the answer is right or wrong, she still can continue to review the answers, and the path to the correct answer is also narrowed down. Kulhavy (1977) researched feedback in information retrieval and application on test-like items. Results showed that when intentionally including feedback, the most important effect of feedback is correcting erroneous responses not to strengthen correct responses. Scantron forms do not offer this type of error correction, whereas the IF-AT clearly does.

Computerized feedback

Mechanical testing in the classroom can be beneficial to students and teachers in the classroom. If instructors use computer testing, they can assess students outside of the classroom, which would save valuable class time. In addition, the computer program can score multiple-choice tests immediately and present feedback to students right away, leaving the teacher with more time for classroom preparation, and giving students information about their performance more quickly. However, computer assisted feedback has not received much attention in the psychology literature, despite its popularity and obvious advantages. Because of this, instructors use feedback during computer testing based more on intuition than on well-founded feedback principles (Narciss & Huth, 2002).

One interesting finding from Epstein et al. (2002) was that they despite the fact that the paper IF-AT was successful at improving knowledge retention, a computerized version of the IF-AT actually led to a decrease in scores in the second administration of the test. Because both of the answer forms contained feedback during the testing session, and researchers suggested that feedback was the mechanism for improving performance,

it does not seem logical that the computerized version of the feedback answer form is not successful. Could presenting feedback mechanically be hindering feedback effectiveness?

Sturges (1978) investigated the applicability of feedback findings in previous literature to computer-assisted testing in an educational setting. Researchers randomly assigned students to one of four feedback conditions: no feedback, 2-second delay, 20-minute delay, and 24-hour delay. After the initial test, participants received feedback according to their random group assignment. Students then returned 1 to 3 weeks later, when they completed a paper-and-pencil test. This post test had some identical items from the original test, plus 17 new items. Results showed a statistically significant increase in the second test compared to the first test for all of the feedback groups, with delayed feedback (i.e. the 20-minute and 24 hour delay conditions) promoting retention better than immediate feedback. The feedback participants scored higher on all of the identical items, yet no statistical differences existed on the 17 new items.

Gilman (1969) taught 75 undergraduates 30 general science concepts. The participants were assigned to one of five conditions: no feedback, feedback of "correct" or "wrong", feedback of the correct response choice, feedback appropriate to the student's response, and a combination of the feedback given in the second, third, and fourth conditions (referred to as the combination condition). The computer went through iterations based on previously incorrectly answered question in all conditions until the student answered all 30 questions correctly. Participants took a paper-and-pencil post test after the computer instruction that contained 30 items similar to those of administered during the computer program. Results showed that the feedback of the correct response choice, feedback appropriate to the student's response (i.e. specific to the answer they

selected), and the combination groups required significantly fewer responses to reach the criterion then the no feedback and the right/wrong feedback groups. Also, the number of iterations the computer had to go through before the student reached criterion was significantly lower for these groups compared to the no feedback and right/wrong feedback groups. The post test results showed that the combination group was significantly superior to all other groups on their post test scores.

Kluger and Adler (1993) investigated the effects of feedback provided by a person versus that provided by a computer. Undergraduate students participated in the experiment where they took a test consisting of seven items from a retired Graduate Record Examination. Researchers randomly assigned participants to receive feedback by a person or via a computer, as well as to the feedback condition (no feedback, automatic feedback after every trial, or feedback only upon request). Participants also completed questionnaires about personality and their opinions about computerized feedback. Results showed that participants were more likely to seek out feedback from a computer than from another person. In addition, performance declined when a person delivered feedback relative to when there is a person in the room yet not giving feedback. This was not the case in the computerized feedback condition.

The above research provides insight on computerized feedback. First, from all three experiments there is evidence that researchers have successfully delivered feedback electronically. Each of these studies found a positive effect despite the fact that a computer was used. Second, feedback delivered electronically results in similar performance to feedback from non-computerized sources. An example is Gilman (1969) who found that the more elaborate the feedback is, the better the outcome. This was also

found in Shute's (2008) meta-analysis of the education literature. Additionally, Sturges (1978) found that feedback delay has an effect on outcome measures. This was demonstrated in both Shute (2008) and Alvero, Austin and Bucklin (2001). If patterns of feedback that were found in feedback research and on feedback delivered electronically, this provides evidence that electronic feedback mimics paper based feedback. Third, computerized feedback can be beneficial in that people are more likely to seek out computerized feedback. Kluger and Adler (1993) lend evidence that people may be more responsive to feedback from a computer than feedback from a person. In addition, performance actually declined when feedback was delivered from a person compared to a person just being in the room not providing any feedback, which actually shows computers can be a beneficial way of delivering feedback. Furthermore, computerized feedback has been shown to be effective in a number of articles (e.g., Sneider & Shugar, 1990; Clariana, Ross & Morrison, 1991; Pridemore & Klein, 1991).These studies in combination start to build an argument that computers do not dilute the effects of feedback and the differences found by Epstein et al (2002) are likely due to an alternative explanation.

One important thing to note about these studies on computerized feedback is that they are fairly dated. Computers have come a long was in just the past few years and the most recent study discussed above was done over 17 years ago. Epstein states,

> I think the reason for the difference between immediate feedback on a computer versus that on a "hardcopy" IF-AT form is a matter of interest and involvement…The scratch-off coating on IF-AT forms is a rubbery, latex-like ink and there is some tactile pleasure in scratching it off; it's kind of fun! Add that to the discovery component of seeking the right answer and the using IF-AT has a game-like quality. Students tend to be more interested and involved when using IF-AT (personal communication, December 2009).

Although this explanation makes sense, if a computerized version is made to mock the paper based version the computerized version would have many of these aspects as well. Also, this explanation does not account for the actual decline in scores on Time 2 tests. Not even the scantron form, which involved the same types of behavior and lack of involvement, had decreased scores (Epstein et al., 2002). Because the computerized IF-AT hindered knowledge despite the fact that previous research has been successful with computerized feedback, further study is necessary.

Current Study

Because there is evidence that feedback in testing can improve learning retention, it is definitely worthy of study. So far, studies have shown that the IF-AT has promise on promoting learning retention above the traditional scantron use. However, only a few have researched the IF-AT. Also, the evidence that computerized version of the IF-AT does not promote learning retention is contradictory to previous studies done on computerized feedback. Because of this, it is important that previous research on the IF-AT be replicated.

