Fall 12-2017

Quizzing and Restudy Dynamics in a TST Paradigm: The (Null) Effect of Feedback and the (Significant) Effects of Metacognition

Francis Anderson
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Psychological & Brain Sciences

Quizzing and Restudy Dynamics in a TST Paradigm: The (Null) Effect of Feedback and the (Significant) Effects of Metacognition

by

Francis T. Anderson

A thesis presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Master of Arts

December 2017
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank my dissertation committee, Mark McDaniel (my mentor), Julie Bugg, and Kathleen McDermott, for taking time out of their busy schedules to read my master's and grill me with tough questions during my defense. I would also like to thank the National Science Foundation for funding me through the Graduate Research Fellowship (DGE-1745038).

I certainly need to thank the rest of the Memory and Complex Learning Lab (Reshma Gouravajhala and Toshi Miyatsu) for giving me much needed critique and advising the restraint of my natural tendencies towards embellishment. Along those same lines, I would also like to thank the people that in general keep me in line; most notably Öykü Üner, who has spent a disproportionate amount of time listening to me rant and ramble about my data and design.

Finally, this project would not have been possible without the aid of several undergraduate research assistants, who have spent many hours running participants and coding data. These undergraduate are Rachel Lowe, Yejin Lee, Nick Fierro, and Alana Dinh. I wish them the best, because they gems.

Francis T. Anderson

Washington University in St. Louis

December 2017

Abstract

Quizzing and Restudy Dynamics in a TST Paradigm: The (Null) Effect of Feedback and the (Significant) Effects of Metacognition

by

Francis T. Anderson

Master of Arts in Psychological & Brain Sciences

Washington University in St. Louis, 2017

Professor Mark McDaniel, Chair

In authentic educational settings, using formative quizzes or tests can improve students' memory by direct strengthening of the memory trace. There are other indirect effects of testing, however, such as improved understanding of what one does and does not know. That is, quizzes can benefit students' metacognitive awareness, which may in turn affect their restudy behaviors. We tested whether different types of feedback (correct/incorrect, correct answer, or minimal) differentially affected students' metacognition, changed their restudy behaviors, and influenced final test performance. We found no effect of feedback type, but were able to better understand quizzing and restudy dynamics in an authentic educational scenario. For example, we show that even with minimal feedback, participants had insight into which concepts they answered incorrectly, because they later chose to restudy those concepts. Additionally, they were especially likely to restudy high-confidence errors, which were the most discrepant from expected performance. Finally, these behaviors appear to be adaptive, in that the items they chose to restudy were more likely to be answered correctly on the final test.

1.1 Introduction

When applying cognitive psychology to education and instruction, no phenomenon has received more attention than the testing effect (Roediger & Karpicke, 2006a). Put simply, this is the finding that taking practice tests or quizzes benefits memory more than spending an equivalent amount of time restudying (Roediger & Karpicke, 2006b). In most laboratory studies, this benefit is assumed to come from *direct* effects of testing; that is, the act of retrieval itself strengthens the memory trace such that people are better able to recall the same information at a later time. This is why the testing effect is often referred to as "retrieval practice": By practicing retrieval people eventually become better at doing so.

1.1.1 The Influence of Feedback

Although the testing effect is a robust phenomenon, there are many factors that can severely impact the size of an effect. One of these critical factors is the presence or absence of feedback provided on the test. For example, feedback (which is not always, but typically provided in the form of displaying the correct answer) is generally beneficial and maximizes testing effects (Rowland, 2014). Initial performance on the test and the type of feedback received, however, can dramatically affect this relationship. For example, feedback especially helps on items answered incorrectly on the test, and therefore in the presence of ceiling performance there may not be an effect of feedback at all (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Butler & Roediger, 2007; Pashler, Cepeda, Wixted, & Rohrer, 2005).

Even more influential is the type of feedback received on the first test. Different types of feedback have often been categorized by the degree of elaboration, ranging from almost nothing at all (e.g., you got 86% correct), to extreme elaboration (e.g., displaying the correct answer,

explaining why it is the correct answer, and showing you the passage in the primary text where that answer could have been found) (Kulhavy, White, Topp, Chan, & Adams, 1985). The previous two feedback types are rarely used, but a common manipulation is the difference between providing correct/incorrect (C/I) feedback (also referred to as knowledge-of-results), and providing correct answer (CA) feedback (also referred to as knowledge-of-correct-response) (Noonan, 1984). Clearly, CA feedback contains more elaboration than C/I feedback, but which causes greater learning on a final test?

Roper (1977) asked just this question, and found that giving participants no feedback at all was the worst, C/I feedback performed better, and CA feedback was the best on a subsequent final test. Many other studies have found this same pattern, with more elaborative feedback leading to higher performance (Bangert- Drowns et al., 1991; Birenbaum & Tatsuoka, 1987; Fazio, Huelser, Johnson, & Marsh, 2010; Gilman, 1969; Hattie & Timperley, 2007; Pashler et al., 2005; Sassenrath & Garverick, 1965; Travers, Van Wagenen, Haygood, & McCormick, 1964). There are some notable exceptions, however. Andre and Thieman (1988), for example, found that feedback only benefitted factual questions, but not application questions, and that the only difference between C/I and CA feedback was that C/I was better for factual questions that were directly repeated from the initial quiz. Additionally, Hanna (1976) found that C/I feedback benefitted high-ability students, but that CA feedback was better for low-ability students. Lee (1985) found no differences between C/I and CA feedback, and Wentling (1973) actually found that C/I feedback was superior to CA feedback (though, if students answered an item incorrectly they were encouraged to respond again until they got the answer correct; however, see Butler, Karpicke, & Roediger, 2007). Similarly, Noonan (1984) found that C/I feedback (with an explanation of why you got a question wrong) was just as good as CA feedback if participants

were allowed to make a second choice. Finally, McDaniel and Fisher (1991) found no effect of further elaboration on CA feedback (generating a reason why the answer is true, rather than just reading the correct answer).

Therefore, although the majority of studies support the idea that more elaborative feedback (typically in the form of providing the CA) is better than less elaboration (typically in the form of C/I feedback), there appears to be some situations in which this is not the case. Furthermore, although C/I feedback is generally beneficial above not providing feedback at all, there are some studies that find no benefit whatsoever (Fazio et al., 2010; Pashler et al., 2005). This should be qualified, however, by noting that most studies examining the effect of feedback type do not allow a period of restudy, and thereby C/I feedback should not be expected to perform better than receiving CA feedback. This is because CA feedback provides the exact information necessary to answer a similar question at a later time, and a C/I feedback condition typically has no opportunity to identify the correct answers on questions they got wrong. Along those same lines, most laboratory studies use identical, repeated questions on the final test, maximizing the overlap between the CA feedback and the required final test response. In authentic domains, this is rarely the case, with only 25% of instructors claiming they use identical questions, and 75% claiming they use similar questions instead (Wooldridge, Bugg, McDaniel, & Liu, 2014). McDaniel, Thomas, Agarwal, McDermott, and Roediger (2013), for example, showed that middle school science students who were quizzed on definitional questions did not improve performance for application questions on a later test. Thus, by using repeated questions and not allowing participants in a C/I feedback condition to restudy (which would be very uncommon in an educational setting), the claimed benefits of CA feedback for authentic classroom studies may be somewhat inflated in the literature.

