THE EFFECTS OF iCLICKER BAR GRAPH FEEDBACK ON TEST PERFORMANCE

by

Elizabeth Driscoll Larson

A thesis submitted in partial fulfillment
of the requirements for the degree

of

Master of Science

in

Psychological Science

MONTANA STATE UNIVERSITY
Bozeman, MT

April 2017

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

This study examined the influence of receiving clicker bar graph feedback on immediate (Experiment 1) and delayed (Experiment 2) test performance. Participants received clicker questions with and without bar graph feedback or did not receive clicker questions during lecture. Overall, relative to the control group, clicker use enhanced immediate test performance. Clicker conditions did not differ in immediate or delayed test performance; however, when the majority of participants answered clicker questions correctly, test performance was enhanced only for those in the feedback group. When the majority answered clicker questions incorrectly, individuals in the no feedback group had marginally better test performance than those in the feedback condition, suggesting that bar graph feedback provokes source misattribution errors and interferes with one's ability to make corrections. Importantly, clicker use appears to enhance test performance, as long as individuals do not receive bar graph feedback when the majority answer clicker questions incorrectly. Thus, instructors should reconsider displaying bar graph feedback during clicker-based lectures. Finally, future research must explore the lack of delayed testing effects observed in this study.

INTRODUCTION

Audience Response Systems, also known as clickers, are hand-held devices commonly used in secondary education to enhance learning of lecture content (Kay & LeSage, 2009). Instructors typically implement such devices by displaying multiple-choice questions on a screen and asking students to use their clickers to select the correct answers. The instructor then provides bar graph feedback showing the frequency of selected choices so that students can compare their answer relative to others in the class. Finally, the instructor gives the correct answer to the question and may or may not discuss the reasoning behind the correct answer and why some individuals could have arrived at the incorrect answer (Hwang & Wolfe, 2010).

One general advantage of clicker use in classroom settings is its influence on student participation and engagement. Student engagement predicts academic achievement, persistent motivation, and scholastic commitment (Baker, Spiezio, & Boland, 2004; Marks, 2000; Shernoff & Hoogstra, 2001), virtues that may not be cultivated by traditional lecturing techniques. Trees and Jackson (2007) suggest that simply taking notes and silently observing the lecture decreases student motivation and, thus, academic performance. Further, offering incentives, especially when contingent on clicker performance, during clicker-based lectures enhances retention of lecture material via classroom engagement (Oswald & Rhoten, 2014).

Enhanced learning by means of active participation likely depends on how closely the properties of information encoding and retrieval match (e.g., Morris, Bransford, & Franks, 1977; Thomson & Tulving, 1970). Researchers have elaborated on the benefits

of initial retrieval by examining the *testing effect*, the finding that retrieving an item from memory has more beneficial effects for later retention than simply re-studying the item (Carrier & Pashler, 1992). Testing effects are generally more robust for final recall tests; however, Chan and McDermott (2007) examined effects of initial testing on final recognition tests and found that initial testing can increase recollection, but does not also increase familiarity, within final recognition. In this vein, receiving initial clicker questions during lecture forces students to retrieve the answer from memory, which could increase the likelihood of later answering corresponding test questions correctly.

The testing effect also appears to strengthen after a delay. For example, Roediger and Karpicke (2006a) had some students complete a recall test immediately after reading study passages and some students, instead, restudy the passages again. Students were given a final recall test on the information from the passages either five minutes, two days, or one week after the learning period. When students received the final test after five minutes, those who restudied the material had better test performance than those who were tested initially, but after delays of two days and one week those who took initial recall tests performed better. Additionally, repeated testing is thought to increase retention of information, relative to repeated studying, because the processes used during the initial retrieval of an answer are repeatedly produced, which assists performance when the same processes are required during a final test (Jacoby, 1978). After a delay, information previously learned will not be readily available or automatically produced due to interference and forgetting, requiring re-processing of that information in order to encode it at the level previously attained (Jacoby, 1978).

Learning can also be enhanced by a mode of active participation known as constructivism (Anderson, 1987), a learning theory that proposes that students learn more effectively when they must actively construct their own comprehension of course concepts. This notion is often implemented via presenting students with questions based on lecture content. When provided questions during class, students are acquainted with important points of the lecture, given the opportunity to reflect on what is known or not known of the content, and instructed to retrieve an answer which can increase future retrieval. Such questions produce better assessment performance when constructed conceptually rather than functionally (Mayer, 1975; Sagerman & Mayer, 1987).

Clicker use encourages all students, even those who tend to shy away from participation, to answer questions asked by the instructor during lecture. This active participation is usually accomplished by mandating clicker use as part of students' overall course grade. Thus, as student engagement appears to increase, student learning, in some cases, also improves. For example, Shapiro and Gordon (2012) examined effects of clicker-supplemented participation on test performance. Students were given typical lectures, each containing several clicker questions and bar graph feedback, over the course of a semester. Students were tested over the material at the end of the semester with an exam that consisted of questions on information that was previously assessed by clickers and information not assessed by clickers. Students correctly answered more exam questions that derived from clicker questions than questions that were not initially tested using clickers.

Perhaps the most accepted inference of why academic performance is enhanced in clicker-implemented courses is that of peer-to-peer and instructor-to-peer discussion. In an analysis of 56 studies, MacArthur and Jones (2008) found that clickers powerfully enhance learning course content only if they foster student discussion. Instructor led- and peer led-discussion are reportedly crucial in boosting conceptual comprehension (Smith et al., 2009) and important in cultivating student engagement and motivation (Boyle & Nicole, 2003). Active learning can be carried out by the presentation of clicker questions which, if implemented properly by the instructor, often leads to discussion of whether certain answers to questions are correct or incorrect.

McDonough and Foote (2015) examined active learning through student collaboration in order to explore effects of individual clicker use compared to shared clicker use in small groups of students. Students were allowed to discuss clicker questions in both conditions; however, students were more likely to collaborate and correctly answer questions when in small groups than when answering questions individually. Thus, sharing clickers in groups facilitated collaborative learning and performance, suggesting that using clickers is secondary to active learning via peer collaboration in enhancing test performance.

Although many studies have resulted in better test performance due to the use of clickers during class lectures (e.g., Kennedy & Cutts, 2005; Morling, McAuliffe, Cohen, & DiLorenzo, 2008; Shaffer & Collura, 2009), numerous studies have not found significant learning gains by utilizing clickers (e.g., Lasry, 2008; Miller, Ashar, & Getz, 2003). For example, Lasry (2008) examined conceptual learning and examination scores

of individuals who either used flashcards or clickers to answer lecture questions. Results indicated no significant differences in conceptual learning gains or test scores between groups, suggesting no additional performance benefit from using clickers during lecture. Miller et al. (2003) surveyed participants who answered lecture questions by clickers or verbal response in order to determine impressions of the lecture and knowledge of the material. Questionnaire responses revealed that those who used clickers rated the quality of the lecture and their level of attention during lecture higher than those who answered questions verbally; however, test performance did not differ significantly between participants who used clickers or answered verbally. In order to accept or refute the success of clickers in the enhancement of academic performance, reviews by both Kay and Lesage (2009) and Atlantis and Cheema (2015) recommend further examination of whether the benefits of active participation initiated by clickers are superficial or actually indicative of learning.

Anthis (2011) sought to determine whether the positive outcomes of test performance in clicker studies is a result of clickers themselves or simply the questions asked. Participants received periodic questions throughout the semester that were similar to subsequent exam questions. Participants either answered lecture questions via clickers or hand-raising. Those who used clickers saw the correct answer displayed on a screen, and those who raised their hands were verbally told the correct answer by the instructor. Anthis concluded that using clickers to answer lecture questions did not enhance test performance above and beyond hand-raising, suggesting that as long as questions are asked, the mechanism of questioning does not matter. Similar effects have also been

demonstrated when comparing clicker use to low-tech answering methods like paper answer cards (Desrochers & Shelnutt, 2012; Fallon & Forrest, 2011). Although conclusions on the effectiveness of clicker use on test performance are mixed, Anthis' (2011) study provides insight that providing students with questions during class, whether implemented by clickers or other means, might account for successful academic achievement.

An ironic phenomenon appears in many of the clicker studies that show no effect on enhanced testing performance; students tend to report an enjoyable experience and a positive perspective on learning outcome when using clickers, despite equal or worse test performance, compared to students who answered class questions by different means or who simply did not answer questions during class. For example, Morgan (2008) reported that studies finding minimal increases in For example, Morgan (2008) reported that studies finding minimal increases in test scores in clicker-based classes also had the majority of students reporting higher levels of enjoyment compared to other treatment and control conditions. Similarly, Sutherlin, Sutherlin, and Akpanudo (2013) found that clickers were not a predictor of student achievement, even though the majority of students felt that they learned more when using clickers. Further, Sun, Martinez, and Seli (2014) compared a traditional clicker-based course to a web-based polling course in which students answered questions provided by the instructor before and after class. Students in the web-based polling course had higher levels of emotional and cognitive engagement than individuals in the traditional clicker-based course even though they

reported lower levels of self-efficacy for learning. These students most likely displayed higher academic achievement, but had a decreased perception of their performance.

Additionally, it could be that the type of feedback presented to students determines later test performance. Roediger and Karpicke (2006b) note that if feedback is given, it must be extensive and purposeful; however, it is unclear how helpful receiving bar graph feedback actually is to test performance. For example, it might be that the act of testing influences learning via consequential means, like implementing feedback to correct misconceptions or solidify answers. How conducive is feedback to learning and test performance, however, if such feedback happens to be misleading?

Several clicker studies have examined the benefits of receiving bar graph feedback of how other students answer clicker questions (e.g., Lantz & Stawiski, 2014; Oswald & Rhoten, 2014). Oswald and Rhoten (2014) sought to verify whether feedback increases learning beyond simply active engagement. All students received the correct answer after responding to clicker questions, but only half of the students received bar graph feedback displaying how students answered each question. Oswald and Rhoten hypothesized that students who received bar graph feedback would perform better on a subsequent quiz than students who only received the correct answer, due to the opportunity to see how well they measured up to their peers and how the majority of the class answered. Results revealed that presenting bar graph feedback increased quiz scores, but the authors speculated that this finding might be attributed to enhanced retrieval of answers that were both initially correct and chosen by the majority of students during the clicker assessment.