In addition to replicating previous work, the current study also looks for evidence as to why the IF-AT performs better than the scantron. Epstein (personal communication, December 2009) stated that the tactile pleasure of scratching might be why the IF-AT is so successful. If this is true, it is important to note and can be important to the future development of testing formats. In order to determine if feedback is still useful without the scratching behavior component, it would be useful to see if feedback incorporated into the traditional scantron form promotes learning retention. Overall, the aim of the current study is to replicate and extend the IF-AT research by examining the differences

in learning retention among the scantron, scantron with feedback, IF-AT and computerized IF-AT testing formats over the course of one week.

II. Method

Participants

Participants (*N* = 189) were undergraduate students at a large public university recruited through the psychology participant pool. Each participant earned two credits of participation credit in an introductory psychology course. Participants also received incentives, such as entries into drawings, for completion of and performance on the post test (see below for details). The overall population of the university had an average age of 20 and was comprised of 60.3% women.

Materials

Testing formats. I used a scantron sheet, an IF-AT, and a computerized version of an IF-AT created with Java software as my testing formats. Scantrons are sheets of paper used to record answers where students darkened in a circle that corresponded to the answer she wanted to select. Scantrons had 50 question blanks on each side of the answer form and contained five choice options, A through E.

When participants responded on IF-AT forms, they scratched off the waxy opaque covering of their choice that reflected the desired answer. If a participant scratched an answer choice that was correct, a star was revealed; however, if the choice was blank the answer was incorrect and the participant could continue to scratch the options until the answer was revealed. The placement of this star was randomized across questions. The IF-AT contained 25 answer blanks and four response options, A through D. Therefore, all questions for the quizzes had four response options.

An expert designed the computerized version of the IF-AT using java. The program presented all questions simultaneously to simulate the paper format. A

participant answered each question by clicking the mouse and dragging it across the box to select the desired response. This method mimicked the scratching needed to respond on the paper IF-AT. After a participant selected an answer, a star or an empty box was revealed to indicate whether the answer was right or wrong. The participant could continue to select options until he or she choose the correct answer.

Reading material and quiz questions. The reading material used in the study was a textbook titled *Behavioral Principles in Everyday Life* (Baldwin & Baldwin, 2001). The topic of the 10-page chapter was science as it relates to human behavior. I conducted the pilot study using the same chapter and quizzes. Based the pilot data, I determined that 30 min would be sufficient time for reading the chapter.

I developed two quizzes for use in this study using the instructor's manual for the textbook (Baldwin & Baldwin, 2001). After selecting a sample of questions from the test bank, I wrote comparable questions for each of those from the test bank. This allowed a large enough question pool to form two separate, yet comparable, quizzes. I combined the original quiz questions from the instructor's manual with the questions that I wrote on each of two quizzes. I then conducted a pilot study to investigate the usability and difficulty of the questions. Again, each question on Quiz 1 corresponded to the content addressed in a question on Quiz 2. I compared the item difficulty (as measured by percentage of respondents answering correctly) of each question across quizzes. If the comparable questions from the two quizzes had a difference of difficulty levels that were 10% or more, I adjusted one of the questions to make it more similar to the other question. If I could not change the question, the question was repeated and made identical on both quizzes. This process resulted in two quizzes that each included both six repeated

questions (i.e., identical questions) and nine novel (i.e., comparable) questions. This allowed me to determine the effects of each testing format on both identical questions as well as comparable questions.

Gift Certificates. Local restaurants donated gift certificates to offer as incentives to participants to complete both parts of the study as well as to use for motivation to do well on the quizzes. These gift certificates included items such as free cookies, free teas, free breadsticks, and free entrees.

Design

The current study used a 4 (test format) x 2 (item type) x 2 (time) mixed design. Participants signed up for a testing session through an online signup system and participated in one of four possible conditions: (a) a scantron with no feedback, (b) scantron with feedback, (c) IF-AT, or (d) computerized version of the IF-AT (CIFAT). After 1 week, all participants returned to complete a post test using a scantron. Participants did not receive feedback on their post test responses unless they contacted me after the experiment for their results.

Procedure

General procedure. As stated above, the study involved participants completing two sessions. Four types of initial testing sessions occurred, each having between 42-50 participants. Which session the participant signed up for determined what condition the participant completed. During the initial session, after an introduction to the study and securing informed consent, I instructed all participants to read a short introductory chapter (described above). During these instructions, I pointed out that they were to take a quiz on the subject matter during the pre test as well as 1 week later when they came

back for the post test. I told all participants that each question they answered correctly on the post test would enter them into a drawing for several gift certificates. Therefore, it was beneficial for the students to attend to the quiz questions and material. This was done to encourage the participants to perform their best on both the pre test and post test, hopefully motivating the participants who had feedback available to them to use the feedback to learn the material. After 30 min elapsed, all participants completed a 15-item quiz about the reading using the format corresponding to the condition to which the participant was assigned. All participants came back exactly 1 week later and answered the second 15-item quiz on a scantron answer form. The scantron (n = 49), IF-AT (n = 50), and CIF-AT (n = 42) participants were given 15 minutes to take the quiz and were instructed to review their answers with their remaining time.

The participants in the scantron with feedback condition (n = 48) were given 10 min to complete the quiz and review their answers. I also told them to write their answers on both the scantron as well as on the paper test form they recieved. After the 10 min elapsed, I collected the scantron forms (they were allowed to keep the answers they wrote on the paper test) and distributed a list of the correct answers. I instructed participants to review the questions and the correct answers for 5 min. I did this because reviewing the correct answer is built into the IF-AT and CIF-AT testing procedures, but not in the scantron with feedback condition.

At the conclusion of the experiment, I entered participants into a drawing for every correct answer they received on the post test. Participants were contacted by e-mail if they won a prize (more details below).

Post test. One week after the conclusion of the first quiz, participants returned to the same classroom at the same time and took the second quiz. All participants responded to this quiz on scantron form. I read directions on how to use the scantron form, and then participants had 15 min to complete the quiz. I reminded participants that every question they answered correctly entered their names into a drawing for several gift certificates. Directions given to participants for each condition as well as during the post test can be found in Appendices A-E.

Dependent variable

Initial accuracy was measured by whether the participant answered the question correctly on the first attempt. With the IF-AT and CIF-AT, participants were able to continue selecting answers until observing the feedback indicating the correct response; however, these responses were not included in initial accuracy measure.