1.1.2 Indirect Benefits of Testing on Students' Metacognition

So far, we have described the direct effects of testing, how feedback can moderate this effect, and how educational settings may be different from typical laboratory studies. Despite some of those differences, there is also a growing literature showing the benefits of testing in classroom settings as well (see McDaniel et al., 2013; McDermott, Agarwal, D'Antonio, Roediger, & McDaniel, 2014 for reviews). Although much of this improvement is likely due to direct effects of testing, many researchers have suggested that there are additional opportunities for *indirect* (or mediated) effects of testing to play a role in students' learning (Bjork, 1994; Bjork, Dunlosky, & Kornell, 2013; Dunlosky & Ariel, 2011; Roediger & Karpicke, 2006a; Son & Kornell, 2008). Indirect effects come from such factors as realizing what types of questions may be asked, which concepts are more or less challenging, and improved understanding of what one does and does not know. By taking tests, students are able to gather a host of information that does nothing to directly benefit memory, but can indirectly assist them in the future.

Of the many indirect benefits, one of the most potent appears to be the metacognitive feedback loop between monitoring and control processes in students' learning behaviors (Nelson & Narens, 1994). Metacognition refers to the higher order knowledge of one's own cognitive processes; metamemory, for example, is the accuracy with which someone can gauge what they are able to remember and forget. Nelson and Narens proposed that our metacognitive system operates by monitoring our own behaviors (e.g., "I could have sworn I've seen that man on the bus before, but I cannot recall where!") and engaging control processes to achieve desired behaviors (e.g., executing a controlled memory search to determine that he is, in fact, your butcher). For example, upon learning (via monitoring) that a particular professor cares little about conceptual learning, and instead tests largely on one's ability to memorize or store

information, one might engage control processes to focus their study solely on definitional content.

Further, in typical classroom settings, there are a variety of features that increase the scope and influence of metacognitive monitoring and control processes (McDaniel, Bugg, Liu, & Brick, 2015). In contrast to most laboratory testing effect studies, students typically do not engage in a one-off learning session, but instead have many opportunities to study, take tests or quizzes, and restudy the material before a final test. Additionally, students often have their books, notes, and quizzes available during such a restudy session. The access to all of this information gives students a large degree of freedom to engage control processes and adjust restudy behaviors to suit their perceived needs.

Many studies have examined the influence of participants' metacognitive judgments on subsequent restudy behaviors, with special importance placed on whether items were initially answered correctly or incorrectly on a quiz. Critically, of course, feedback in some form is often provided, which gives participants' even more information to base their restudy decisions on. For example, many studies have shown that participants choose to restudy items they initially answered incorrectly on a quiz. This finding was first documented by Zacks (1969), and later by Kulhavy and Stock (1989). Since then, it has been termed the *discrepancy reduction hypothesis* (Dunlosky & Hertzog, 1998), and states that people seek to reduce the discrepancy between their beliefs about what they think they know and what they actually appear to know by focusing their efforts on items answered incorrectly. The discrepancy reduction hypothesis gains additional nuance when examining participants' item-level (for each question) confidence ratings: The greatest discrepancy should occur for high-confidence items answered incorrectly, and the lowest discrepancy should occur for high-confidence items answered correctly (Kulhavy & Stock,

1989). Likelihood to restudy (and/or amount of time spent restudying) is positively correlated with discrepancy, such that high-discrepancy items are more likely to be restudied than low-discrepancy items, and this study allocation strategy is beneficial for final test performance. The discrepancy reduction hypothesis has been replicated in many studies (Dunlosky & Thiede, 1998; Kulhavy, Yekovich, & Dyer, 1976; 1979; Son & Metcalfe, 2000; Thiede & Dunlosky, 1999), and is supported by other tangential findings such as people choosing to drop-out items from restudy that they feel are well-learned (Pyc & Rawson, 2011).

Although the discrepancy reduction hypothesis has garnered much support, there does appear to be particular boundary conditions influencing the likelihood of people choosing to adopt such a strategy, as well as whether such a strategy is in fact beneficial. Notably, Nelson and Leonesio (1988) found that for some particularly difficult items, no matter how long people spend restudying they are unable to answer such questions correctly on a final test. Termed the *labor-in-vain* effect, this finding challenges the general assumption that it is beneficial for students to continue studying items they initially answered incorrectly. Mazzoni & Cornoldi (1993) also found labor-in-vain effects, but note that they disappeared when experimenter-imposed time limits were enforced.

Findings such as these, in conjunction with the discrepancy reduction literature, spurred Metcalfe (2009) to introduce her *region of proximal learning hypothesis*, which states that students allocate their study time according to a proximal region of items that have steep learning curves. According to Metcalfe, this involves quickly eliminating items identified as particularly easy (i.e., items that one is sure to answer correctly on the final test), deliberately ignoring items judged to be extremely difficult, and focusing the majority of their time restudying items judged as being unknown but easily-learnable. The proximal learning hypothesis is supported by studies

showing that when time constraints are imposed, and when all items are presented simultaneously (as opposed to one at a time), students do tend to follow the general pattern of ignoring especially easy or difficult items (Son & Metcalfe, 2000; Thiede & Dunlosky, 1999). Therefore, although restudying based on a discrepancy reduction control process holds true as a general rule, there are conditions in which hard items may be ignored (e.g., students may choose to cram right before a test, or knowingly aim to achieve a grade of B) in favor of studying items they believe are going to be more efficiently learned.

Of further interest are findings that show an interaction between effects of metacognitive control and the presence or type of feedback received on the initial quiz. We have already noted that feedback is especially beneficial for items that were initially answered incorrectly (Bangert-Drowns et al., 1991), but the metacognitive reasons for why this occurs can be determined by additionally looking at item-level confidence. For example, when CA feedback is provided, items which are answered incorrectly with high confidence (e.g., high discrepancy) are especially likely to be answered correctly on the final test (regardless of an opportunity to restudy). This effect is referred to as hypercorrection, and has been shown in a number of experiments (Butterfield & Metcalfe, 2001; Kulhavy et al., 1976; 1979). Along a related vein, Butler, Karpicke, and Roediger (2008) showed that when CA feedback is provided, low-confidence correct answers were also more likely to persevere as correct answers on a final test than when no feedback was provided (see also, Fazio et al., 2010; cf. Kulhavy et al., 1976). Additionally, Kulhavy and Stock (1989) proposed that more elaborative feedback should benefit high-discrepancy items the most, given that it provides the learner with more information with which to correct their initially wrong responses. Finally, Kulhavy et al. (1985) found that increasing elaboration above and beyond simply providing CA feedback had little effect on final

test performance, but increased study time nonetheless. That is, they found that CA feedback was the most efficient in terms of time spent studying and the subsequent benefits on final test performance.

A final note regarding the potential influences of metacognitive monitoring on control processes during restudy concerns making a particular kind of metacognitive judgment, called a judgment of learning (JOL). JOLs typically take the form of asking participants to look at an item they have previously answered (or studied) and to determine how well they think they will do on a final test for a similar question. JOLs are similar but distinct from confidence judgments, in that both provide a measure of how well one believes they know the information, but JOLs additionally ask participants to project that meta-knowledge into beliefs about later performance. For example, Mazzoni, Cornoldi, and Marchitelli (1990) found that participants were more likely to restudy items they gave low JOLs for (see also Metcalfe & Kornell, 2005). Additionally, when Nelson, Dunlosky, Grag, and Narens (1994) asked participants to make item-level JOLs, and later gave them the items to restudy they rated as having a low JOL, this boosted final test performance (see also Kornell & Metcalfe, 2006 and Thiede, 1999). These findings support a discrepancy reduction view, but Kornell and Metcalfe (2006) additionally showed that when participants were only allowed to restudy items they answered incorrectly on the quiz, participants chose to study items they gave the *highest* JOLs for. They interpret this as support for the proximal learning hypothesis, because participants appeared to selectively restudy items they believed they could learn well.