Similarly, Lantz and Stawiski (2014) hypothesized that receiving only bar graph feedback (but not the correct answer) would enhance memory for lecture material compared to conditions receiving feedback only of the correct answer and receiving no questions throughout the lecture (no feedback). Students in both feedback conditions performed better on a subsequent test than students who did not receive questions during the lecture; however, the feedback groups did not differ significantly in test performance. However, Lantz and Stawiski noted that the majority of students answered nearly every clicker question correctly, so students who did not initially receive feedback of the correct answer based on the most frequently selected answer displayed on the histogram. This speculation suggests that students are likely to assume that the most common answer to a clicker question is also the correct answer, but this inference could prove inaccurate if the majority of the class answers incorrectly.

To my knowledge, no studies have explored the possible impedance of bar graph feedback on test performance in clicker-based lectures. For example, it is possible that students receiving bar graph feedback could incorrectly remember the most common answer (if also incorrect), even if they were given the correct answer by the instructor. Thus, receiving bar graph feedback indicating how other students answered could jeopardize test performance because, rather than an increase in correct retrieval, these individuals could have an increase in false retrieval due to misattributing such memory to an incorrect source (i.e., an incorrect majority response received via bar graph feedback) (McCabe & Geraci, 2009).

To understand the reasoning behind this interpretation, source monitoring theory (Johnson, Hashtroudi, & Lindsay, 1993) must be considered. Source monitoring is the ability to remember specific details about the source of a memory. In the event of a source monitoring error, for example, people might forget where an event took place or to whom they told a bit of information. Such errors are more likely to occur when recollection of the memory is vague, when discriminability between potential sources is low, and when recollection is made rapidly with non-deliberate decision making (Lindsay, 1990).

The rate at which source monitoring errors occur could be motivated by exercising the distinctiveness heuristic (Dodson & Schacter, 2002; Schacter, Israel, & Racine, 1999), a technique that aids in evaluating which items in recollection are part of an actual experience. Items that are recollected with vivid details are assumed to have occurred within an experience, but items retrieved without recognizable characteristics are typically interpreted as being novel. Source monitoring errors can occur when an individual erroneously encodes a novel bit of information as existing knowledge by confusing new details with vivid details of a former experience (Dodson & Schacter, 2002; Schacter, Israel, & Racine, 1999).

McCabe and Geraci (2009) assessed the source misattribution account by investigating whether items that were presented to participants before receiving a list of to-be-remembered words could be cued by items in a recognition test. Source misattributions occur when an individual falsely remembers an item due to recollection of items or item features from a context other than the study list. Thus, McCabe and

Geraci predicted that that participants would falsely misattribute pre-study items as study items. Further, they suspected that pre-study items would more likely be misattributed as study items when pre-study and study processing are similar (e.g., both tests requiring gender judgement) than when less similar (e.g., one test requiring a gender judgement; the other test requiring a pleasantness judgement). In three experiments, McCabe and Geraci first exposed participants to pre-study items two days prior to administering the study list, and both lists were processed in a dissimilar manner. In a second experiment, participants processed pre-study items, presented immediately before the study list, in a similar manner as study items. In the final experiment, participants received pre-study items two days before the study list was administered; however, in contrast to the first experiment, half of the participants processed both lists similarly, whereas the other half of participants processed both lists in a dissimilar manner. McCabe and Geraci's (2009) predictions were supported, such that participants erroneously remembered more items that were presented in pre-study lists than novel items that were not previously processed. Additionally, when pre-study and study lists were processed in similar manners, relative to dissimilar manners, false remembering increased. These findings suggest that misattributing sources can contribute to incorrectly remembering an item, a bit of information, or an event.

Additionally, receiving misleading information after viewing an event can result in source monitoring errors as measured by subsequent memory tests. For example, in a series of studies conducted by Loftus, Miller, and Burns (1978), participants were shown 30 picture slides depicting a car hitting a pedestrian and were then given either an

accurate description of the event or an account detailed with misleading information. Loftus and colleagues found that participants who received the misleading account of the viewed event produced more source monitoring errors on the memory test, suggesting that participants had a difficult time knowing whether they were remembering a detail from the slides or from the misleading description.

The increase in source monitoring errors due to misleading accounts was further assessed by Meade and Roediger (2002). In a series of four experiments, they sought to determine whether factors of false memory are generalizable to the social contagion paradigm (Roediger, Meade, & Bergman, 2001), a manifestation in which a participant adopts false recollections provided by a confederate. Results from Meade and Roediger's study, particularly those that are relevant to my study, revealed that even when participants were reminded that their partner (confederate) could have made a recollection mistake during the joint recall session, they still recollected the incorrect information presented by their partner, and even attributed that information to the scene rather than their partner. These findings suggest that, under certain conditions, social interactions can have a powerful effect on recognition and source monitoring.

The underlying premise of source monitoring errors produced in such experiments is that participants experience optimal conditions that hinder differentiating the incorrect source of their memory from the correct source. For example, a participant of Loftus et al.'s (1978) study might have incorrectly remembered seeing a stop sign because the false description of the event mentioned a stop sign, even though the participant actually received the version of the slides that displayed a yield sign. Similarly, receiving a bar

graph showing the actual answer to a clicker question could be comparable to the viewing of the slides in Loftus et al.'s study; whereas receiving bar graph feedback of how other students answered could be analogous to the commentary following the viewing of the slides. For example, if students receive bar graph feedback of a clicker question showing that the majority of students in the class answered 'A' and are then presented with a bar graph displaying the correct answer 'B', students have two potential memory sources (i.e., graphs) when retrieving the answer to that question on the subsequent test. In contrast, students who only receive the bar graph showing the actual answer, but not the bar graph feedback, have just one source of their memory when retrieving the answer to that question on the test. Therefore, memory retrieval during test might be hindered by multiple sources offered to the student (i.e., majority answer and correct answer), creating confusion and possibly poorer performance.

After a delay, effects of source monitoring errors are thought to worsen. For example, Underwood and Pezdek (1998) found that participants are more likely to recognize or retain errors in source monitoring one month after initial encoding. Similar to Loftus et al.'s (1978) design, participants viewed slides of two different scenes and then read subsequent narratives, either from a credible source (memory psychologist) or a low-credible source (four year old), that included misinformation describing the previously viewed scenes. After a delay, a recognition test revealed that participants recognize more misinformation, especially when told that the narrative was from a four year old, than when tested immediately after receiving the narratives. Underwood and Pezdek (1998) explain their findings by the availability valence hypothesis which

demonstrates that people view low-credible sources as more valid over time because the source (four year old) and message (narrative) become less associated, so it is more difficult to determine whether the information that is remembered is likely correct or incorrect; however, under a short delay, it is easier to remember that the misinformation received is from a four year old and should, therefore, be less trusted (Mazursky & Shul, 1987). Thus, this misinformation is more likely to remain in memory over time, compared to correctly placed information, and the source of that misinformation is likely to be misinterpreted.

The availability valence hypothesis is also commonly referred to as the sleeper effect, a classic phenomenon first reported by Hovland, Lumsdaine, and Sheffield (1949). Half of the soldiers who participated in Hovland et al.'s study watched a World War II propaganda film, and the other half did not. Relative beliefs of all participants on this matter were measured at five days or nine weeks. Results indicated that differences in beliefs between groups increased between five days and nine weeks, suggesting that soldiers who watched the propaganda film immediately realized the bias, but after nine weeks, these soldiers might have forgotten the source, yet remembered the persuasive information presented in the film.

Accordingly, after a delay, students might be more likely to answer test questions by recalling information from the incorrect and less-credible source; bar graph feedback. Over time, students are less able to distinguish details between the bar graph of the actual results and the bar graph of feedback, causing confusion in selecting the source of their recollection. Further, information from the low-credible source (student answers) is

more likely to be determined as correct compared to information from the credible source (instructor) after a delay as opposed to immediately after receiving the lecture. In contrast, if tested immediately after receiving the lecture, students might make fewer source monitoring errors because they are better able to disassociate the correct answers presented by the instructor from the displayed answers of fellow students.

Frost, Ingraham, and Wilson (2002) explain this effect of delay on source monitoring errors by noting that over a short interval, the differences between the verbal information from the narrative and the observable details of the slides is more noticeable, but as the time interval increases, one tends to forget many of the details related to the slides. The forgetting of perceptual details makes distinguishing between sources more difficult and mistaking information from the narrative as information from the slides more likely. Additionally, Zaragoza and Lane (1994) argue that, when given a narrative after viewing slides of a scene, participants tend to visualize the events of the narrative. Even after short delays, constructing visual details of misinformation increases the likelihood of committing source monitoring errors.

In clicker-based lectures, students typically receive the correct answer to the question as well as bar graph feedback. The correct answer acts as information that is critical for the students to remember; however, students also receive information in the form of bar graph feedback that is not necessary to remember. The recollection for the correct answers, therefore, will likely reduce across a delay, but the unstudied (unnecessary) bar graph information will likely remain salient when the student must recall the material during a subsequent test. Thus, students might mistakenly choose the

most common answer viewed via bar graph feedback as the answer to test questions. Such a mistake could prove favorable if the majority of students answer the initial clicker question correctly or detrimental if the majority of students answer the clicker question incorrectly. Realistically, it is not uncommon for students to answer initial clicker questions incorrectly because, assumingly, the information taught in lectures is novel and purposefully challenging in order to foster more advanced learning. In this case, then, receiving bar graph feedback could be harmful to test performance. Thus, this study examined the influence of receiving clicker questions, with and without bar graph feedback, on subsequent test performance.