Gift certificate drawing

After all post test quizzes had been scored, participants were entered into a drawing for several prizes. The number of correct responses on the post test determined who was entered into the drawing and how often their name was entered. Therefore, if a participant correctly answered all 15 items, he or she had 15 chances to win a prize. A participant who answered seven items correctly would only have seven chances to win a prize. The participants earned several prizes such as free buffets, entrees, cookies, and drinks donated from local restaurants. I entered participants into a Microsoft Excel spreadsheet and selected using the random number generator feature. The gift certificate drawing was a tool used to motivate people to attend the post test as well as try their hardest on the pre test so they can get more items correct on the post test. Knowledge of

the drawing and was given in the instructions at the beginning of the pre test. Selected participants collected their prizes by showing up to a predetermined room.

III. Results

The data were analyzed using a 4 (testing format: scantron, scantron with feedback, IF-AT, and CIF-AT) x 2 (time: pre- and post test) x 2 (item type: identical and comparable) mixed ANOVA. Prior to running the ANOVA, assumptions were tested. After visual inspection and examining skewness and kurtosis of the variables, I determined that the data were approximately normally distributed. The Levene's test of equality of error variances was not significant ($p > .05$) for all variables, meaning that the assumption of homogeneity of variance was not violated.

The results of the mixed ANOVA showed that all effects were significant at the $p < .05$ level. There were main effects of testing format ($F[3, 185] = 7.42$, $p < .01$, $r = .20$), item type ($F[1, 185] = 88.91$, $p < .01$, $r = .57$), and time ($F[1, 185]) = 87.94$, $p < .01$, $r = .57$) as well as two interactions (described below). The three-way interaction was nonsignificant.

There was a significant two-way interaction between time and testing format ($F[3,185] = 7.3$, $p < .001$, $r = .19$), indicating that the improvement seen from pre test to post test was dependent on the testing format. Figure 2 shows a graph of average items correct for the four testing formats across time. The error bars represent 95% confidence intervals. Visual inspection (recommended by Goldstein & Healy, 1995) of the overlap in confidence intervals was completed to detect difference among the groups. As the graph shows, participants in the three testing formats did not score significantly differently from each other on the pre test. However, on the post test, participants who used the Scantron with feedback ($M = 10.92$, $SD = 2.15$), IFAT ($M = 9.58$, $SD = 2.47$), and CIFAT ($M = 9.69$, $SD = 2.02$) significantly outperformed the scantron group ($M = 8.27$, $SD = 2.49$).

The CIFAT and IFAT users scored similarly on the post test, whereas the scantron with feedback users significantly outperformed all of the other testing formats. See Table 1 for descriptive statistics.

There was also an interaction between item type and testing format ($F[3,185] = 7.41$, $p < .001$, $r = .20$). Figure 3 which shows a separate graph for each testing formats that compare the mean percent of correct identical and comparable items. The scantron graph shows that there was no significant increase in the percentage of correct identical or comparable items from pre test to post test. For the scantron with feedback group, the percentage of correct items for both the identical items and the comparable items increased from pre test to post test. For both the IFAT and the CIFAT, there was no increase in the percent of correct comparable scores from pre test to post test; however, the percentage correct identical items increased over time. See Table 1 for descriptive statistics.

IV. Discussion

The current study showed that participants in the scantron conditions did not show any evidence of learning retention a week after their initial quiz. However, participants in the CIF-AT and IF-AT testing formats performed similarly, and significantly outperformed the scantron participants. Finally, the scantron with feedback participants performed significantly better than the participants in the other three conditions. Results also showed that participants scored differently on comparable and identical items depending on which testing format they used. The scantron users did not score significantly better on either comparable or identical items, the CIF-AT and IF-AT participants only scored significantly better on the identical items, whereas the scantron with feedback users scored significantly better on both the identical and comparable items.

The current study replicated a portion of Epstein et al.'s (2002) results. First, the results showed that the IF-AT testing format significantly outperformed the scantron testing format. On the post test, participants in the IF-AT condition outperformed the scantron condition by 9 percentage points with the scantron users average score equivalent to an F letter grade and the IF-AT users average score equivalent to a D letter grade. This is a problem because a D in a college course is typically not a favored grade and reflects poor mastery of material. Therefore, although the IF-AT group scored significantly better, the performance would not be indicative of mastery of material. Unlike Epstein's et al.'s research where IF-AT participants demonstrated learning on both identical items and comparable items, the current study found that IF-AT participants only scored significantly higher on the identical items after the one-week

delay, there was no evidence that they scored higher on the comparable items. This could mean that participant knowledge is not generalizing to new items when using the IF-AT and participants are simply able to remember the old items better with the IF-AT. The cause for the discrepancy across studies could be due to the lack of detail in Epstein et al.'s definition of what it means to be a comparable item. In the current study, comparable items share the same content as well as had the same difficulty level during pilot testing.

Unlike in Epstein et al.'s (2002) study, the CIF-AT group did score comparably to those in the IF-AT group. The IF-AT and C-IFAT groups' average percentage correct on the post test were 64% and 65%, respectively, both of which would be equivalent to a D letter grade. The participants in the CIF-AT group also mirrored the participants in the IF-AT group by scoring significantly higher on the identical items but not on the comparable items. This discrepancy could exist because of the lack of information Epstein et al. provided about their CIF-AT. It is unlikely that the two were the same as there is a nine year gap between the two studies during which technology has drastically changed. In personal communication with Epstein he suggested that the difference between the paper and computerized testing formats could be to tactile pleasure of scratching the IF-AT or perhaps that the paper version had a game-like quality. Because the CIF-AT used in the current study mimicked the IF-AT as closely as possible and had comparable results to the IF-AT, it is likely that Epstein's theory is not the case. Previous literature also suggests that computers are successful at delivering feedback (e.g., Sneider & Shugar, 1990; Clariana, Ross & Morrison, 1991; Pridemore & Klein, 1991), providing further evidence that Epstein may not be accurate in his conclusion.

The current study also extended Epstein et al.'s (2002) research by incorporating a fourth testing format type—the scantron with feedback condition. This condition was added to the current study in order to understand if the scratching behaviors using the IF-AT are what led to higher post test scores. This study did not support that scratching behaviors improved post test scores. In fact, the study showed that scantron with feedback testing format resulted in post test performance that exceeded that of the other testing formats. Participants in the scantron with feedback condition scored an average of 72%, which is both statistically and practically higher than the other three testing formats. Participants in the scantron with feedback condition scored equivalent to a C letter grade and scored significantly higher on both comparable and identical items, unlike the CIF-At and IF-AT (referred to as C/IF-AT throughout the rest of the paper) conditions.