Though most of the studies reviewed here do not convey perfect metacognitive monitoring, the observed effects do lend credence to the idea that students have some understanding of what they do and do not know, and that they are able to engage beneficial

control processes (cf. labor-in-vain effects). This is not always the case, however, as there is also evidence that students are quite over-confident in their monitoring behavior. Rawson and Dunlosky (2007), for example, asked participants to recall key definitions from previously studied passages; critically, they were also asked to self-score their responses either with or without CA feedback provided to help them (see also, Dunlosky, Rawson, & Middleton, 2005). The degree to which their self-scored responses matched an objective scoring scheme served as the measure of monitoring ability. They proposed two hypotheses: 1- the absence of standard hypothesis, which predicts that students' monitoring ability should be impaired when no feedback is provided, and 2- the limited competence hypothesis, which predicts that students simply *have* impaired monitoring ability, regardless of whether or not the correct answers are provided. They found support for both hypotheses, with participants displaying better metacognitive accuracy in the feedback condition relative to no feedback, but they were still remarkably over-confident even when they had the correct answers at their disposal. Therefore, it should be noted that monitoring ability is by no means perfect even with CA feedback, and students' ability to grade their own responses is especially poor without feedback.

1.1.3 The Present Study

In sum, it appears as though both feedback and metacognitive awareness can have far-reaching effects on both participants' restudy behaviors and final test performance. In addition to the direct effects of testing, there are many indirect effects that can influence the dynamic interplay between taking a quiz, receiving feedback, making restudy decisions based on quiz performance, and eventual performance on a final test. A study by McDaniel et al. (2015) sought to better understand this interplay in an authentic laboratory setting (i.e., using realistic study and test materials, having multiple learning opportunities, and changing the testing/quizzing format)

by splitting participants into a repeated study condition (SSS), a test-restudy-test condition (TST), and a repeated testing condition (TTT). Five days after this, they received a final, short-answer test. In their second experiment, when C/I feedback was provided and participants had access to their quiz (and feedback) during restudy (for the SSS and TST conditions), the TST condition outperformed the TTT condition. They further showed that the number of tested items that were highlighted during restudy was significantly greater for the TST condition than for the SSS condition. These results suggest that the TST condition was more focused in their restudy than the SSS condition, and that these indirect effects of testing on restudy boosted performance above the condition in which the direct effects of testing were greatest (TTT).

In the present experiment, we seek to adapt the McDaniel et al. (2015) methodology to more directly examine the potential indirect effects of taking a test on restudy behaviors, and to determine what impact, if any, feedback has on these effects. Specifically, using the same materials (excerpts from a psychology research methods textbook), all participants took a multiple-choice quiz on this information, and we manipulated the degree of elaboration in feedback they received. One condition received minimal feedback in the form of proportion correct on the quiz, the other condition received C/I feedback, and the final condition received CA feedback. Participants were then allowed to restudy, and we tracked which concepts from the quiz were highlighted to determine whether quiz performance and restudy behavior affected accuracy on the final test. In addition, we obtained metacognitive ratings of confidence for each quiz response, and more broad (not item-level) post-restudy JOLs to determine whether participants had accurate knowledge of their own learning.

One might predict that providing CA feedback will lead to the best retention because it contains the exact information necessary to answer a similar question. However, one could also

predict that C/I feedback will produce superior retention because it could positively influence restudying behavior. Specifically, C/I feedback might help guide restudy by facilitating the correction of wrong answers. Because students are not given the correct answer, they must return to the text, determine the source of their earlier misconception, and generate a new possible correct answer. In doing so, they might develop a more complete understanding of the underlying concepts that future test questions are based on. Along these lines, one could also predict that the minimal feedback condition will obtain the greatest performance, in that they must additionally determine which items they answered incorrectly, and then return to the text to generate a new possible answer. The large degree of generative processing could benefit their underlying conceptual knowledge and translate into high performance on the final test. Alternatively, the minimal feedback condition may not have enough environmental support, causing them to restudy ineffectively, and hurt final test performance. Certainly, the majority of the literature predicts that receiving minimal feedback should lead to the worst performance.

Further, by associating highlighted concepts in the text to both quiz and final test questions, we can directly observe whether people are selectively restudying poorly learned information, and whether this restudying affects final test performance. Finally, confidence ratings and JOLs will allow us to test a variety of metacognitive effects in an authentic laboratory setting. Specifically, we can determine whether participants seem to favor a discrepancy reduction or proximal learning approach, whether students hypercorrect their answers, if feedback positively impacts low-confidence correct answers, and if students make accurate JOLs. Most of the reviewed literature suggests that participants should be able to effectively monitor and engage control processes, but it is unclear whether these effects will be present

using authentic educational materials, or when no feedback is provided (Rawson & Dunlosky, 2007).

1.2 Method

1.2.1 Participants and Design

119 students at Washington University participated in the study for either course credit or monetary compensation ($25). Six participants failed to show up to take the final test 2 days later, leaving our total sample at 113 participants. Our design was a 2 x 2 x 3 mixed factorial containing the within-subjects variables question type (definition, application) and question stem (same, different), as well as the between-subjects variable feedback condition (C/I, CA, minimal). Participants were randomly assigned to each feedback condition, resulting in 35 participants in the C/I condition, 40 participants in the CA condition, and 38 participants in the minimal feedback condition.

1.2.2 Materials

Our materials were borrowed and adapted from McDaniel et al. (2015), and used the same 20 concepts found in a 38 page packet excerpted from *Research Methods in Psychology, 3rd edition* (Heiman, 2002). The packet was presented as a PDF using Adobe Reader. Two versions of the quiz (A, B) were created containing 40 multiple-choice questions, with both a definition and application question for each of the 20 concepts. As an example, a definition question for the concept reliability was:

What is reliability?

(a) The extent to which a procedure measures what it is intended to measure.

(b) The extent to which our results generalize to other participants and other situations.

(c) The extent to which a measurement reflects the hypothetical construct of interest.

(d) The extent to which a measurement is consistent, can be reproduced, and avoids error.

By contrast, the application question for reliability was:

A student complains to her professor that her essay makes the same points as her friend's but she got a lower grade than her friend. She is complaining that the grading lacks _____.

(a) Internal validity

(b) External validity

(c) Reliability

(d) Concurrent validity

Both versions of the quiz tested identical concepts in the same order of presentation. However, quiz versions differed by changing the question stem. For example, the first question in quiz version A asked:

_____ is the extent to which a procedure measures what it is intended to measure.

(a) Validity

(b) Reliability

(c) External Validity

(d) Subject Attrition

The first question in quiz version B, by contrast, asked:

> What is validity?
>
> (a) The extent to which a procedure measures what it is intended to measure.
>
> (b) The extent to which our results generalize to other participants and other situations.
>
> (c) The extent to which a measurement reflects the hypothetical construct of interest.
>
> (d) The extent to which a measurement is consistent, can be reproduced, and avoids error.