EXPERIMENT 1

Hypotheses

Based on previous research suggesting that supplementing lecture content with clicker questions can enhance subsequent test performance on such material via increased participation and testing effects, my first hypothesis was that participants who received clicker questions would have enhanced test performance. Further, evidence suggests that receiving bar graph feedback after answering a clicker question can possibly negatively influence later responses to that same question. Therefore, my second hypothesis was that individuals who answered clicker questions and received the correct answer, but did not receive bar graph feedback, would have better test performance because these individuals would not only benefit from active participation, but would have less sources (i.e., graphs) to monitor for the subsequent test. Specifically, individuals who received bar graph feedback might misremember the majority response observed on this graph, rather than the graph of the actual results. This could improve performance on test questions in which the majority of individuals answer initial clicker questions correctly, but impair performance on test questions in which the majority of the individuals answer initial clicker questions incorrectly. Thus, my third hypothesis was that test performance would be hindered for those who received bar graph feedback when the majority of individuals answered initial clicker questions incorrectly, due to misattributing the bar graph feedback as the correct answer on corresponding test questions. Should this occur, the average proportion of correct test answers of individuals who received bar graph

feedback would be less than that of individuals who did not receive bar graph feedback, because initial majority response would have no effect on those that did not view others' responses. In contrast, the average proportion correct should be greater for those who received bar graph feedback when the majority of individuals answered the initial question correctly, because both sources would point to the same answer.

Participants and Design

One hundred thirty-three Montana State University undergraduates who reported owning iClickers (i>clicker, 2016) participated to obtain Introduction to Psychology research credits. The study used a between-subjects design and was constructed to mock a live classroom setting. Conditions were randomly assigned to participants depending on the day that they voluntarily chose to participate. In total, 45 received the feedback condition (exposure to clicker questions, the correct answer to each question, and bar graph feedback), 44 participants received the no feedback condition (exposure to clicker questions and the correct answer to each question, but not bar graph feedback), and 44 participants received the control condition (no exposure to clicker questions or feedback). On average, "classes" consisted of approximately 20 students each.

Materials

Two 20-minute versions of a Microsoft PowerPoint (PowerPoint, 2013) lecture were created to accommodate the different conditions. Both versions were identical in content and design, except that both clicker conditions also included slides that displayed

a clicker question immediately after descriptions of each study. Dr. Keith Hutchison narrated the lecture, entitled "Effects of attentional control, list content, and item-type on semantic priming," which consisted of six studies on psycholinguistics. This lecture topic was used based on the assumption that most Introduction to Psychology students have not been formerly exposed to this information, thereby normalizing the general level of understanding. Additionally, the version of the lecture used in the clicker conditions implemented REEF polling (REEF Polling by i>clicker, 2016), an iClicker software add-in which collects and records participants' answers to clicker questions. The REEF polling add-in was also used to display bar graph feedback of participants' answers to clicker questions in the feedback condition only. Both the bar graph feedback and actual result slides contained bar graphs with either only blue bars (feedback) or both blue and yellow bars (actual results).

Immediately after the lecture, participants in each condition received a 10-minute filler-task consisting of one multiplication worksheet and one long division worksheet. Participants were told to complete all of the math problems as accurately as possible in the allotted time. After the filler-task, participants in each condition took a test based on the lecture. The test, worth 6 points, consisted of six multiple choice questions derived directly from the clicker questions given to participants in the feedback and no feedback conditions. Answers to the questions on the test were verbatim to corresponding answers of clicker questions. The order of questions and answers on the test, however, were randomized as to not follow the lecture chronologically.

Procedure

After completing consent forms, participants were told that they would receive a lecture on psycholinguistics and instructed to pay close attention to the content as they would be tested on the material. Participants in the control condition simply watched and listened to the lecture and did not partake in any classroom participation. Participants in both clicker conditions watched and listened to the lecture, but they were also given identical multiple choice clicker questions; one question following each of the six lecture topics. Each topic described a study, and participants were asked to make a prediction as to how the previously described study resulted. Questions were formatted this way to engage conceptual, rather than functional, thinking to elicit an optimal learning experience (Mayer, 1975; Sagerman & Mayer, 1987). Participants were given as much time as needed to choose an answer; however, participants typically answered within 10 seconds of receiving the questions and choices. Participants indicated that they made a selection by placing their clickers on their desks. Those in the clicker conditions submitted their predictions to each clicker question, but only individuals in the feedback condition were shown bar graph feedback of how participants in the class answered each question. Participants in all three conditions were shown a bar graph displaying the actual results (i.e., the correct answer); those in the control condition were given the actual results immediately after the description of each study, those in the no feedback condition were given the actual results immediately after answering the clicker question, and those in the feedback condition received the actual results immediately after

answering the clicker question and viewing the bar graph feedback. Once the lecture ended, participants in each condition completed a 10-minute filler-task.

Following the filler-task, participants were given a 6-item multiple choice test based on the six studies described in the lecture. Each question, consisting of a brief description of the study, asked participants to choose the best answer explaining the final result of that study. Participants had approximately 25 minutes to complete the test; however most participants finished the test within 10 minutes. Once finished, participants submitted their tests and were debriefed of the experiment.

Results

My first hypothesis was that test performance would be better for participants who received clicker questions, due to increased engagement, than for participants who simply listened to the lecture. Secondly, in comparing both clicker conditions, I hypothesized that individuals who received bar graph feedback might perform worse on the subsequent test than individuals who only received the correct answer because of the additional graphs viewed, resulting in source confusion when recalling answers during the test. Although the proportion of correct test answers for each condition presented in the predicted direction (see Figure 1), a one-way analysis of variance (ANOVA) showed no significant effect of condition on test performance, $F(2, 130) = 1.287$, $p = .280$, $\eta_p^2 = .019$. There was, however, a marginal difference in test performance between the control and no feedback conditions, $t(86) = 1.66$, $p = .099$.

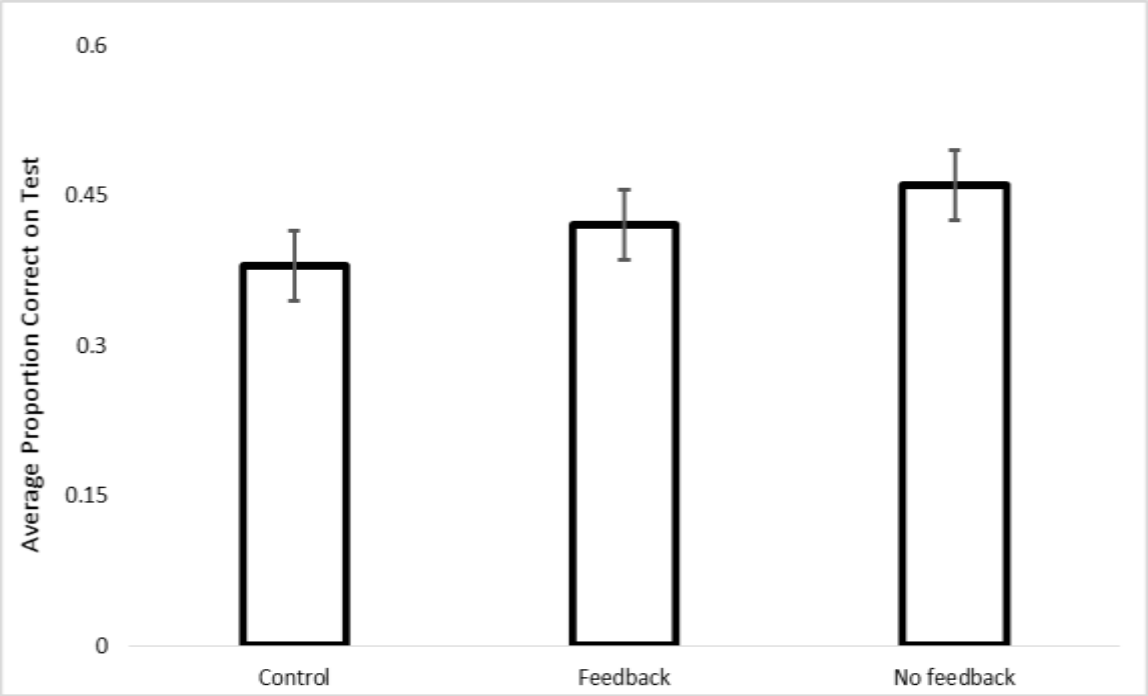

Figure 1. Average proportion of correct test answers for each clicker condition. Error bars reflect standard error of the mean.

A 2 (feedback, no feedback) x 2 (majority, non-majority) ANOVA assessed my third hypothesis that test performance would be hindered for those who received bar graph feedback when the majority of individuals answered initial clicker questions incorrectly, due to source misattribution errors, and would benefit from receiving bar graph feedback when the initial majority response was correct. A marginal interaction supported my hypothesis, $F(1, 87) = .382$, $p = .079$, $\eta_p^2 = .035$, such that individuals who received bar graph feedback performed non-significantly 7.2 ± 10.3% better on the test when the majority of participants answered the initial clicker questions correctly (see Figure 2), $t(44) = -1.502$, $p = .140$, whereas those who received no bar graph feedback performed non-significantly 5.9 ± 10.4% worse when the majority of participants initially

answered the clicker questions correctly, $t(43) = 1.049$, $p = .301$ [± = 95% confidence interval].

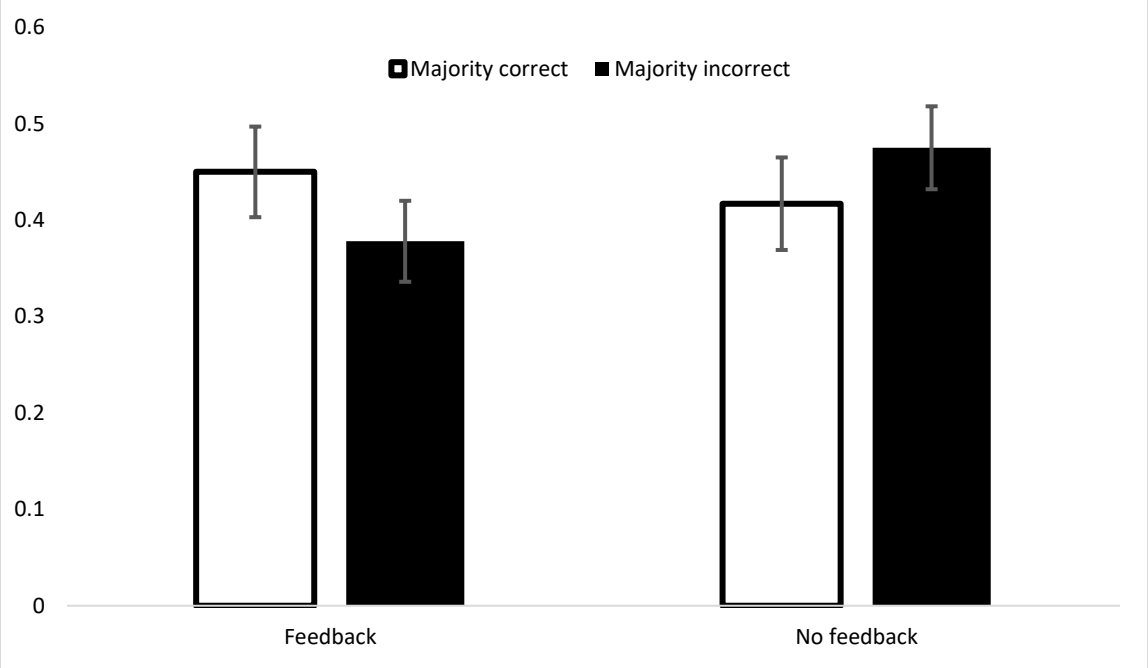

Figure 2. Average proportion of correct test answers of each condition by majority correct and incorrect clicker responses. Error bars reflect standard error of the mean.