I hypothesize that perhaps the participants are using the feedback differently across conditions. It is likely that feedback can be used in myriad ways because previous research cannot come to a consensus about what the exact function of feedback is (Agnew, 1998; Agnew & Redmon, 1992; Daniels & Daniels, 2006; Komaki, Barwick and Scott, 1978) nor can they decide on an exact definition (Alvero, Austin, & Bucklin, 2001). While feedback is given in the C/IF-AT and scantron with feedback condition, the participants are given different instruction from feedback. Participants in the C/IF-AT are receiving a star even if they do not get the answer right until the third try. This star could potentially be a cue to the test taker that they can move on. With the scantron with feedback format, after participants took the quiz they looked at the test as a whole and realized exactly what items and how many items they answered incorrectly, which could

be a cue to stop and review the incorrect material. Therefore, in the two formats the feedback is cuing different behavior from the participants.

Another hypothesis is that the timing of feedback could affect learning retention. First, the C/IF-AT allows the student to read each question, mark an answer and, if he or she answered incorrectly, he or she can then select a different answer until selecting the correct answer. This procedure allows the participants to get feedback after every question. The scantron with feedback on the other hand, allows users to go through the whole test, and then compare their answers to the answer key. With this format, users get feedback only at the end of the test. According to Alvero, Austin, and Bucklin (2001) and Daniels and Daniels (2006), the more frequent the feedback, the better. However, in the scantron with feedback group, the feedback is actually further delayed, which, in theory, would mean that the IF-AT should outperform the scantron with feedback. However, this is not the case. Although this seems counterintuitive from previous research, perhaps because the timing is only minutes apart (Alvero, Austin, and Bucklin look at timing days, weeks, and months apart) the effect of timing is diminished.

Limitations

Although this study does provide evidence that the scantron with feedback condition is superior in lab settings to the C/IF-AT, teachers will be interested in the performance of students in real classroom settings. Certainly, the variables in classroom contexts are more complex then lab settings. With the IF-AT, users know they received a wrong answer so the user is forced to check their answers as they go along. If a student takes a test and is asked to review their answers using a key, he or she might not check

their answers because the quiz is over with and they might not see the utility in checking their answers (or they might be more likely to check their answers than in the lab situation). In addition, instructors may opt to allow extra points with the IF-AT (i.e. they may get a point off if they do not get the answer correct on the first try, two points off if they cannot get the answer correct in two tries) With the scantron with feedback, this type of scoring is not an option.

Another limitation could be that I did not collect participants' demographic information. Because I was unable to use random assignment, perhaps my sample is not representative and could potentially confound the study. Had I collected demographics, I could compare the demographic to the university's population to make sure that the sample was representative.

The fact that the scantron with feedback did not mock every aspect of feedback that the IF-AT has is another limitation. Given how answers are selected on the scantron, it is impossible to give feedback to the participants after each question. Since there is no star to denote the right answer, the researcher would have to give the participant the correct answer after every question. If this is done, the participant could change their answer as there is no indicator (such as a star on the scantron) to which the participant scratched last (assuming they stopped scratching once they received the correct answer on the IF-AT). Asking the participants to mark their answers down on the quiz and scantron, then collecting the scantron and allowing the participants to compare their quiz questions with the key is the only feasible way to incorporate feedback into the scantron.

Also, since review time is built into the IF-AT testing form, I had to allow scantron with feedback users time in which to review their feedback as well. Because of this, there is a possibility time with material could impact the findings. Again, since the only way to incorporate feedback into the scantron is to present it after the participant has taken the whole test, there was no way to include review time and testing time in the same interval. Since error correction seems to be the most important feature of feedback (Kulhavy, 1977), making sure that the student got time to review their answers was important. Due to these limitations, time and timing of feedback could be confounding variable that need to be untangled in future research.

Future Direction

In conclusion, there is evidence that adding feedback to testing formats can increase students' performance on identical items over time. However, whether or not participants could correctly answer the comparable items may depend on how the instructor incorporates feedback into the testing procedure. Researchers should conduct future studies in order to determine how the timing affects learning retention. Researchers can accomplish this by taking the scantron with feedback and letting students compare their answers to the key after each individual item, after the entire test, or not at all using both comparable and identical items. This could help gather evidence that timing has a effect on scoring higher on comparable items.

The function of feedback could also account for the differences between the scantron with feedback and C/IF-AT. Researchers could examine if both testing formats fit Daniels and Daniels (2006) two function definition of feedback. Another way could be

to get students to complete both testing formats during a “think-a-loud”, asking each student to verbalize their internal behaviors. This would give researchers a good starting point to see if participants are using the feedback incorporated into each testing format differently. The reasons why the differences between the scantron with feedback and IF-AT perform differently is not currently known and is worthy of more research.

Another direction to take this study is an applied setting. As stated above, this is a lab-based study and the results might not generalize to a context where more variables are in place. It would be interesting to investigate if the scantron with feedback significantly outperforms the C/IF-AT in a classroom setting, where more complex variables exist. In addition, I would want to make sure that the CIF-AT and IF-AT are still comparable testing formats in the applied setting.

Researchers should conduct more research to look at how other phenomena can be incorporated into testing procedures to promote student learning. The current study specifically looked at putting feedback into answer-until-correct multiple choice testing procedures. People can insert feedback into other testing formats such as constructed response, matching, and fill-in-the-blank questions. In addition, other phenomena such as inserting a token economy into testing procedures could also help optimize student learning, as it has been successful in facilitating classroom behaviors in previous research (Kazdin & Bootzin, 1972).

Conclusion

Over the past few years, people have been researching new teaching formats that promote academic achievement in students, however new testing formats remain largely

unexplored. The current study showed evidence that testing procedures can have an effect on student learning. Incorporating phenomenon such as feedback into a testing procedure can increase students learning retention. However, how researchers implement feedback into the procedure seems to be important and could affect the extent to which students answer identical and comparable items across tests. Researchers should do further research to improve testing procedures in order to optimize student learning.