For application questions, changing the question stem meant that the two quiz versions received different problems or scenarios that targeted the same concept.

The first half of the quiz contained one question per concept (half definition and half application), and the second half of the quiz tested the same concepts but changed whether they were definition or application questions (i.e., those that were definition questions in the first half were application questions in the second half). The presentation order of the questions was initially randomized within each half to create the two quiz versions, but all participants received questions in the same order. That is, the presentation order of quiz questions was not randomized on a case-by-case basis (e.g., the first question for all participants was a definition question about validity; only the question stem changed). However, response options for each question were randomized on a case-by-case basis.

Following submission of the quiz (via Blackboard educational system software), participants received feedback in accordance with their condition. Participants in the minimal feedback condition were shown the quiz questions, the answers they selected, and an overall score on the quiz (e.g., 34 out of 40). Participants in the C/I condition were shown the overall

score, the quiz questions, the answers they selected, and whether each question was correct or incorrect. Participants in the CA condition were shown all of this information, but a green checkmark was located next to correct answer (and red X's next to each incorrect answer).

The final test contained 40 short-answer questions, with a definition and an application question for each of the 20 concepts. One question (the application question on confounding variables) was eliminated because it did not accurately assess knowledge of the intended concept, leaving 39 final test questions. As in McDaniel et al. (2015), eight definition questions and eight application questions used the same question stems as in the quizzes. Twelve definition questions and 12 application questions used different question stems from those in the quizzes. McDaniel et al. (2015) randomly determined a priori whether each question on the final test used the same-stem or a different-stem from the initial quiz. Therefore, 16 items from the quiz (eight definition and eight application) were always presented as same-stem questions and the remaining 24 items (12 definition and 12 application) were always presented as different stem questions (see above for examples of changing the question stem), though which of these items were same-stem or different-stem changed depending on which version of the quiz the participant received. In an effort to change the order in which each concept was presented between quiz and final test, we randomly determined a priori the position of each question in the final test (not on a case-by-case basis). That is, every participant received the final test questions in the same order, but this order was randomly determined and different from the order of the quiz questions.

1.2.3 Procedure

Participants came to the lab and were initially asked whether or not they had taken a class on research methods (notable examples in the psychology department were provided). They were

not allowed to participate if they answered yes to this question. Next, they were informed that they would be studying a packet on research methods in psychology, would take a quiz on this information, have access to their quiz while they restudied the packet, and come back 2 days later for a final test.

Participants were given 1 hour to read a PDF of the textbook excerpts presented on the computer and were encouraged to simply read through, rather than study as they went, as the packet was fairly long (38 pages). If participants finished early, they were told they could go back and reread sections if they wanted, and if they did not quite finish reading the entire packet, they were told not to worry.

Following the initial study session, participants played Tetris for 5 min before proceeding to take the quiz. The quiz was administered using Blackboard educational system software, with a scrollable window. Therefore, they were allowed to go back and change their answers, as would be the case in most educational settings. The quiz was untimed, but took most participants approximately 15-20 min. Following each quiz response, participants rated their confidence on a 1-10 scale, and were encouraged to use the full range of the scale, rather than simply respond using 1s and 10s only.

After they submitted the quiz, feedback was provided by Blackboard according to their condition (see Materials). We had participants sitting at a computer with dual-monitors (one with the textbook excerpts and one with the quiz), so the experimenter turned on the other monitor with the textbook excerpts, and informed participants that they could look at their quiz if they chose while they restudied. Fifteen minutes were provided to restudy, and participants were asked to highlight the paragraphs they restudied using the Adobe Reader highlight tool. Passages

within the text were coded according to what concept was being targeted, as well as whether the concept was being defined or explained using applied examples. In this way, we were able to code highlighting behavior as targeting the definition or application level of each concept.

Following restudy, participants were asked to make category-level JOLs regarding the information they had studied, been quizzed on, and restudied. Specifically, on a scale of 1-10 they were asked to rate how well they thought they would perform on the final test for questions concerning: Reliability and Validity (e.g., external vs. internal), Research Designs (e.g., between vs. within), Potential Confounds (e.g., variables that vary systematically with the independent variable), Research Variables (e.g., independent vs. dependent), and Conditions (e.g., control condition). After 2 days they returned to take the final short-answer test, which was untimed, but took most participants approximately 30-40 min. Total time for both sessions took approximately 2.5 hours (see Appendix 1.6.1 for a complete list of instructions).

1.3 Results

1.3.1 Analysis Methods

For all analyses, the alpha level was set to .05. In order to examine mean differences in final test performance among our between-subjects feedback conditions (C/I, CA, minimal) by the within-subjects variables question type (definition, application) and question stem (same, different), we employed a 2 x 2 x 3 mixed analysis of variance (ANOVA). In order to look at the item level influence of quiz performance (incorrect, 0; correct, 1) on restudy behavior (no restudy, 0; restudy, 1), as well as the influence of quiz performance and restudy behavior on final test performance (incorrect, 0; correct, 1), we used multilevel logistic regression with participant intercepts as a random factor. We used logistic regression because our dependent outcomes were

all binary, and we allowed intercepts to vary between participants because responses were nested within each subject. Because of the nesting within the data, the assumption of independence would be violated using a standard logistic model (i.e., individual cases should be correlated within participants). Additionally, the coding scheme for the variable question type was definition (0) and application (1), and the scheme for question stem was same-stem (0) and different-stem (1). Multilevel modeling was performed using the *R* package *lme4* (Bates, Maechler, Bolker, &Walker, 2015) and regression figures were generated using the package *sjPlot* (Lüdecke, 2017).

Quiz performance, restudy behavior, and final test performance were all matched by concept and question type. In this way, for example, we were able to see if a participant answered a definition question on validity correctly, whether they studied the section of the text providing the definition, and whether they answered a same or similar final test question on the definition of validity correctly. Final short-answer test coding was done by the principal investigator, whereas quiz and restudy behavior were coded from a rubric by two different research assistants. Partial credit was available on the final test, and scores generated by this method were used for the ANOVA analyses. For our logistic regression analyses, however, we decided to use a strict coding scheme because of the generally high performance on the final test (M = .75, SE = .01), and to maintain a binary format. We decided to use a strict rather than lenient coding scheme based on overall final test performance, but before running our logistic models. Strict coding meant that answers which received partial credit (.5) were coded as incorrect (0).

1.3.2 ANOVA Analyses on Final Test Accuracy

To determine what effect our three feedback conditions had on final test performance, as well as whether our feedback conditions differentially impacted our within subjects variables question type and stem, we ran a 2 x 2 x 3 mixed ANOVA. There was a main effect of question type, $F(1, 103) = 6.21$, $p < .05$, $\eta^2 = .06$, indicating that application questions were more often answered correctly ($M = .76$, $SE = .01$) than definition questions ($M = .74$, $SE = .02$). There was also a main effect of question stem, $F(1, 103) = 4.28$, $p < .05$, $\eta^2 = .04$, indicating that same-stem questions were more often answered correctly ($M = .77$, $SE = .02$) than different-stem questions ($M = .74$, $SE = .02$). There was no effect of condition, nor was there any interaction between condition and question type or question stem, all F's < 1. Therefore, it appears as though our feedback manipulation had no effect on final test performance (see Figure 1).