Discussion

Hypotheses 1 and 2 were not statistically supported, such that individuals who received clicker questions did not perform significantly better on the test than individuals in the control condition, and individuals without bar graph feedback did not perform significantly better on the test than individuals who received bar graph feedback; however, those who did not receive bar graph feedback trended toward better test performance than individuals in the control condition. Moreover, the predicted pattern of test performance was observed, numerically. My third hypothesis was marginally

supported, such that individuals who received bar graph feedback had non-significantly better or worse performance than those not receiving feedback, depending upon whether the initial majority response was correct or incorrect, respectively. Finally, overall test performance was substantially low (M = .42, SE = .03) but average test scores were still above chance level (.25) for the control (M = .38, SE = .03), $t(43) = 3.94$, p = .000, feedback (M = .42, SE = .04), $t(44) = 4.37$, p = .000, and no feedback (M = .46, SE = .03), $t(43) = 6.28$, p = .000, conditions.

As noted above, individuals who received bar graph feedback appeared to benefit from this feedback only when the majority of participants answered initial clicker questions correctly. This benefit, although non-significant, suggests that viewing other students' answers might be beneficial for test performance only when the majority of students answer clicker questions correctly. Viewing clicker results in which the majority of students answer incorrectly could possibly hamper test performance. The marginal interaction between initial majority response and feedback condition, as well as the generally low test scores, influenced the procedural additions of Experiment 2.

Because evidence suggests robust testing effects after a delay (e.g., Jacoby, 1978; Roediger & Karpicke, 2006a), a 48-hour delay condition was included in Experiment 2. Additionally, average test scores in Experiment 1 were generally low, so, following Oswald and Rhoten (2014), a monetary incentive was offered to participants in Experiment 2 in order to increase motivation to perform well. Participants received a motivation questionnaire which measured self-reported effort and importance in regard to comprehension of lecture content and test performance.

EXPERIMENT 2

Hypotheses

Hypotheses 1, 2, and 3 were identical to those in Experiment 1; however, I included an additional hypothesis in Experiment 2 in order to account for effects of delay. Therefore, my fourth hypothesis posited that effects of condition should be greater after a delay. Consistent with effects of testing (Jacoby, 1978; Roediger & Karpicke, 2006a), individuals who received clicker questions should have better delayed test performance than individuals in the control condition, because answering clicker questions requires retrieval, and prior retrieval should result in more accessible retrieval for corresponding test questions after a delay. Further, evidence suggests that source confusion is usually greater following a longer interval between processing and retrieval (e.g., Underwood & Pezdek, 1998). Accordingly, individuals who received bar graph feedback should perform worse on the delayed test than individuals who did not receive bar graph feedback, due to increased source monitoring errors. Furthermore, assuming that the clicker questions were rather difficult, the majority of individuals were expected to usually answer incorrectly. In fact, of the majority responses to clicker questions in Experiment 1, 3.57 (*SE* = .073) out of 6 clicker questions (59.5%) were answered incorrectly, on average, whereas 2.43 (*SE* = .073) out of 6 clicker questions (40.5%) were answered correctly. This difference was significant, $t(176) = 11.10$, $p = .000$. Because of this, the average proportion of correct test answers of individuals who received bar graph

feedback was predicted to be lower after a delay than that of individuals who did not receive bar graph feedback.

Finally, I suspected test performance to increase from Experiment 1 to Experiment 2 due to monetary incentive. It is possible that testing only has an effect among those who care about performing well. Enhanced test performance, in turn, might correlate with greater self-reports of effort and importance, reflecting an accurate correspondence of self-competency and actual achievement. Test performance and self-reported motivation could, however, associate negatively or non-significantly, reflecting a division between perceived and actual achievement.

Participants and Design

Two hundred seventy-six Montana State University undergraduates who reported owning iClickers participated to obtain Introduction to Psychology research credit. The design of Experiment 2 mirrored Experiment 1; however, half of the participants were randomly assigned to a 48-hour delay condition in which these individuals were released after receiving the lecture and returned two days later to complete the test. Participants who received the 48-hour delay, therefore, did not receive the filler-task. The other half of the participants were randomly assigned to an immediate condition in which participants followed the same general procedure as Experiment 1. All participants were offered a monetary reward for high test performance. Additionally, all participants received a motivation questionnaire immediately after completing the test. Experiment 2 consisted of six groups: immediate control (n = 49), immediate feedback (n = 42),

immediate no feedback (n = 43), delay control (n = 48), delay feedback (n = 48), and delay no feedback (n = 46). We used a 3 (control, feedback, no feedback) x 2 (immediate, 48-hour delay) mixed factorial analysis of variance to observe effects of both condition and delay on test performance.

Materials

Materials were identical to Experiment 1; however, all participants were told within the consent form that the individual with the highest test score would receive 50 dollars or, in the case of a tie, those highest-scoring individuals would be put into a drawing to determine the lucky winner. Additionally, all participants received the Student Opinion Scale (SOS) (Sundre, 2007), a 10-item self-reported motivation instrument, after the test to assess individual reports of exerted effort and perceived importance during the lecture. The SOS has high internal validity and is especially useful when administered after low-stakes tests, such as that used in this study, which pose minimal threat or personal consequences to test takers. The highest score possibly obtained is 25 on both subscales (i.e., effort and importance). The higher an individual scores on either subscale, the higher that individual perceives his or her exertion of effort or sense of importance during the lecture and test (Sundre, 2007).

Procedure

The procedure mirrored that of Experiment 1; however, participants received the SOS immediately after taking the test and before debriefing. Those participants in the

48-hour condition were released after receiving the lecture and returned 48 hours later to take the test, complete the SOS, and receive debriefing.

Results

A 3 (condition) x 2 (time) mixed factorial analysis of variance assessed my first, second, and third hypotheses. Hypotheses 1 and 2 posited that individuals who received clicker questions would have higher test performance, on average, than individuals who did not receive clicker questions, and, secondly, of individuals who received clicker questions, those who did not receive bar graph feedback would have higher test scores, on average, than those who received bar graph feedback. The overall main effects of condition, $F(2, 270) = 4.921$, $p = .008$, $\eta_p^2 = .035$, and time, $F(1, 270) = 18.703$, $p = .000$, $\eta_p^2 = .065$, were significant. Hypothesis 1 was supported, such that, relative to the control group, test performance was marginally ($p = .075$) and significantly ($p = .012$) better for individuals who received clicker questions with bar graph feedback and without bar graph feedback, respectively. Further, my second hypothesis was not supported, such that test performance did not differ significantly between bar graph feedback and no bar graph feedback conditions, $p = .777$. Despite a non-significant difference in test performance between clicker conditions, all three conditions again presented numerically in the predicted direction (see Figure 3).

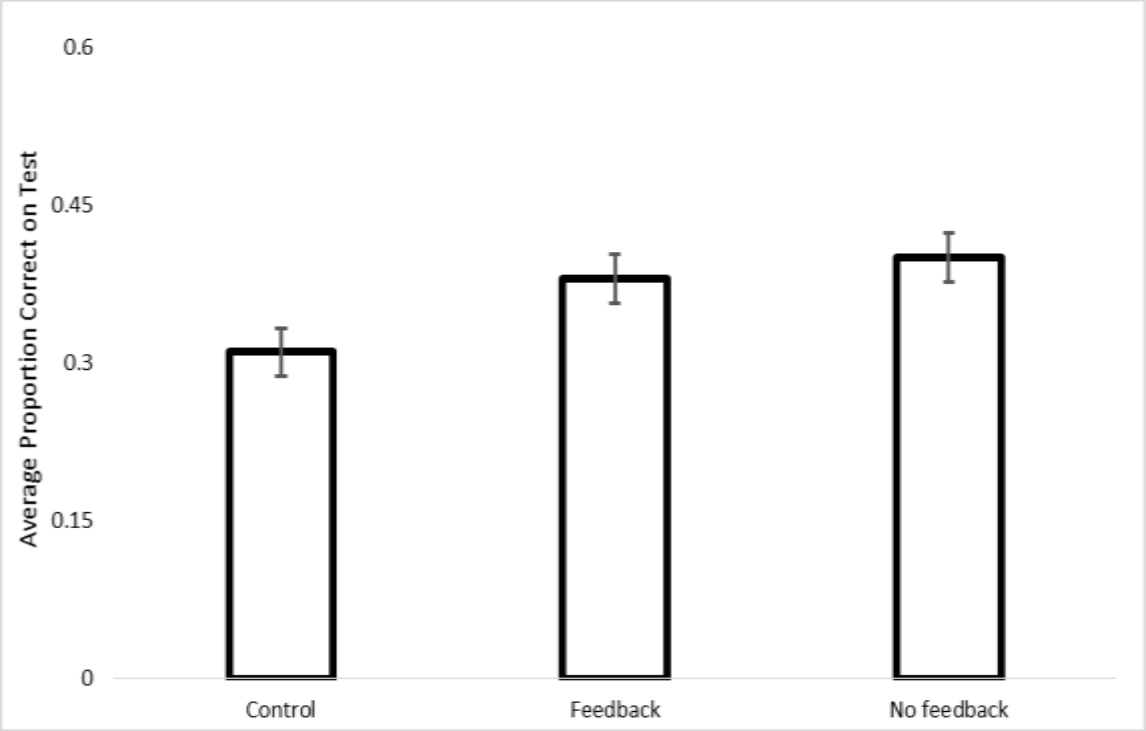

Figure 3. Average proportion of correct test answers of each clicker condition in Experiment 2. Error bars reflect standard error of the mean.