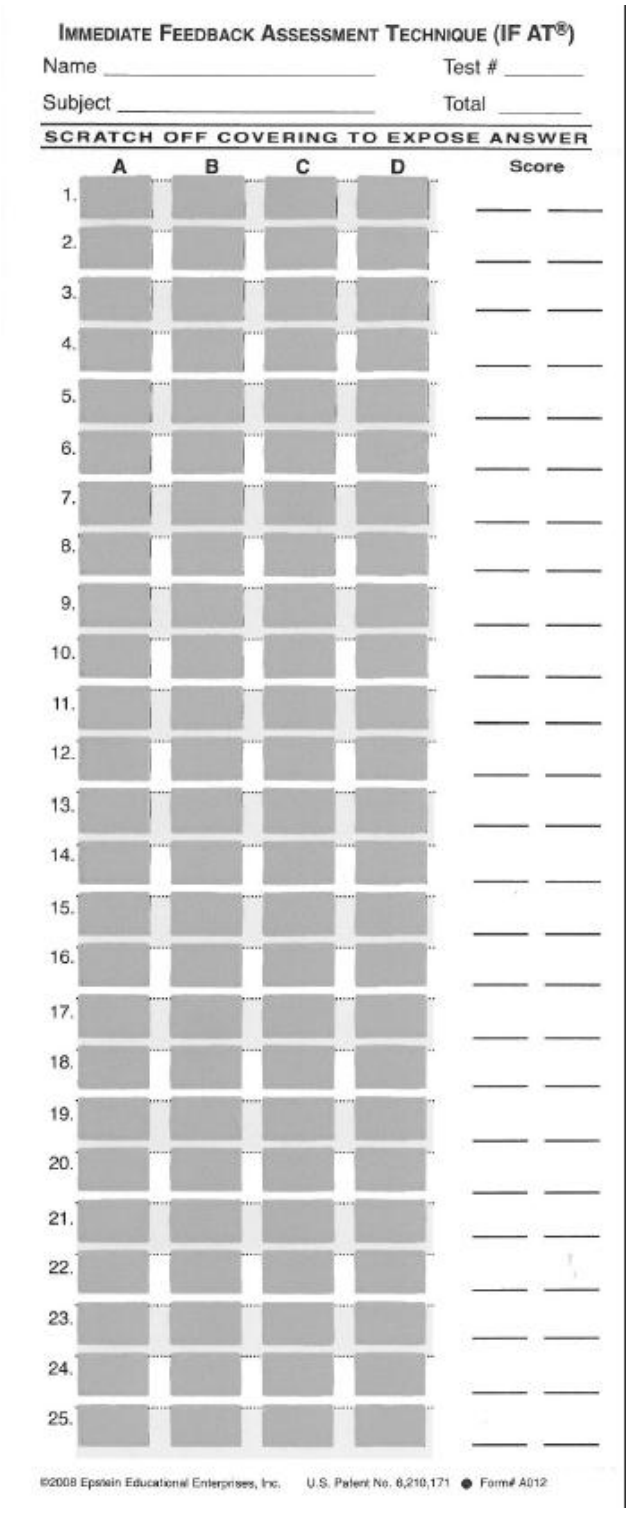
Table 1.

Descriptive statistics for Pre test and Post test scores across conditions.

		Pre test Mean	Pre test SD	N	Upper bound	Lower bound	Post test Mean	Post test SD	N	Upper bound	Lower Bound
Scantron		7.76	2.38	49	8.42	7.09	8.27	2.49	49	8.40	7.62
	Identical	38.44%	17.08%				38.44%	20.47%			
	Comparable	60.54%	20.55%				66.21%	21.51%			
Scantron with Feedback		8.44	2.53	48	9.12	7.77	10.92	2.152	48	9.09	10.27
	Identical	51.39%	19.40%				68.40%	18.91%			
	Comparable	59.49%	20.94%				75.69%	17.25%			
IFAT		7.92	2.18	50	8.57	7.26	9.58	2.467	50	8.56	8.94
	Identical	41.00%	20.54%				58.00%	20.00%			
	Comparable	60.67%	19.60%				67.78%	18.95%			
CIFAT		7.92	2.35	42	8.63	7.21	9.69	2.02	42	8.62	9.00
	Identical	42.86%	21.20%				54.37%	20.84%			
	Comparable	64.55%	20.86%				71.43%	14.13%			

Figure 1.

Immediate Feedback Assessment Technique (IF-AT) 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.						

©2008 Epstein Educational Enterprises, Inc. U.S. Patent No. 6,210,171 ● Form# A012

Figure 2.