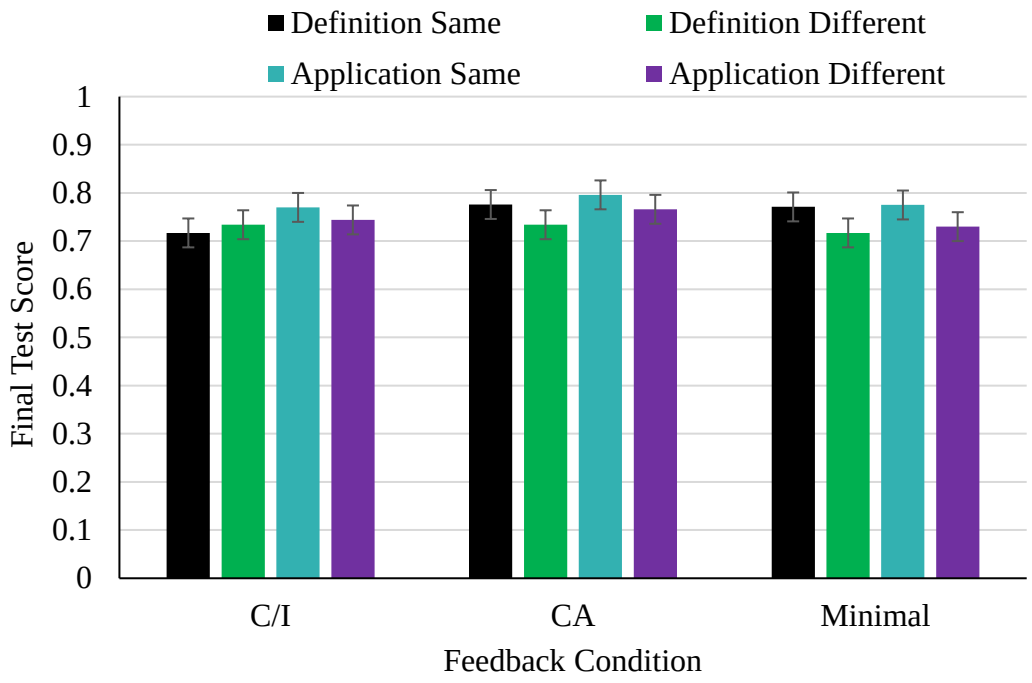

Figure 1. Final test performance by feedback condition [correct/incorrect (C/I), correct answer (CA), minimal], question type (definition, application), and question stem (same, different).

There was no effect of feedback condition, nor any interactions with condition. There was a significant advantage for application over definition questions, and a significant advantage for same-stem over different-stem questions.

1.3.3 Multilevel Logistic Regression Analyses on Final Test Accuracy

Model 1. We ran a logistic mixed effects regression analysis to determine whether initial quiz performance influenced restudy behavior, and whether this in turn influenced final test performance (see Analysis Methods section for coding details). That is, we allowed quiz accuracy and highlighting behavior to interact when predicting final test accuracy. Further, we included both question type and question stem in our model as factors to determine whether definition or application questions were more likely to be answered correctly, as well as whether the use of same or different stems from the quizzes affected performance. Finally, we tested whether our 3 feedback conditions obtained different patterns in our model. Condition did not interact with quiz accuracy or highlighting, so we included it only as an individual factor. Therefore, all results are collapsed across feedback conditions.

The results of model 1 can be seen in Table 1. Significant results are best interpreted in terms of the odds ratio (OR), which are the odds that an outcome will occur (in this case, answering an item correctly or restudying an item), compared to the odds of the outcome not occurring (in this case, answering an item incorrectly, or not restudying an item). In instances of a negative association (i.e., an OR < 1), we reversed the directionality to maintain a consistent format (e.g., we changed *more* likely to *less* likely). We found no significant effect of highlighting on final test accuracy ($OR = 1.04$, $p > .05$), suggesting that participants were just as likely to answer a final test question correctly regardless of whether or not they had highlighted the concept. Quiz accuracy was a significant predictor ($OR = 2.69$, $p < .001$), showing that

participants were 2.67 times more likely to answer a final test question correctly if they also answered it correctly on the quiz. As can be seen in Figure 2, there was also a significant interaction between quiz accuracy and highlighting behavior, which indicates that if participants answered a quiz question correctly, they were far more likely to also answer it correctly on the final test if they did *not* restudy it. There was a significant effect of question stem (*OR* = 1.19, $p < .05$) indicating that participants were 1.19 times more likely to answer same-stem questions correctly than different-stem questions. Finally, neither question type (*OR* = .94, $p > .05$) nor condition (*OR* = 1.02, $p > .05$) appeared to significantly impact final test performance.

Table 1.

Results from Model 1, showing the interaction between quiz accuracy and highlighting behavior on final test accuracy.

	Final Test Accuracy		
	Odds Ratio	*CI*	*p*
Fixed Parts			
(Intercept)	1.52	0.94 – 2.45	.089
Highlight (no restudy, restudy)	1.04	0.74 – 1.45	.831
QuizAccuracy (incorrect, correct)	2.69	2.14 – 3.37	**<.001**
QuestionStem (same, different)	0.84	0.73 – 0.97	**.016**
QuestionType (definition, application)	0.95	0.82 – 1.10	.516
Condition	1.02	0.86 – 1.21	.815
Highlight:QuizAccuracy	0.46	0.32 – 0.67	**<.001**
Random Parts			
$\tau_{00,\ Subject}$		0.371	
$N_{Subject}$		106	

$ICC_{Subject}$	0.101
AIC	4704.9
Observations	4129
Deviance	4476.607

Note. ICC = Intraclass correlation coefficient; a measure of the within-subjects correlation in responses. τ_{00} = the variance between subjects. AIC = Akaike information criterion; a measure of model fit (when comparing related models, lower values indicate better fit). Odds Ratio = the odds that an outcome will occur (in this case, answering an item correctly or restudying an item), compared to the odds of the outcome not occurring (in this case, answering an item incorrectly, or not restudying an item). See Analysis Methods section for coding details. Significant effects are noted in bold.

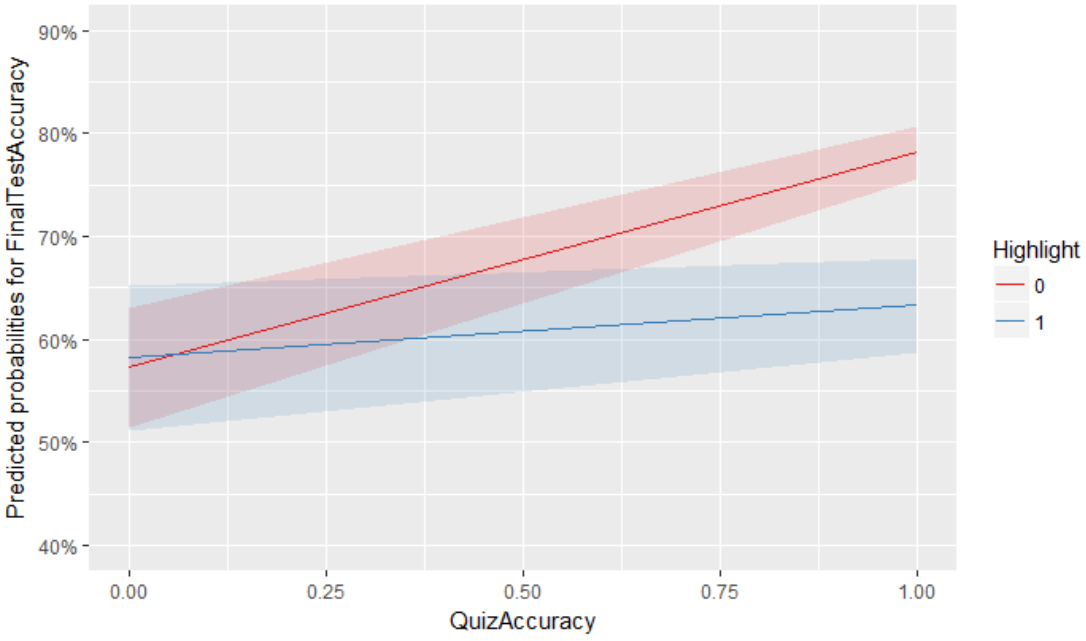

Figure 2. The interaction between quiz accuracy (0, 1) and highlighting behavior (0, 1) on final test performance (0, 1). The interaction shows that participants were more likely to answer final test questions correctly for items that were initially right on the quiz but were not restudied.