A 2 (bar graph feedback, no bar graph feedback) x 2 (time) x 2 (majority correct, majority incorrect) mixed factorial analysis of variance assessed my third hypothesis that test performance would be hindered for those who received bar graph feedback when the majority of individuals answered initial clicker questions incorrectly, but that initial majority response would have no effect for those who did not receive bar graph feedback. As in Experiment 1, this hypothesis again received marginal support, $F(1, 175) = 3.265$, $p = .072$, $\eta_p^2 = .018$ (see Figure 4). Specifically, individuals who received bar graph feedback performed marginally 7.7 ± 8.1% better when the majority of individuals answered initial clicker questions correctly as opposed to incorrectly, $t(89) = -1.82$, $p = .072$. In contrast, there was no effect of initial majority response among those who did

not receive bar graph feedback, $t(88) = .743$, $p = .459$. Additionally, there was a trending interaction between time and majority, $F(1, 175) = 3.318$, $p = .070$, $\eta_p^2 = .019$, such that individuals who were tested immediately performed marginally 7.7 ± 8.4% better when the majority answered initial clicker questions correctly, rather than incorrectly, $p = .072$, but there was no effect on initial majority response when the test was delayed ($p = .459$). The 3-way interaction between condition, time, and majority was not significant, $F(1, 175) = .141$, $p = .708$, $\eta_p^2 = .001$. Finally, Hypothesis 4 was not supported, such that the interaction between condition and time of test was not significant, $F(2, 270) = .569$, $p = .567$, $\eta_p^2 = .004$, indicating that effects of clicker condition did not differ between immediate and delayed testing.

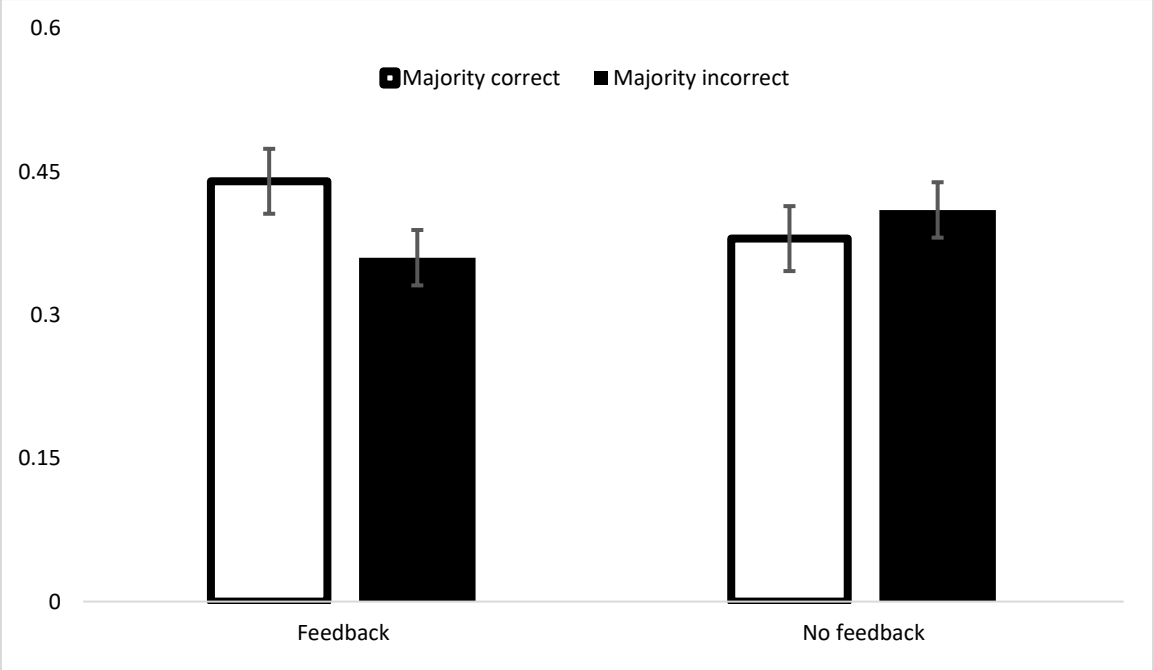

Figure 4. Average proportion of correct test answers of each condition by majority correct and incorrect clicker responses in Experiment 2. Error bars reflect standard error of the mean.

Further, I anticipated enhanced test scores in Experiment 2, compared to Experiment 1, due to monetary incentive. However, an independent samples *t*-test showed no significant difference in test performance between these groups, $t(265) = .152$, $p = .879$. Participants were not randomly assigned to either group, however, so this finding cannot be concluded as causal.

Additionally, I assessed the possible positive, negative, or non-significant relationship between test performance and self-reported motivation. To test this, I first conducted correlations between test performance, self-reported effort, and self-reported importance. As can be seen in Table 1, all 3 correlations were significantly positive, such that higher test scores were associated with higher self-reported effort and importance. Importantly, however, a one-way analysis of covariance showed that effects of condition, $F(2, 269) = 5.360$, $p = .005$, and time, $F(1, 269) = 19.602$, $p = .000$, were still significant even after controlling for effort. Similarly, effects of condition, $F(2, 269) = 5.608$, $p = .004$, and time, $F(1, 269) = 14.838$, $p = .000$, still reached significance after controlling for importance. Further, there were no significant interactions between condition and effort, $F(27, 229) = .847$, $p = .686$, and condition and importance, $F(31, 223) = 1.150$, $p = .277$, on test performance. Therefore, although effort and importance were associated with increased test performance, effects of condition or time of testing on test performance were independent of self-reported effort or importance, and the effect of condition on test performance did not depend on participant motivation.

	Test Performance	Importance	Effort
Test Performance	1	.158*	.194*
Importance		1	.375*
Effort			1

Table 1. Pearson *R* correlations between test performance, self-reported effort, and self-reported importance.
*Correlation is significant at the .01 level (2-tailed).

I also assessed effects of condition and time on self-reported effort and importance to examine whether using clickers enhances motivation. A one-way ANOVA revealed no significant main effects of condition, $F(2, 270) = .067$, $p = .935$, or time, $F(1, 270) = .007$, $p = .93$, on self-reported effort. Additionally, there was no significant main effect of condition on self-reported importance, $F(2, 270) = 1.661$, $p = .192$; however, there was a significant main effect of time on self-reported importance, $F(1, 270) = 9.635$, $p = .002$, such that individuals who were tested immediately, relative to delayed testing, reported higher levels of perceived importance. There was no significant interactions between condition and time on either self-reported effort, $F(2, 270) = 2.029$, $p = .133$, or self-reported importance, $F(2, 270) = .643$, $p = .527$.

Combined Analysis

Finally, because both Hypotheses 1 and 3 received at least marginal support in both Experiments 1 and 2, these hypotheses were tested with the combined data to increase power to detect the effects. First, independent *t*-tests examined whether overall performance across experiments was better for those that used clickers than for those in

the control group. Overall, performance was indeed 5.7 ± 5.1% better for those who received bar graph feedback than those in the control group, $t(274) = 2.188$, $p = .03$, and 8.6 ± 5.2% better for those who received no bar graph feedback than those in the control group, $t(272) = 3.247$, $p = .001$. Again, there was no difference between the two clicker groups, $t(266) = 1.057$, $p = .291$. Next, a 2 (bar graph feedback, no bar graph feedback) x 2 (majority response) mixed factorial analysis of variance assessed the effects of majority response of initial clicker questions on test performance of individuals in both Experiment 1 and 2 who either received bar graph feedback or did not. There was indeed a significant interaction between majority response and condition, $F(1, 266) = 6.081$, $p = .014$, $\eta_p^2 = .022$. Overall, as predicted by Hypothesis 3, individuals who received bar graph feedback performed significantly 7.3 ± 6.4% better on the test when the majority answered initial clicker questions correctly rather than incorrectly, $t(134) = -2.339$, $p = .021$; however, there was no effect of initial majority response among those who did not receive bar graph feedback, $t(132) = 1.197$, $p = .234$. Independent group *t*-tests also showed that those in the no feedback group marginally outperformed those in the feedback group when initial majority responses were incorrect, $t(266) = -1.933$, $p = .054$, but there was no difference when initial majority responses were correct, $t(266) = 1.205$, $p = .229$.

Discussion

In Experiment 2, Hypothesis 1 was significantly supported, such that individuals who answered clicker questions performed significantly better on the test than individuals who did not answer clicker questions. Consistent with Experiment 1 results, Hypothesis 2 was not supported, such that individuals without bar graph feedback following clicker questions did not perform significantly better on the test than individuals who received bar graph feedback. Thus, Experiment 2 results suggest that test performance benefited from receiving clicker questions during lecture but was not significantly affected by whether participants received bar graph feedback. Numerically, however, individuals who did not receive bar graph feedback performed better on the test than individuals who received bar graph feedback, prompting speculation that individuals who received bar graph feedback were possibly confused by the multiple graphs that they viewed. Consistent with Hypothesis 3, the potential benefit or harm in receiving bar graph feedback depends upon whether the bar graph feedback is consistent or inconsistent with the correct response. In Experiment 2, this hypothesis was again marginally supported, such that test performance was marginally enhanced when individuals viewed the correct majority response.

Importantly, when the data for Experiment 1and 2 were combined, those who received bar graph feedback benefitted from viewing correct, as opposed to incorrect, majority responses to initial clicker questions, whereas test performance of individuals who did not receive bar graph feedback was not significantly affected by majority response. These results suggest that participants' test performance was indeed influenced

by viewing bar graph feedback, but seeing how others answered initial clicker questions was only advantageous if the majority of individuals answered correctly.

Additionally, when individuals were tested immediately, test performance marginally increased when the majority of participants answered clicker questions correctly. The interaction between condition, time, and majority, however, did not significantly impact test performance. Thus, my prediction that individuals who received bar graph feedback would perform significantly worse after a delay when the majority of individuals answered initial clicker questions incorrectly, relative to individuals who did not receive bar graph feedback, was not significantly supported.