Percent of items correct for each condition across time

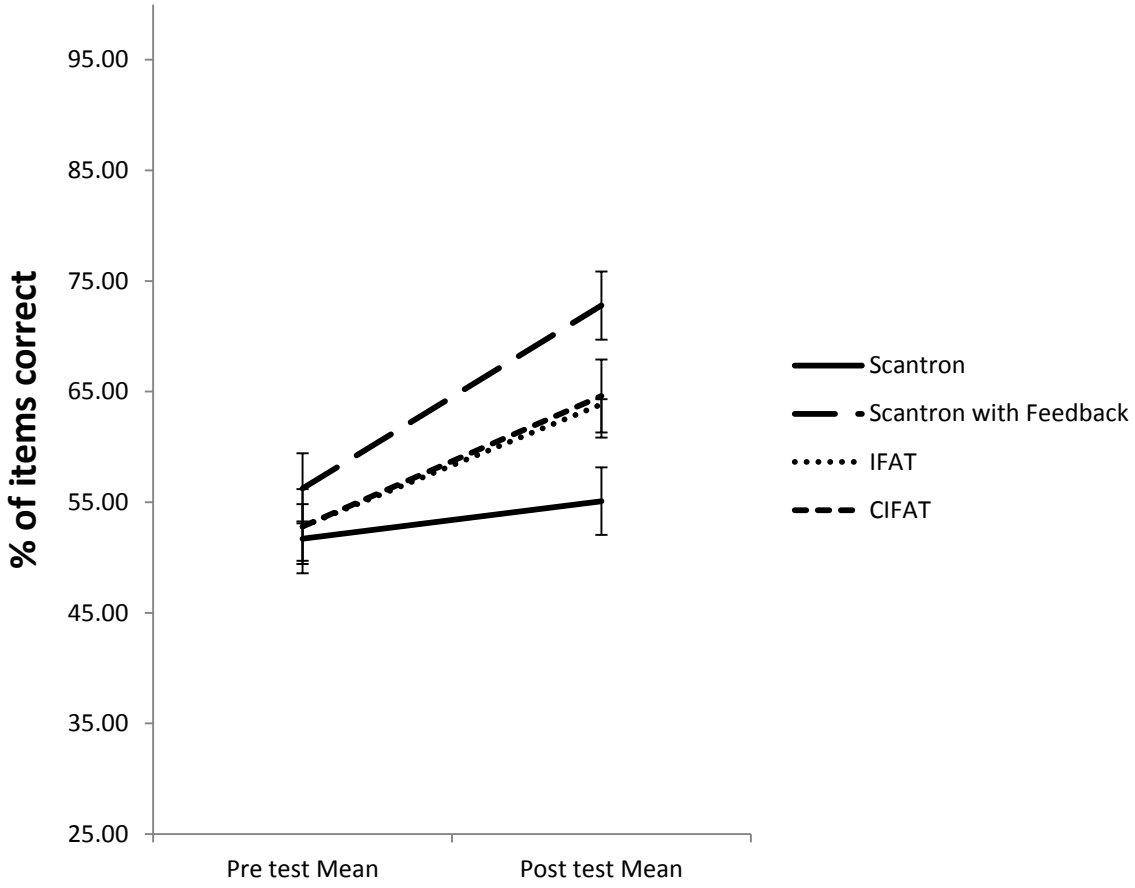
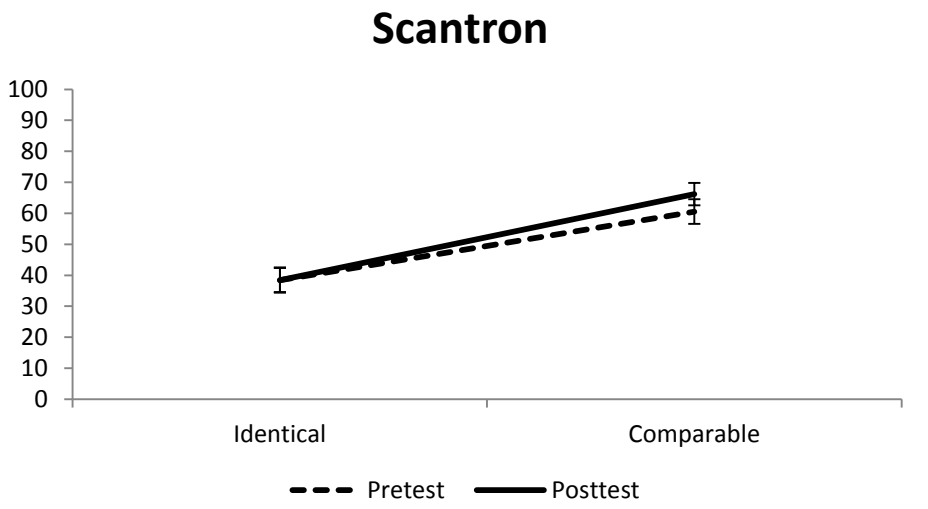
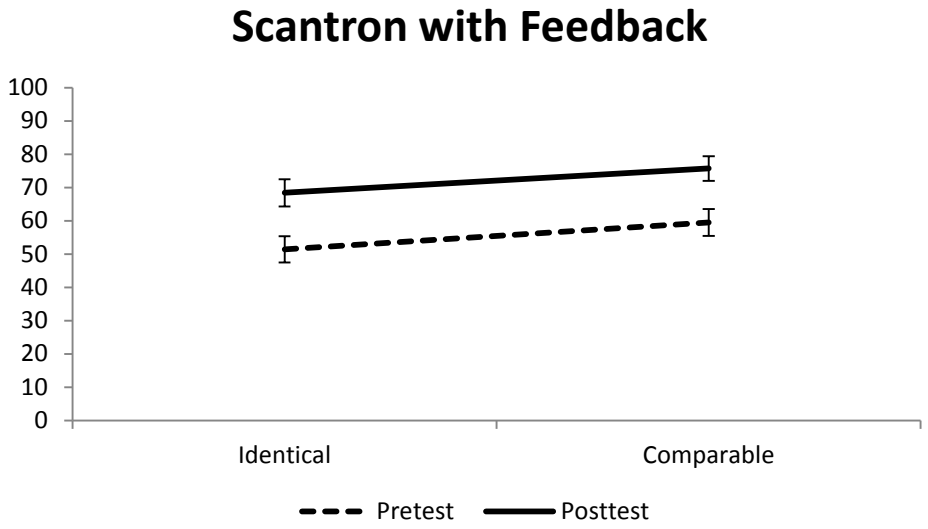


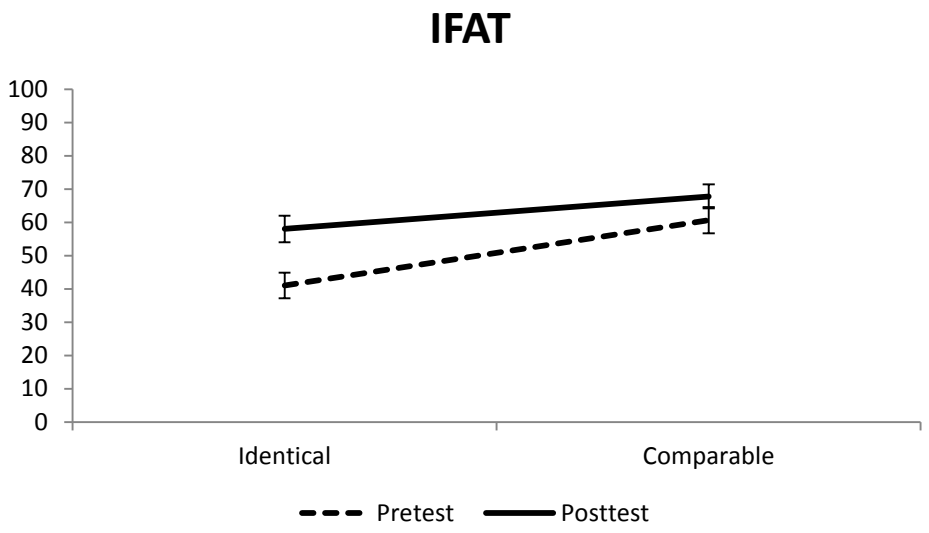
Figure 3.