Model 2. At first glance, the highlighting results from model 1 appear to contradict what one might intuitively expect. That is, it looks as though restudying items (that were initially answered correctly on the quiz) actually hurt performance on the final test. Based on this observation, we theorized that there could be a bias in the data, such that participants selectively chose to restudy more difficult items, regardless of whether or not they answered that item correctly on the initial quiz. To test this, we averaged across participants the final test performance for each individual item, and tested whether item average score predicted the likelihood to restudy. There was indeed a bias, with participants being 11.39 times more likely to highlight difficult concepts than easier ones.

Therefore, we decided to incorporate item average score as a factor in model 2 so that all other effects could be interpreted at average levels of item difficulty. All other factors were identical to model 1. The results of model 2 are presented in Table 2, and initial inspection of the Akaike information criterion (AIC; lower values indicate better model fit) indicated that model 2 fit the data dramatically better. Further, the intraclass correlation coefficient (ICC; the within-person correlation) increased from .10 to .21, suggesting that our model better accounted for within-person variance.

Table 2.

Results from Model 2, showing the interaction between quiz accuracy and highlighting behavior on final test accuracy, at average levels of item difficulty.

	Final Test Accuracy

	Odds Ratio	*CI*	*p*
Fixed Parts			
(Intercept)	0.03	0.01 – 0.06	**<.001**
Highlight (no restudy, restudy)	1.57	1.05 – 2.34	**.027**
QuizAccuracy (incorrect, correct)	2.24	1.73 – 2.91	**<.001**
QuestionStem (same, different)	0.81	0.69 – 0.96	**.017**
QuestionType (definition, application)	1.03	0.87 – 1.23	.703
Condition	1.00	0.79 – 1.28	.981
ItemAvgScore	374.64	244.44 – 574.20	**<.001**
Highlight:QuizAccuracy	0.62	0.40 – 0.98	**.039**
Random Parts			
$\tau_{00,\ Subject}$		0.858	
$N_{Subject}$		106	
$ICC_{Subject}$		0.207	
AIC		3662.8	
Observations		4129	
Deviance		3382.303	

Note. ICC = Intraclass correlation coefficient; a measure of the within-subjects correlation in responses. τ_{00} = the variance between subjects. AIC = Akaike information criterion; a measure of model fit (when comparing related models, lower values indicate better fit). Odds Ratio = the odds that an outcome will occur (in this case, answering an item correctly or restudying an item), compared to the odds of the outcome not occurring (in this case, answering an item incorrectly, or not restudying an item). See Analysis Methods section for coding details. Significant effects are noted in bold.

Unlike in model 1, we now found a significant effect of highlighting (*OR* = 1.57, $p < .05$), indicating that participants were 1.57 times more likely to answer a final test question

correctly if they had highlighted that concept. We again found a significant effect of quiz accuracy (*OR* = 2.23, $p < .001$), meaning that participants were 2.23 times more likely to answer a final test question correctly if they answered the quiz item correctly. As seen in Figure 3, quiz accuracy and highlighting behavior significantly interacted, which displays that participants were more likely to answer a final test question correctly if they highlighted a concept after getting it wrong on the quiz (at average levels of item difficulty). Again, participants were more likely to answer same-stem questions correctly than different-stem questions (*OR* = 1.23, $p < .05$). As in model 1, neither question type (*OR* = 1.02, $p > .05$) nor condition (*OR* = 1.00, $p > .05$) appeared to significantly impact final test performance.

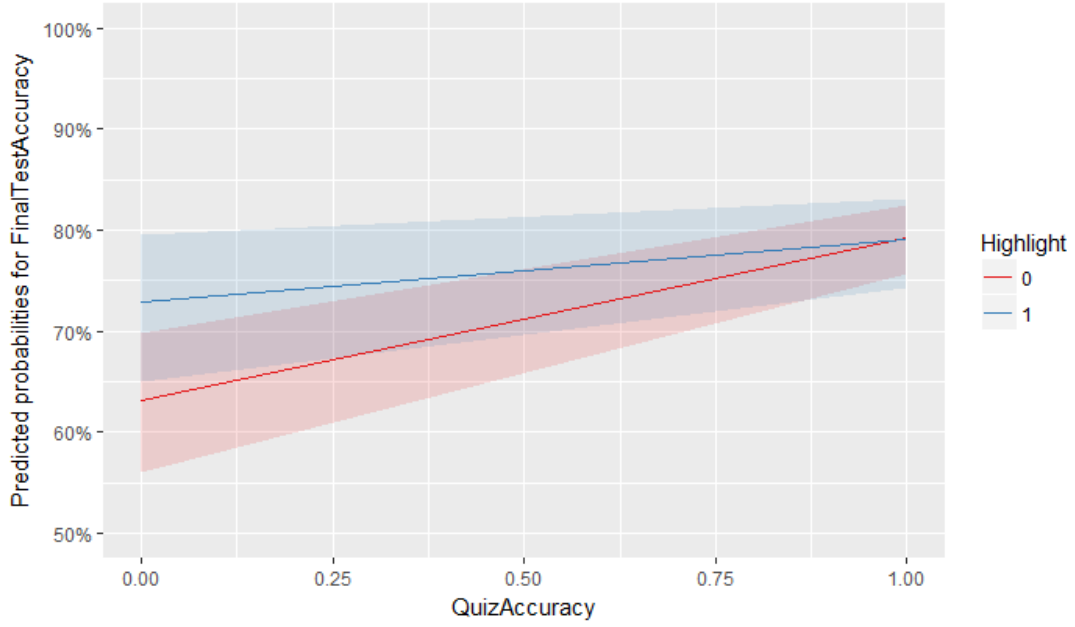

Figure 3. The interaction between quiz accuracy (0, 1) and highlighting behavior (0, 1) on final test performance (0, 1) at average levels of item difficulty. The interaction shows that participants were more likely to answer final test questions correctly for items that were initially wrong on the quiz if they were restudied.