Further, Hypothesis 4 was not supported, such that effects of condition were not greater after a delay. This finding suggests that plausible source monitoring errors produced by individuals who received bar graph feedback were not exacerbated during the delay. In fact, although the interaction with delay was not significant, when tested separately, the effects of condition was significant for those who tested immediately (p = .032), but not for those who tested after a delay (p = .249). Thus, all participants performed relatively uniform on the delayed test, regardless of the condition in which they belonged.

Furthermore, offering a monetary incentive did not significantly boost test performance. Average test performance between individuals in Experiment 1 and Experiment 2 (immediate testing) did not differ significantly, tentatively indicating that test performance of individuals in Experiment 2 was not influenced by the prospect of receiving money. Additionally, higher test scores were significantly positively associated

with higher reports of effort and importance, suggesting a relationship between perceived motivation and actual achievement. Alternatively, people could have used their test performance to infer motivation. Thus, it is likely that high self-reports of motivation were not determined by monetary incentive, but perhaps personal motives, like an innate fear of failing or being judged as unintelligent. Although motivation played a significant role in test performance, it did not alter the significant effects of condition and time. That is, even when controlling for effort and importance, condition and testing time continued to significantly influence test performance.

Finally, although individuals who tested immediately, relative to individuals who tested after a delay, reported significantly greater levels of perceived importance, and despite the finding that test performance significantly positively correlated with self-reported motivation, self-reported levels of effort and importance cannot be explained by condition. There were no significant interactions between condition and time on self-reported effort or importance. Therefore, the use of clickers and/or the presentation of initial bar graph feedback did not influence motivation, suggesting that using clickers did not boost participants' determination to perform well.

GENERAL DISCUSSION

This study examined the influence of receiving clicker bar graph feedback on subsequent test performance. Based on extent literature suggesting that classroom participation enhances scholastic performance, I first hypothesized that individuals who received clicker questions would perform better on the subsequent test than individuals who did not receive clicker questions and, thus, did not actively participate during lecture. This hypothesis was marginally supported in Experiment 1, such that individuals who received clicker questions and did not receive bar graph feedback performed numerically better than individuals who did not actively participate, and was significantly supported in Experiment 2, such that individuals in both clicker conditions had greater test performance, on average, than individuals in the control condition. When Experiments 1 and 2 were combined, Hypothesis 1 was supported such that both clicker groups outperformed the control group.

Results supporting my first hypothesis are congruent with Trees and Jackson's (2007) findings, suggesting that receiving clicker questions during lecture, as opposed to no application of student engagement, enhances subsequent test performance. Thus, in correspondence with extent clicker literature (e.g., Kennedy & Cutts, 2005; Morling, McAuliffe, Cohen, & DiLorenzo, 2008; Shaffer & Collura, 2009; Shapiro & Gordon, 2012; Trees & Jackson, 2007) this study reinforces the claim that clickers can be useful tools in encouraging participation and boosting test performance, at least under appropriate learning conditions. Additionally, results of Hypothesis 1, that test performance is enhanced by active participation, indicate that individuals who received

clicker questions during the encoding of lecture content were better able to retrieve the information from memory when taking the subsequent test, presumably because of the match in context and processing between encoding and test (e.g., Morris, Bransford, & Franks, 1977; Tomson & Tulving, 1970). Finally, results are consistent with the testing effect (Carrier & Pashler, 1992) because individuals who received clicker questions (i.e., initial test) performed better on the subsequent test than individuals who simply listened to the lecture and did not receive initial testing.

Secondly, I hypothesized that test performance would be worse for individuals who received clicker questions with bar graph feedback, relative to those who received clicker questions without bar graph feedback, due to having an extra source (i.e., graph) to monitor. Hence, while taking the test, individuals who received bar graph feedback might become confused as to whether their recollection of an answer to a test question came from the graph displaying the correct answer (relevant source) or from the graph displaying the bar graph feedback of how other students answered (irrelevant source). This hypothesis was not supported, such that test performance of individuals who received clicker questions during lecture was not influenced by the presence or absence of bar graph feedback.

Like Lantz and Stawiski (2014), I found no significant difference in test performance between clicker conditions. A non-significant difference in test performance between clicker conditions could have occurred because of the differently colored bar graphs between the graph of the actual results and the graph of the frequency of individual answers. The lack of support for Hypothesis 2 can be explained by

Zaragoza and Lane (1994)'s argument that individuals tend to visualize details of irrelevant or misleading information and, even after short delays, constructing visual details of misinformation increases the likelihood of committing source monitoring errors. Thus, because individuals who received bar graph feedback not only processed the visual details of the graph displaying the correct answer, but also the graph of the frequencies of individual answers, they might have noticed the different colors that distinguished the two graphs. Clearer discrimination between graphs would allow for more accurate source attributions, resulting in similar test performance between individuals who received and those who did not receive bar graph feedback.

Further, in agreement with Anthis (2011), Lasry (2008), and Miller et al. (2003), as long as individuals receive relevant, thought provoking questions during lecture and corrective feedback, maybe other mechanisms are useless in cultivating learning. For example, as long as individuals received clicker questions and corrective feedback, perhaps bar graph feedback was simply extra information that neither significantly helped nor hindered test performance. As observed in analyses of majority response, however, bar graph feedback does appear to play a significant role in determining test performance.

Individuals who received bar graph feedback encountered the possibility of misattributing their recollection to the irrelevant source (bar graph feedback) rather than the relevant source (correct answer). If source confusion occurred, test performance of these individuals could benefit if the majority of students also answered correctly; however if the majority of students answered incorrectly, these individuals would more likely suffer from the ramifications of monitoring multiple sources. Therefore, my third

hypothesis was that if the majority of individuals who received bar graph feedback answered initial clicker questions incorrectly, then the average proportion of correct test answers should decrease due to misattributing the bar graph feedback as the correct answer on corresponding test questions. This hypothesis was supported. Only individuals in the feedback group performed better on test questions in which the initial majority answered correctly, and, thus, the two potential sources of memory pointed to the same answer. The fact that those who did not receive bar graph feedback performed similarly, regardless of initial majority response, suggests that they internally corrected their initial incorrect answers upon receiving the correct answer (Metcalfe & Miele, 2014). A follow-up analysis of Experiments 1 and 2 combined shows that those who did not receive bar graph feedback marginally outperformed those receiving bar graph feedback when the initial majority was incorrect. This marginal effect is ironic, considering that, on average, individuals who did not receive bar graph feedback corrected themselves, whereas those who did receive bar graph feedback did not correct themselves. Apparently, incorrect majority response interfered with participants' ability to correct their initial clicker response.

There was a marginal interaction between time of test and majority response in Experiment 2, such that performance benefited from correct majority responses only when participants were tested immediately, rather than after a delay. Overall, participants performed marginally better on immediate test questions in which the majority was correct on initial clicker questions, but numerically better on delayed test

questions in which the majority was incorrect on initial clicker questions, suggesting a hyper-correction of initially incorrect responses (Metcalfe & Miele, 2014).

Finally, my fourth hypothesis was that effects of condition would be greater after a delay, such that superior test performance of individuals who received clicker questions compared to individuals in the control condition should be even greater after a delay, due to testing effects. Further, superior test performance of individuals who did not receive bar graph feedback, relative to those who did, should be even greater after a delay because of increased source monitoring errors over time. This hypothesis was not supported. The difference in average performance between clicker conditions did not differ significantly when individuals were tested immediately or after a delay.

Results of Hypothesis 4 are contrary to Roediger and Karpicke's (2006a) finding that testing effects strengthen after a delay of just two days. Accordingly, individuals in the clicker conditions should have performed significantly better after a delay than individuals in the control condition because, even though performance was expected to weaken after a delay for all participants, the advantage of receiving initial clicker questions should have broadened the gap in average performance between clicker and control conditions. Contrarily, results of the current study showed no significant difference between clicker and control conditions. It is possible that the delay was too short to detect any strengthening of testing effects. This prospect is unlikely, however, because if anything, a significant difference in test performance between these conditions should have been maintained at such a short delay. Instead, after just 48-hours, these conditions did not differ significantly, and presumably, this difference would only

continue to be non-significant after a longer delay. Additionally, it could be that the benefit of testing requires a longer lag between processing the description of the study and answering its corresponding clicker question. According to Godbole, Delaney, and Verkoeijen (2014), information is better remembered when repeated processing is spaced, rather than massed, and this spacing effect persists after a delay. Further, because there were no effects of testing after a delay, it is possible that the immediate effects of testing were not due to retrieval practices, but instead due to pre-exposure to questions that would subsequently appear on the test. Thus, without practicing retrieval, memory for initial clicker questions and their correct answers would likely fade over time. To control for this, a different control group could present participants with clicker questions that simultaneously include their highlighted correct answer along with the incorrect lures, which would likely prevent students from having to come up with the answer.

Additionally, my results disconfirm Underwood and Pezdek's (1998) finding that, over a delay, source monitoring errors increase, especially when participants receive a narrative from a low-credible source, because the source and the message of the narrative become less associated over time. Because individuals who received more graphs to monitor did not perform significantly worse than individuals who did not receive bar graph feedback in the current study, it is unlikely that individuals who received bar graph feedback were subject to the sleeper effect, as reported by Hovland et al. (1949). The delay in the current study was only 48-hours, a much shorter time interval than those used in Underwood and Pezdek's and Hovland et al.'s studies. Consistent with Frost et al.'s (2002) explanation of the effects of delay on source monitoring, it is possible that not

enough time elapsed for it to become difficult to discriminate the bar graph with the correct answer from the irrelevant bar graph feedback. Thus, participants would still be able to attribute the correct answer to the correct source and, therefore, would perform similarly to individuals who did not have to monitor two graphs.

Most importantly, a combined analysis of majority response and condition on test performance between individuals who received clicker questions in Experiments 1 and 2 found a significant interaction. Individuals only benefited from receiving bar graph feedback if the majority answered initial clicker questions correctly. There was no significant effect of majority response on test performance for individuals who did not receive bar graph feedback. These expected results indicate that individuals who received bar graph feedback were, indeed, influenced by viewing how other participants answered clicker questions. Test performance, however, only benefited from viewing correct majority responses, suggesting that individuals who received bar graph feedback tended to misattribute the majority answer as the correct answer. By committing this source misattribution error, test performance increased only when the majority response was correct. Conversely, test performance was not enhanced if the majority of individuals answered incorrectly. In fact, the bar graph feedback group performed marginally worse than the no feedback group when the majority was initially incorrect. The non-significant effect of majority response on test performance for those who did not receive bar graph feedback makes sense because these individuals did not view others' responses to clicker questions and, therefore, should not be susceptible to the effects of having multiple graphs to monitor. Additionally, this non-significant effect on those who

did not receive bar graph feedback provides evidence that the bar graph feedback disrupted performance when the majority response was incorrect.