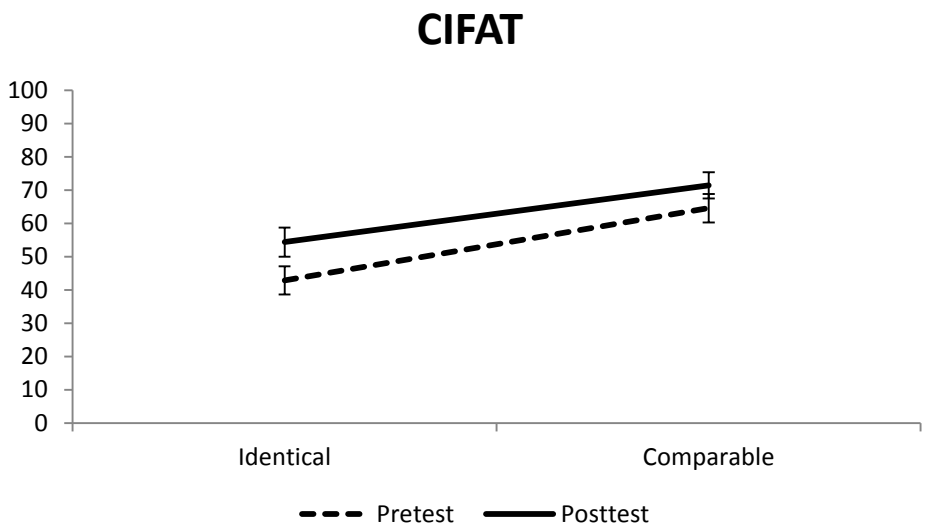
Percent of correct items by item type for each condition





88888


IFAT
100
90
80
70
60
50
40
30
20
10
0
Identical
Comparable
Pretest
Posttest


CIFAT
100
90
80
70
60
50
40
30
20
10
0
Identical
Comparable
Pretest
Posttest

V. Appendix A

Pre test Scantron Instructions

Welcome everyone. Today, you will be asked to read a 10-page chapter from a textbook, that I will pass out to you. You will be given 30 minutes to read the chapter. After the 30 minutes elapses, I will collect all chapters and distribute you a small quiz and a scantron form. When you receive the quiz, please put your ID number at the top of the Scantron form. You will have 15 minutes to complete the test and review the answers you have selected. In order to select your answers, please bubble the correct letter on the scantron that corresponds with the letter choice on the quiz. Make sure you read every question in the quiz carefully and accurately. When you return in a few days, you will take another quiz on this subject matter. For every correct question you get on the 2nd quiz, your name will be entered into a drawing for a number of gift certificates. Therefore, if you try hard on this initial test, you may be able to recall information better on the 2nd quiz. Remember, if you score all the items correct on the 2nd quiz, you have twice the chance of winning a prize as someone who only scores half of the items correct. Please do not write on the article or put your name on the Scantron form, only your ID number is required. At the end of the 15 minutes, I will collect your quizzes and you will be free to leave. Are there any questions?

VI. Appendix B

Pre test Scantron with Feedback Instructions

Welcome everyone. Today, you will be asked to read a 10-page chapter from a textbook, that I will pass out to you. You will be given 30 minutes to read the chapter. After the 30 minutes elapses, I will collect all chapters and distribute you a small quiz and a scantron form. When you receive the quiz, please put your ID number at the top of the Scantron form. You will have 10 minutes to complete the quiz. In order to select your answers, please bubble the correct letter on the scantron that corresponds with the letter choice on the quiz. Make sure you read every question in the quiz carefully and accurately. When you return in a few days, you will take another quiz on this subject matter. For every correct question you get on the 2nd quiz, your name will be entered into a drawing for a number of gift certificates. Therefore if you utilize the feedback and review the quiz with the correct answers, the more likely you are to get the material on the 2nd quiz correct. Remember, if you score all the items correct, you have twice the chance of winning a prize as someone who only scores half of the items correct. Please do not write on the article or put your name on the Scantron form, only your ID number is required. At the end of the 10 minutes, I will collect your responses and distribute you the correct answers to the quiz where you will be given 5 minutes to look over the correct answers. After you review your answers, I will collect your quizzes and you will be free to leave. Are there any questions?

VII. Appendix C

Pre test IF-AT Instructions

Welcome everyone. Today, you will be asked to read a 10-page chapter from a textbook, that I will pass out to you. You will be given 30 minutes to read the chapter. After the 30 minutes elapses, I will collect all chapters and distribute you a small quiz and a scantron form. When you receive the quiz, please put your ID number at the top of the Scantron form. You will have 15 minutes to complete the quiz. In order to select your answers, scratch off the waxy covering of the letter choice on the IF-AT that corresponds to the answer you want to select. Please be careful with the IF-AT, because a scratch on any portion on the form counts as an answer. If you get the answer correct, a star will be revealed. The star for the correct answer can appear anywhere within the box. It is not always in the same location; scratch the entire box. If you select an incorrect answer, a blank box will appear. If you get a blank box, please continue to scratch other choices until you get the answer correct. You can use any of a number of "tools" to scratch your answers; I recommend a fingernail, pen, or penny. Make sure you read every question in the quiz carefully and accurately. When you return in a few days, you will take another quiz on this subject matter. For every correct question you get on the 2nd quiz, your name will be entered into a drawing for a number of gift certificates. Since the IF-AT lets you scratch the boxes until the correct answer is revealed, it is in your benefit to utilize this feature. Remember, if you score all the items correct, you have twice the chance of winning a prize as someone who only scores half of the items correct. Please do not write on the article or put your name on the Scantron form, only your ID number is required.

At the end of the 15 minutes, I will collect your quizzes and you will be free to leave. Are there any questions?

VIII. Appendix D

Pre test Computerized IF-AT Instructions

Welcome everyone. Today, you will have 30 minutes to read a 10-page chapter from a textbook. You will be given 30 minutes to read the chapter. After the 30 minutes elapses, I will collect all chapters and distribute you a link that you will enter into your computer browser that will take you to a small quiz. When your page loads, please make sure you put your ID number in the designated box. You will have 15 minutes to complete the quiz. In order to select your answers, press and hold the right-click button and wave your mouse over the letter choice on the IF-AT that corresponds to the answer you want to select. If you get the answer correct, a star will be revealed. The star for the correct answer can appear anywhere within the box. If you select an incorrect answer, a blank box will appear. If you get a blank box, please continue to scratch other choices until you get the answer correct. Make sure you read every question in the quiz carefully and accurately. When you return in a few days, you will take another quiz on this subject matter. For every correct question you get on the 2nd quiz, your name will be entered into a drawing for a number of gift certificates. Since this computerized IF-AT lets you scratch the boxes until the correct answer is revealed, it is in your benefit to utilize this feature. Remember, Therefore, if you score all the items correct, you have twice the chance of winning a prize as someone who only scores half of the items correct. Please do not write on the article or enter you name anywhere on the computer, only your ID number is required. Again, you have 15 minutes to complete the quiz, once the 15 minute time elapses, you are free to leave. If you have not submitted the quiz within 15 minutes, the computer will submit the quiz automatically. Are there any questions?