1.3.4 Multilevel Logistic Regression Analyses on Metacognitive Influences

Discrepancy Reduction. In order to test the discrepancy reduction hypothesis of restudy behavior, we sought to determine whether participants selectively restudied items they answered incorrectly on the quiz (Dunlosky & Hertzog, 1998). A one-factor model predicting highlighting behavior by quiz accuracy confirmed this hypothesis, indicating that participants were 2.06 times more likely to restudy items they answered incorrectly on the quiz ($OR = 2.06$, $p < .001$). Additionally, the discrepancy reduction hypothesis states that the highest discrepancy should be for high-confidence items answered incorrectly, and the lowest discrepancy should be for high-confidence items answered correctly (Kulhavy & Stock, 1989). Indeed, as can be seen in Figure 4, there was a significant interaction between confidence and quiz accuracy on highlighting behavior ($p < .001$). The interaction shows that participants were more likely to restudy high-confidence errors on the quiz than high-confidence successes, supporting discrepancy reduction. Interestingly, the interaction also reveals that low-confidence successes were more likely to be restudied than low-confidence errors. There was no interaction with feedback condition, $p > .05$.

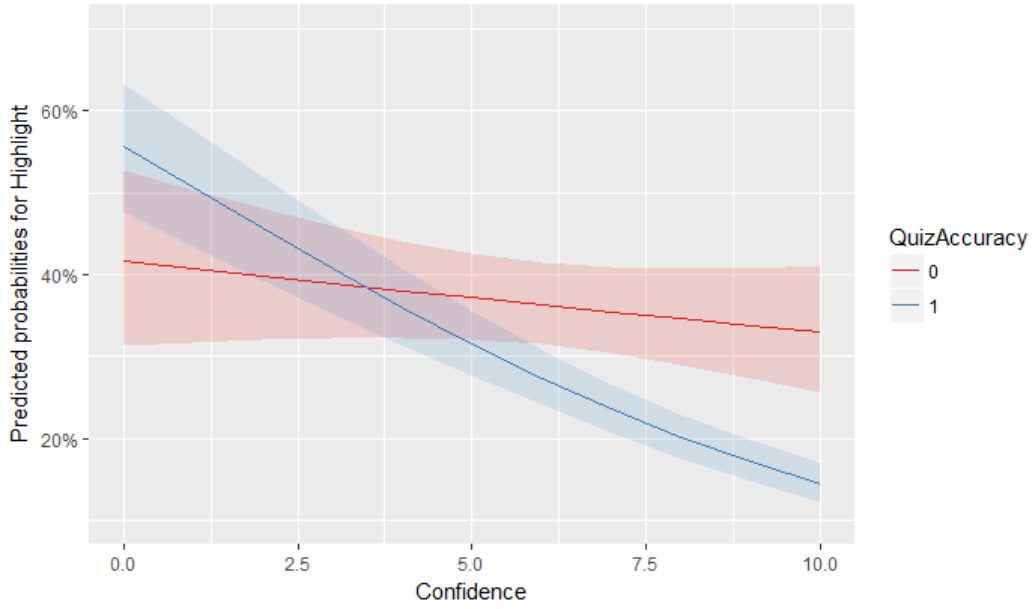

Figure 4. The interaction between quiz accuracy (0, 1) and confidence (1-10) on highlighting behavior (0, 1). The interaction shows that participants were more likely to restudy concepts they answered incorrectly with high confidence, relative to high-confidence correct answers. Additionally, participants were more likely to restudy low-confidence correct answers relative to low-confidence incorrect answers.

Region of Proximal Learning. To test the region of proximal learning hypothesis, we wanted to see if restudy behavior was impacted by item difficulty. Specifically, the proximal learning hypothesis (Metcalfe, 2009) states that participants should selectively restudy items that are neither too difficult nor too easy. Therefore, we set quiz accuracy to interact with item average score in predicting highlighting behavior. The interaction was significant ($p < .01$), and as seen in Figure 5, participants were more likely to study easy items when they answered them incorrectly on the quiz than when they answered them correctly. On one hand, participants do seem to quickly eliminate the particularly easy items (i.e., choosing not to study easy items

answered correctly). As we described in the Model 2 section, however, there also appears to be a general bias towards restudying more difficult items. This finding is in contrast to the proximal learning hypothesis and more in line with a discrepancy reduction view. There was no interaction with feedback condition, $p > .05$.

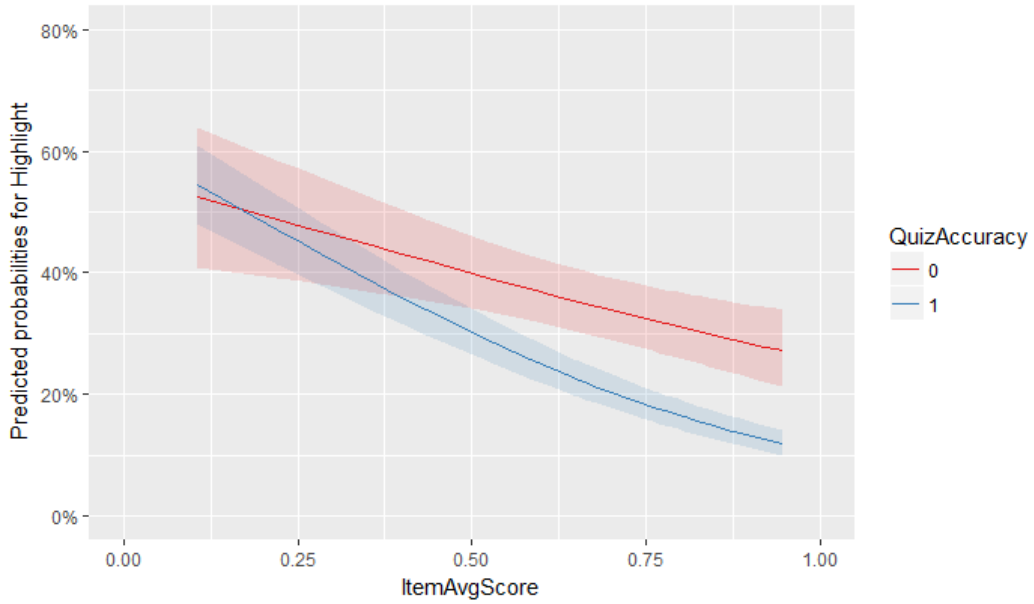

Figure 5. The interaction between quiz accuracy (0, 1) and item average score on highlighting behavior (0, 1). The interaction shows that participants were more likely to restudy easier concepts that they answered incorrectly, relative to easier concepts answered correctly.

Hypercorrection and Perseverance of Low-Confidence Successes. Butterfield and Metcalfe (2001) note that high-confidence failures are more likely to be corrected on the final test, a finding referred to as hypercorrection. Additionally, Butler et al. (2008) found that CA feedback enabled low-confidence successes to perseverate into success on a final test. To test whether our results replicated these findings, we looked at the interaction between quiz accuracy and confidence on final test performance ($p < .01$). As seen in Figure 6, participants were more

likely to answer a question correctly on the final test if they got it wrong on the quiz with high confidence than if they got it wrong with low confidence (e.g., participants hypercorrected their responses). Butler and colleagues found perseveration of low-confidence successes when comparing responses that received CA feedback to responses that received no feedback. Therefore, we let confidence, quiz accuracy, and feedback condition interact. We failed to replicate the Butler et al. (2008) findings, in that low-confidence successes were no more likely to be answered correctly than low-confidence errors. Further, as can be seen in Figure 7, there were no differences in this pattern between the CA and minimal feedback conditions ($p > .05$).