Experiment 2 also explored supplementary predictions positing that offering a monetary incentive would help boost test scores, and self-reported motivation may, or may not be, positively or negatively correlated with test performance. Contrary to Oswald and Rhoten's (2014) finding that offering incentives, especially when contingent on clicker performance, enhanced retention of lecture material via classroom engagement, current results revealed that the prospect of receiving money did not significantly enhance test scores. Perhaps the possibility of receiving 50 dollars was not a large enough sum to convince students to perform at their best abilities. Further, it is possible that the incentive was not properly contingent upon performance, such that an effect might have been observed if participants were increasingly rewarded for the number of clicker questions and test questions that were correctly answered. Additionally, although self-reported levels of effort and importance significantly positively correlated with test scores, when controlled for, these self-reported levels of motivation did not influence the effects of condition and time on test performance. Further, condition and time of test did not significantly affect self-reported effort. Therefore, using clickers does not improve self-reported motivation. Self-reported importance was not significantly affected by condition; however, individuals who were tested immediately reported significantly higher levels of perceived importance than individuals who were tested after a delay. Participants may have been more likely to report high levels of importance only when tested immediately, rather than after a delay,

because they were not given time to overcome the novelty of participating in an experiment that is necessary in order to obtain participation credits for class. This rationalization, however, does not explain why individuals did not also report high levels of effort when tested immediately. The positive correlations between self-reported levels of motivation and test performance disconfirm findings by Morgan (2008), Sun et al. (2014), and Sutherlin et al. (2013) in that, in the current study, as test scores increased self-reported levels of effort and motivation also increased. Thus, in general, participants' perceived competency paralleled their actual achievement.

Ultimately, these results encourage the use of clickers in classroom settings due to the response system's ability to enhance student participation and, consequently, test performance. In light of marginal support when Experiments 1 and 2 were considered separately and significant support when both experiments were combined, results indicated a detrimental effect on test performance when bar graph feedback was displayed and initial majority responses were incorrect. Individuals only benefited from receiving bar graph feedback if the majority of students answered clicker questions correctly. As seen in the current study, when questions are challenging and thought provoking, it is not common for the majority of individuals to consistently answer clicker questions correctly. Thus, receiving bar graph feedback might not be useful in enhancing test performance in clicker-based classes when questions are formatted in an academically-preferred manner that challenges students to think constructively. Consequently, instructors should be cautious of the possible issues that could arise when students are given multiple sources to observe.

Although the results of the current study encourage instructors to reevaluate displaying bar graph feedback during clicker-based lectures, several techniques can be implemented that could possibly safeguard against negative effects of bar graph feedback. For example, instead of showing students how they answered relative to the rest of the class, perhaps only the instructor should view the bar graph feedback. This way, the instructor could determine how well the class is comprehending the information in real-time and adjust the lecture accordingly and, in turn, the students will not be exposed to possibly detrimental sources of information.

Another way to attempt to curb the detrimental effects of source monitoring errors in clicker-based lectures is by adding an additional review test after initial clicker questions and corrective feedback are received. For example, Metcalfe and Miele (2014) tested the effects of hypercorrection (Butterfield & Metcalfe, 2001), a phenomenon in which erroneous answers that were resolved with confidence are more easily corrected than those that were made with less confidence, after the course of a week interval. Metcalfe and Miele found that high-confidence errors reemerged after a delay when items were not immediately tested. Further, participants were less likely to mistake their original error as being correct and more likely to know that the correct answer was correct if such questions were intervened by a test immediately after receiving corrective feedback. In this vein, adding an extra review "test" of clicker questions immediately after the initial clicker questions are asked and corrective feedback is provided could, perhaps, hamper the possibility of incorrect information from being recalled at the time of final testing.

Despite the implications of this study, several limitations are observed. The first issue involves the color schemes of the correct answer and bar graph feedback histograms. The actual results, or the correct answer, of the studies described in the lecture were displayed with yellow and blue bars; whereas the bar graph feedback was displayed with only blue bars. Individuals who received bar graph feedback might have been able to better distinguish between these sources (i.e., correct answer bar graph and feedback bar graph), because of their distinct colors, and attribute their test answer to the correct source (correct answer bar graph). Source monitoring theory (e.g., Johnson et al., 1993; Johnson & Raye, 1981) poses that perceptual features of target memories act as cues to select the source of that memory, and if perceptual details are similar between two sources of information, distinguishing the correct source of a memory becomes more difficult (Foley & Johnson, 1985; Johnson, Foley, Suengas, & Raye, 1988). For example, Ferguson, Hashtroudi, and Johnson (1992) found that the sources of spoken words are more accurately remembered when the two speakers are male and female (dissimilar voices and features) than when the speakers are two females (similar voices and features). Similarly, if both graphs had been the same blue bars, then individuals would have more difficulty discriminating the graphs, allowing for a greater effect to be observed between both clicker conditions.

The second drawback of this study concerns the inconsistent number of individuals who participated within each "class." The ideal class size was approximately 20 students; however, because students independently signed up to participate at their own convenience, some classes did not have an optimal turn-out. Smaller classes in this

study, consisting of less than five students, for example, do not represent typical clicker-based courses on college campuses. Additionally, the social atmosphere might have been different between small class sizes and large class sizes. For instance, an individual might have felt more confident after answering correctly, especially if the rest of the class answered incorrectly and even more so if the class size was large, indicating that he or she understood the content better than many people in the class. Similarly, an individual might have been embarrassed seeing that he or she answered incorrectly against the majority. This embarrassment could be exacerbated by a large class size, eliciting shame of being the only person in the class to answer incorrectly. Contrastingly, an individual might have a stronger emotional response to being correct or incorrect in a smaller class size due to less perceived anonymity.

In this way, test performance on specific questions could be affected by cases in which an individual has an emotional response associated to a clicker response. For example, Wang (2015) found that positive arousal experienced after the encoding of words enhanced delayed recognition of those words. In a related study, Wang and Sun (2017) reported analogous findings after participants experienced negative arousal. Interestingly, both studies found that post-encoding arousal had no effect on the consolidation of source memory (Wang, 2015; Wang & Sun, 2017). Thus, experiencing feelings of pride or embarrassment after realizing an answer was correct or incorrect, respectively, relative to answers of peers could enhance recognition for that specific test question; however, such emotional arousal might not help individuals in knowing the correct source of their memory (i.e., correct answer or bar graph feedback).

Research must be carried out in order to extend the findings of this study and resolve the discord of feedback presentation in clicker-based lectures. First, this study should be replicated using graphs with identical bars. The yellow and blue histogram bars representing the correct answer were not edited to match the blue histogram bars displaying how individuals in the class answered. This easy fix would increase the visual similarities between graphs, magnifying the effect, if any, of committing source monitoring errors.

Additionally, individual answers on both clicker and test questions should be assessed contingent with the majority clicker answer. Participants were asked to bring personal clickers to the study. Unfortunately, because the mock "classes" were not actual courses in which the participants were registered, individual answers to clicker questions were not recorded by the REEF polling system. Only the frequency of each selection (i.e., A, B, C, or D) was provided. It would be interesting to observe how individuals answered the initial clicker question and their answer to that same question on the subsequent test depending on how the majority of individuals in the class answered the initial clicker question. For example, a participant might have answered a clicker question incorrectly, saw that the majority of students answered differently, yet still incorrectly, and answered the coinciding test question with the majority answer, despite having received corrective feedback, because it was misattributed as coming from the correct, relevant source. This correction could be implemented by administering a clicker-based experiment in an actual university course or by using a set of pre-registered clickers provided for participants solely for the purpose of the study. By analyzing

individual data, the pattern of keeping or switching answers based on viewing the majority response might provide a stronger basis of the effect of receiving bar graph feedback.

Further, this study only offered individuals answering clicker questions one testing opportunity (initial clicker questions) before the subsequent test. This procedure is typical in actual classes in that clicker questions are usually constructed for the information learned during that class period and not revisited during the following class. However, based on what is known about the testing effect and its enhancement after a delay (Carrier & Pashler, 1992; Jacoby, 1978; Modigliani, 1976), clicker research should explore the outcome of test performance after specific clicker questions are repeated each week along with new clicker questions in order to truly observe the consequences of repeated retrieval on performance. (Carrier & Pashler, 1992; Jacoby, 1978; Modigliani, 1976). Additionally, in order to determine if clicker-based lectures enhance testing effects after a delay, this study should be modified by administering clicker questions at the end of class after all studies have been described, rather than immediately after the relevant information is taught, so that there is a longer lag between processing and initial retrieval.

Finally, majority response analyses revealed that test performance was, indeed, influenced by receiving bar graph feedback; however, it is conceivable that the effects of observing bar graph feedback are based on social conformity, and not simply source misattribution errors. Social conformity occurs when an individual alters his or her behavior, thoughts, values, etc. in order to cohere with the majority (Asch, 1951; Kelman,

1958). Clicker studies that have addressed the concept of social conformity propose that the use of clickers actually reduce the likelihood of students answering with the majority due to the technology's anonymity (Stowell & Nelson, 2007; Stowell, Oldham, & Bennett, 2010); however, such studies did not display bar graph feedback after clicker questions were asked. It is possible that viewing how the majority of students in a class answer a clicker question could provoke some individuals to answer the corresponding test question in conjunction with the majority, due to social pressure, even if the correct answer was received. This prospect is unlikely because the final test was completed individually and confidentially, so any pressure to conform to others' responses should be mitigated at that point; however, it is still possible that individuals might feel compelled to privately answer test questions with initial majority responses, despite having received corrective feedback. Additionally, extent literature on peer-to-peer instruction suggests that enhancement in academic performance is specifically due to discussion (e.g., MacArthur & Jones, 2008; McDonough & Foote, 2015); however, it could be possible that discussion might foster social contagion and prompt source attribution errors during testing. Thus, social conformity and social contagion, as well as other possible determinants of reduced test performance due to receiving bar graph feedback, must be examined in order to establish the appropriateness, or lack thereof, in presenting bar graph feedback during clicker-based lectures.