IX. Appendix E

Post test instructions

Welcome everyone. Last week you read a 10 page chapter from a textbook. Today, you will take another quiz on the same material. I will distribute another short quiz and a scantron form. When you receive the quiz, please put your ID number at the top of the Scantron form. You will have 15 minutes to complete the test and review the answers you have selected. In order to select your answers, please bubble the correct letter on the scantron that corresponds with the letter choice on the quiz. Make sure you read every question carefully. Remember that on this quiz, your ID number will be entered into a drawing for a number of gift certificates. The more items you answer correctly, the more times your number will be entered into the drawing for gift certificates. For example, if you answer all of the items correctly, you will have twice the chance of winning a prize as someone who only answers half of the items correctly. Please do not put your name on the Scantron form, only your ID number is required. At the end of the 15 minutes, I will collect your quizzes and you will be free to leave, after collecting your debriefing form. Are there any questions?

X. References

Agnew, J. L. (1998). The establishing operation in organizational behavior management. *Journal of Organizational Behavior Management,* 18, 7-9.

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.

Alvero, A. M., Bucklin, B. R., & Austin, J. (2001). An objective review of the effectiveness of essential characteristics of performance feedback in organizational settings (1985-1998), *Journal of Organizational Behavior Management,* 21, 3-29.

Balcazar, F. E., Shupert, M. K., Daniels, A. C., Mawhinney, T. C. & Hopkins, B. O. (1989). An objective review and analysis of ten year os publication in the *Journal of Organizational Behavior Management*. *Journal of Organizational Behavior Management,* 10, 7-38.

Baldwin, J. D. & Baldwin, J. I. (2001) *Behavior Principles in Everyday Life*. Upper Saddle River, NJ: Prentice Hall.

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effectof feedback in test-like events. *Review of Educational Research, 61,* 213-238.

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

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing, 14*, 191-205.

Clariana, R. B., Ross, S. M., & Morrison, G. R. (1991). The effects of different feedback strategies using computer-administered multiple-choice questions as instruction. *Educational Technology Research and Development, 39(2),* 5-17.

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

Davis, B. G. (1993). *Tools for teaching.* San Francisco: Jossey-Bass.

DiBattista, D., Mitterer, J. O., & Goose, L. (2004). Acceptance by undergraduates of the immediate feedback assessment teachnique for multiple-choice testing. *Teaching in Higher Education, 9,* 17-28.

Epstein, M. L., Brosvic, G.M., & Costner, K. L. (2003). Effectiveness of feedback during the testing of preschool children, elementary school children, and adolescents with developmental delays. *The Psychological Record, 53,* 177-195.

Epstein, M. L., & Brosvic, G. M. (2002). Students prefer the immediate feedback assessment technique. *Psychological Reports, 90,* 1136-1138.

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.

Gaynor, P. (1981). The effect of feedback delay on retention of computer-based mathematical material. *Journal of Computer-Based Instruction, 8(2),* 28-34.

Gilman, D. A. (1969). Comparison of several feedback methods for correcting errors by computer-assessted instruction. *Journal of Educational Psychology, 60(6),* 503-506.

Kluger, A. N. & Adler. S. (1993). Person-versus computer-mediated feedback. *Computers in Human Behavior, 9*, 1-16.

Komaki, J., & Barnett, F. (1977). A behavioral approach to coaching football: Improving play execution of the offensive backfield on a youth football team. *Journal of Applied Behavior Analysis,* 10, 657-664.

Komaki, J., Barwick, K. D., & Scott, L. R. (1978). A behavioral approach to occupational safety: Pinpointing and reinforcing safe performance in a food manufacturing plant. *Journal of Applied Psychology, 63 (4),* 434-445.

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research, 47*, 211-232.

McKeachie, W. J. (2002). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers* (11th ed.). Boston: Houghton-Mifflin.

Mislevy, R. J. (1991). A framework for studying differences between multiple-choice and free-response test items. Cited from R. D. Bennett & W. C. Ward (1993), construction vs. choice in cognitive reassessment. Hillsdale, NJ: Erlbaum.

Narciss, S. & Huth, K. (2004). How to design informative tutoring feedback for multi-media learning. In Niegemann, M. H., Leutner, D. & Brunken, R. (Eds.) *Instructional Design for Multimedia Learning.* Munster: Waxmann.

Nolan, T., Jerema, K., &Austin, J. (1999). An objective review of the *Journal of Organizational Behavior Management*: 1986-1997. *Journal of Organizational Behavior Management*, 19, 83-114.

Phye, G. D., & Sanders, C. E. (1994). Advice and feedback: Elements of practice for problem solving. *Contemporary Educational Psychology*, 19(3), 286-301.

Pridemore, D. R., & Klein, J. D. (1991). Control of feedback in computer-assisted instruction. *Educational Technology Research and Development, 39(4)*, 27-32.

Pridemore, D. R., & Klein, J. D. (1995). Control of practice and level of feedback in computer-based instruction. *Contemporary Educational Psychology, 20*, 444-450.

Rummler, G. A., & Brache, A. P. (1995). *Improving performance: Managing the white space on the organizational chart.* San Francisco: Jossey-Bass.

Schneider, H. G., & Shugar, G. J. (1990). Audience and feedback effects in computer learning. *Computers in Human Behavior, 6*, 315-321.

Seligman, C., & Darley, J. M. (1977). Feedback as a means of decreasing residential energy consumption, *Journal of Applied Psychology*, 62, 363-368.

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

Sturges, P. T. (1978) Delay of informative feedback in computer-assisted testing, *Journal of Educational Psychology*, 70, 378-387.

Sulzer-Azaroff, B. (1978). Behavioral ecology and accident prevention. *Journal of Organizational Behavior Management, 2,* 11-44.

Sulzer-Azaroff, B. & Mayer, G. R. (1991). *Behavior analysis for lasting change.* Fort Worth. TX: Harcout Brace College.