Clicker-based lectures are advantageous in enhancing academic performance, likely due to their facilitation of active participation, but this approach can also prove detrimental if implemented unfavorably. The results of this study indicate that clickers

are useful tools in enhancing test performance; however, displaying bar graph feedback likely hinders test performance, especially when clicker questions are constructed in a manner that promotes conceptual thinking. When clicker questions are challenging, it is likely that the majority of students will answer incorrectly. Thus, viewing incorrect answers via bar graph feedback may influence increased source monitoring errors and decreased answer correction after receiving corrective feedback, likely resulting in poor test performance. Additionally, the small benefit of clickers was not significant after a delay, suggesting that the immediate benefit of using clickers might simply be due to previewing the questions that later appear on the test. If clicker-based lectures do not, in fact, enhance effects of testing over time, perhaps clickers are not worth implementing in courses that typically last several months. Future studies should replicate the current study by using identically colored bar graphs, analyzing individual data in order to assess the pattern of selecting answers based on specific majority responses, and including an intervening clicker test in order to determine if source monitoring errors can be abated. Additionally, research should investigate test performance when clicker questions are administered at the end of class and repeated throughout an entire course in order to observe testing effects after a delay. Finally, other possible causes of the disadvantages of viewing bar graph feedback, such as social conformity or contagion, should be assessed.

REFERENCES CITED

Anderson, C. W. (1987). Strategic teaching in science. In B. F. Jones (Ed.), *Strategic teaching and learning: Cognitive instruction in the content areas*, (pp. 73-91). Alexandria, VA: Association for Supervision and Curriculum Development.

Anthis, K. (2011). Is it the clicker, or is it the question: Untangling the effects of student response system use. *Technology and Teaching*, *38*(3), 189-193.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (Ed.), *Groups, leadership and men*. Pittsburg, PA: Carnegie Press.

Atlantis, E., & Cheema, B. S. (2015). Effect of audience response system technology on learning outcomes in health students and professional: An updated systematic review. *International Journal of Evidence-Based Healthcare*, *13*, 3-8.

Baker, K. Q., Spiezio, K. E., & Boland, K. (2004). Student engagement: Transference of attitudes and skills to the workplace, profession, and community. *The Industrial-organizational Psychologist*, *42*(2), 101-107.

Boyle, J. T., & Nicol, D. J. (2003). Using classroom communication systems to support interaction and discussion in large class settings. *Association of Learning Technology Journal*, *11*, 43-57.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*, 633-642.

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *33*, 431-437.

Desrochers, M. N., & Shelnutt, J. M. (2012). Effect of answer format and review method on college students' learning. *Computers & Education*, *59*, 946-951.

Dodson, C. S., & Schacter, D. L. (2002). Aging and strategic retrieval processes: Reducing false memories with a distinctiveness heuristic. *Psychology & Aging*, *17*, 405-415.

Fallon, M., & Forrest, S. L. (2011). High-tech versus low-tech instructional strategies: A comparison of clickers and handheld response cards. *Technology & Teaching*, *38*(3), 194-198.

Ferguson, S. A., Hashtroudi, S., & Johnson, M. K. (1992). Age differences in using source relevant cues. *Psychology & Aging, 7,* 443- 452.

Foley, M. A., & Johnson, M. K. (1985). Confusions between memories for performed and imagined actions. *Child Development, 56,* 1145-1155.

Frost, P., Ingraham, M., & Wilson, B. (2002). Why misinformation is more likely to Recognized over time: A source monitoring account. *Memory, 10*(3), 179-185l.

Godbole, N. R., Delaney, P. F., & Verkoeijen, P. P. J. L. (2014). The spacing effect in immediate and delayed free recall. *Memory, 22*(5), 462-469.

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. (1949). *Experiments in mass communication.* Princeton, NJ: Princeton University Press.

Hwang, J. H., & Wolfe, K. *(*2010*).* Implications of using the electronic response system in a large class. *Journal of Teaching in Travel and Tourism, 10,* 265–279.

I>clicker (Version 7) [Computer software]. (2016). Macmillan.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning & Verbal Behavior, 17,* 649-667.

Johnson, M. K., Foley, M. A., Suengas, A. G., & Raye, C. L. (1988). Phenomenal characteristics of memories for perceived and imagined autobiographical events. *Journal of Experimental Psychology: General, 117,* 371-376.

Johnson, M. K., Hashtroudi, S., & Lindsay, S. D. (1993). Source monitoring. *Psychological Bulletin, 114(*1*),* 3-28.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review, 88,* 67 85.

Kay, R. H., & LeSage, A. (2009). Examining the benefits and challenges of using audience response systems: A review of the literature. *Computers & Education, 53*(3), 819-827.

Kelman, H. C. (1958). Compliance, identification, and internalization: Three processes of attitude change. *Journal of Conflict Resolution, 2,* 51-60.

Kennedy, G., & Cutts, Q. (2005). The association between students' use of an electronic voting system and their learning outcomes. *Journal of Computer Assisted Learning, 21,* 260-268.

Lasry, N. (2008). Clickers or flashcards: Is there really a difference? *The Physics Teacher, 46,* 242-244.

Lindsay, S. D. (1990). Misleading suggestions can impair eyewitnesses' ability to remember event details. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(6), 1077-1083.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4*(1), 19-31.

MacArthur, J. R., & Jones, L. L. (2008). A review of literature reports of clickers applicable to college chemistry classrooms. *Chemical Education, Research and Practice, 9,* 187-195.

Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American Educational Research Journal, 37*(1), 153-184.

Mayer, R. E. (1975). Forward transfer of different reading strategies due to test-like events in mathematics text. *Journal of Educational Psychology, 67,* 165-169.

Mazursky, D., & Shul, Y. (1987). The effects of advertisement encoding on the failure to discount information: Implications for the sleeper effect. *Journal of Consumer Research, 15,* 24-35.

McCabe, D. P., & Geraci, L. (2009). The role of extralist associations in false remembering: A source misattribution account. *Memory & Cognition, 37*(2), 130-142.

McDonough, K., & Foote, J. A. (2015). The impact of individual and shared clicker use on students' collaborative learning. *Computer & Education, 86,* 236-249.

Miller, R. G., Ashar, B. H., & Getz, K. J. (2003). Evaluation of an audience response system for the continuing education of health professionals. *Journal of Continuing Education in the Health Professions, 23,* 109-115.

Morgan, R. K. (2008). Exploring the pedagogical effectiveness of clickers. *InSight: A Journal of Scholarly Teaching, 3,* 31-36.

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

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer-appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519-533.

Oswald, K. M., & Rhoten, S. E. (2014). Improving classroom clicker practices: Effects of incentives and feedback on retention. *North American Journal of Psychology*, *16*(1), 79.

PowerPoint (Version 15.0) [Computer software]. (2013). Santa Rosa, CA: Microsoft.

REEF Polling by i>clicker [Computer software]. (2016). Macmillan.

Roediger, H. L., III., & Karpicke, J. D. (2006a). Test enhanced learning: Taking memory tests improves long-term retention. Psychological Science, 17, 249-255.

Roediger, H. L., III., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181-210.

Sagerman, N., & Mayer, R. E. (1987). Forward transfer of different reading strategies evoked by adjunct questions in science text. *Journal of Educational Psychology*, *79*, 189-191.

Schacter, D. L., Israel, L., & Racine, C. A. (1999). Suppressing false recognition in younger and older adults: The distinctiveness heuristic. *Journal of Memory & Language*, *40*, 1-24.

Shaffer, D. M., & Collura, M. J. (2009). Evaluating the effectiveness of personal response system in the classroom. *Teaching of Psychology*, *36*, 273-277.

Shapiro, A. M., & Gordon, L. T. (2012). A controlled study of clicker-assisted memory enhancement in college classrooms. *Applied Cognitive Psychology*, *26*, 635-643.

Shernoff, D. S., & Hoogstra, L. (2001). Continuing motivation beyond the high school classroom. *New Directions in Child and Adolescent Development*, *93*, 73-87.

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science, 323*, 122-124.

Stowell, J. R., & Nelson, J. M. (2007). Benefits of electronic audience response systems on student participant, learning, and emotion. *Teaching of Psychology*, *34*, 253-258.

Stowell, J. R., Oldham, T., & Bennett, D. (2010). Using student response systems ("clickers") to combat conformity and shyness. *Teaching of Psychology, 37*, 135-140.

Sun, J. C., -Y., Martinez, Z. B., & Seli, H. (2014). Just-in-time or plenty-of-time teaching? Different electronic feedback devices and their effect on student engagement. *Educational Technology & Society, 17*(2), 234-244.

Sundre, D. L. (2007). The student opinion scale (SOS): A measure of examinee motivation [Test manual]. Harrisburg, VA: The Center for Assessment & Research Studies.

Sutherlin, A. L., Sutherlin, G. R., & Akpanudo, U. M. (2013). The effect of clickers in university science courses. *Journal of Science Education and Technology, 22*, 651-666.

Tomson, D. M., & Tulving, E. (1970). Associative encoding and retrieval: Weak and strong cues. *Journal of Experimental Psychology, 86*(2), 255-262.

Trees, A. R., & Jackson, M. H. (2007). The learning environment in clicker classrooms: Student processes of learning and involvement in large university-level courses using student response systems. *Learning, Media and Technology, 32*(1), 21-40.

Underwood, J., & Pezdek, K. (1998). Memory suggestibility as an example of the sleeper effect. *Psychonomic Bulletin and Review, 5*, 449-453.

Wang, B. (2015). Positive arousal enhances the consolidation of item memory. *Swiss Journal of Psychology, 74*(2), 91-104.

Wang, B., & Sun, B. (2017). Post-encoding emotional arousal enhances consolidation of item memory, but not reality-monitoring source memory. *The Quarterly Journal of Experimental Psychology, 70*(3), 461-472.

Zaragoza, M. S., & Lane, S. M. (1994). Source misattribution and the suggestibility of eyewitness memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(4), 934-945.