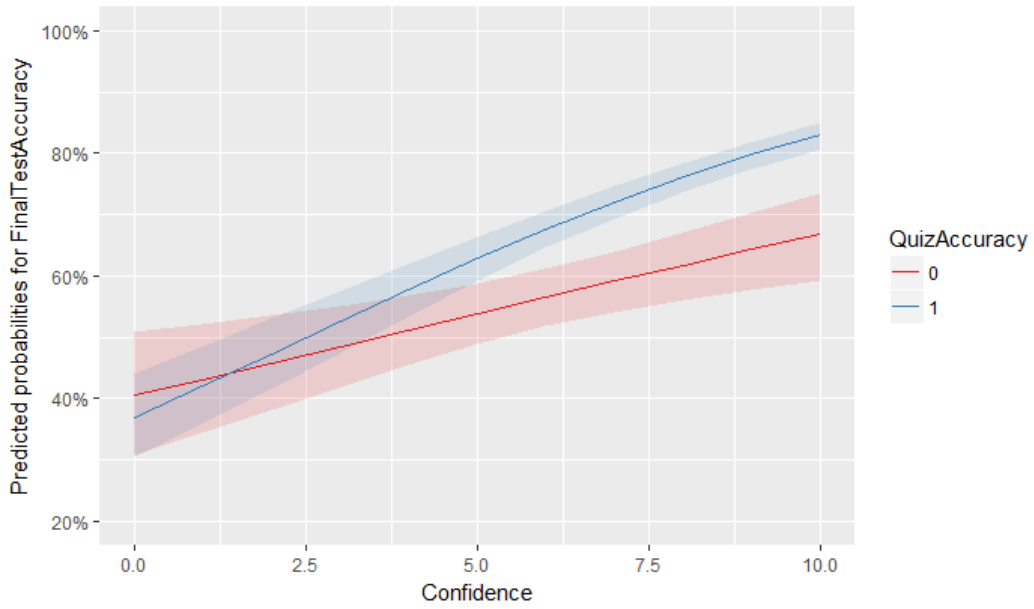

Figure 6. The interaction between quiz accuracy (0, 1) and confidence (1-10) on final test accuracy (0, 1). The interaction shows that participants were more likely to answer final test questions correctly for high-confidence successes on the quiz, relative to high-confidence errors on the quiz.

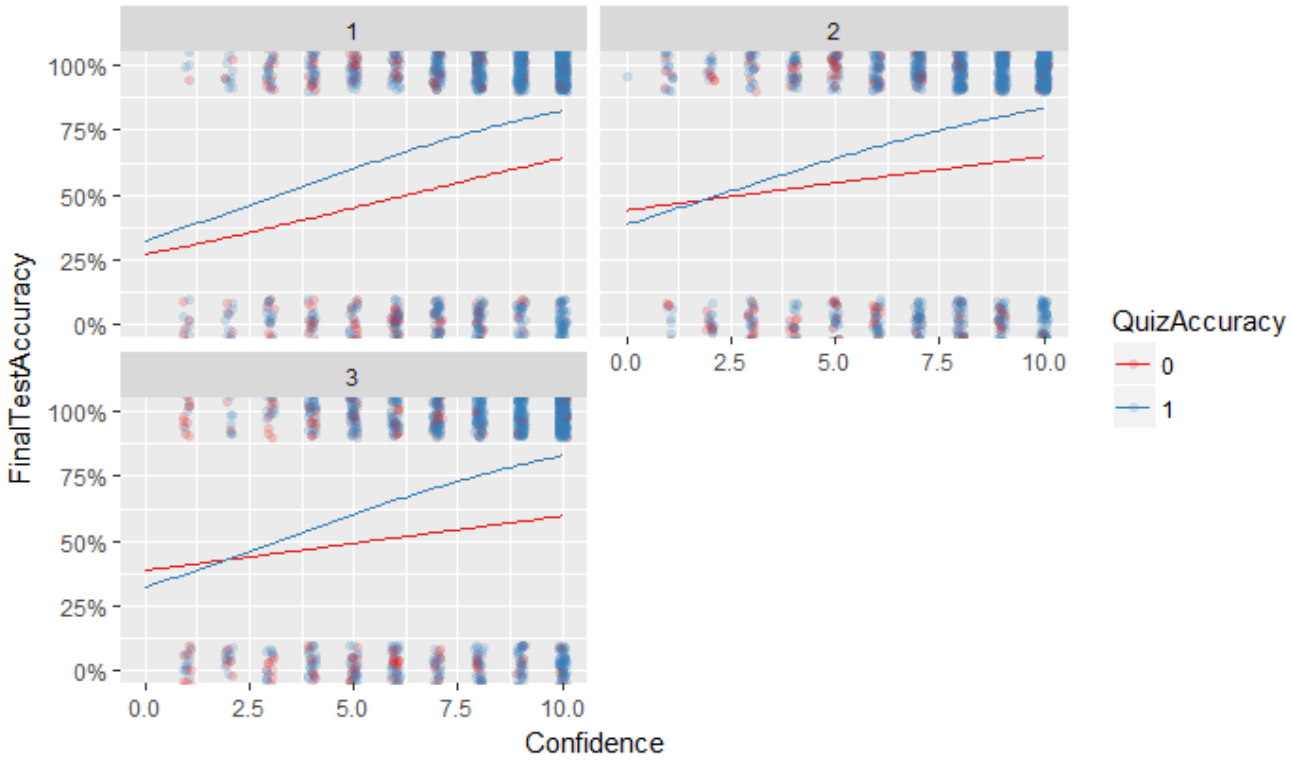

Figure 7. The interaction between quiz accuracy (0, 1) and confidence (1-10) on final test accuracy (0, 1). Panel 1 = C/I feedback, Panel 2 = CA feedback, Panel 3 = minimal feedback. This figure displays that there were no differences among conditions for the interaction effect in Figure 6.

Judgments of Learning. To determine whether or not participants' JOLs were accurate predictors of their final test performance, we coded each item on the final test as to whether or not it matched our five broad JOL categories (reliability and validity, research designs, potential confounds, variables, and conditions). Not all questions were included, such as questions concerning simple random sampling or the use of confederates. Because JOLs were given after restudy (on a scale of 1-10), we were only interested in whether participants' JOLs were correlated with final test accuracy by concept category. First, we found using a one-way

ANOVA that there were no differences in average JOL ratings among conditions, $F(2, 112) = .07$, $p > .05$ (C/I: $M = 7.85$, $SE = .17$; CA: = 7.80, $SE = .20$; minimal = 7.89, $SE = .17$). Next, we sought to determine whether participants JOLs were well-calibrated with actual performance on the final test. Briefly, calibration is the correlation between predicted performance and actual performance across participants (e.g., for JOL ratings of 5, perfect calibration would predict average performance of 50%).

For all categories there was a significant positive correlation between JOLs (divided by 10) and performance on questions targeting that concept (all p's $< .05$). However, there was also a positive correlation with nearly all other concept categories (e.g., JOLs for Reliability and Validity were highly correlated with questions concerning Experimental Designs), and all JOLs were intercorrelated. Therefore, we collapsed across concept category, and found that average JOLs were substantially correlated with final test accuracy, $r(113) = .57$, $p < .001$. This finding did not differ by feedback condition, $p > .05$. When we plot these results as a calibration graph (see Figure 8), the diagonal line represents perfect calibration, with points above the diagonal representing under-confidence, and those underneath the diagonal representing over-confidence. Based on the graph, participants seemed to make fairly accurate predictions of final test performance based on their average JOL, though they were somewhat under-confident.

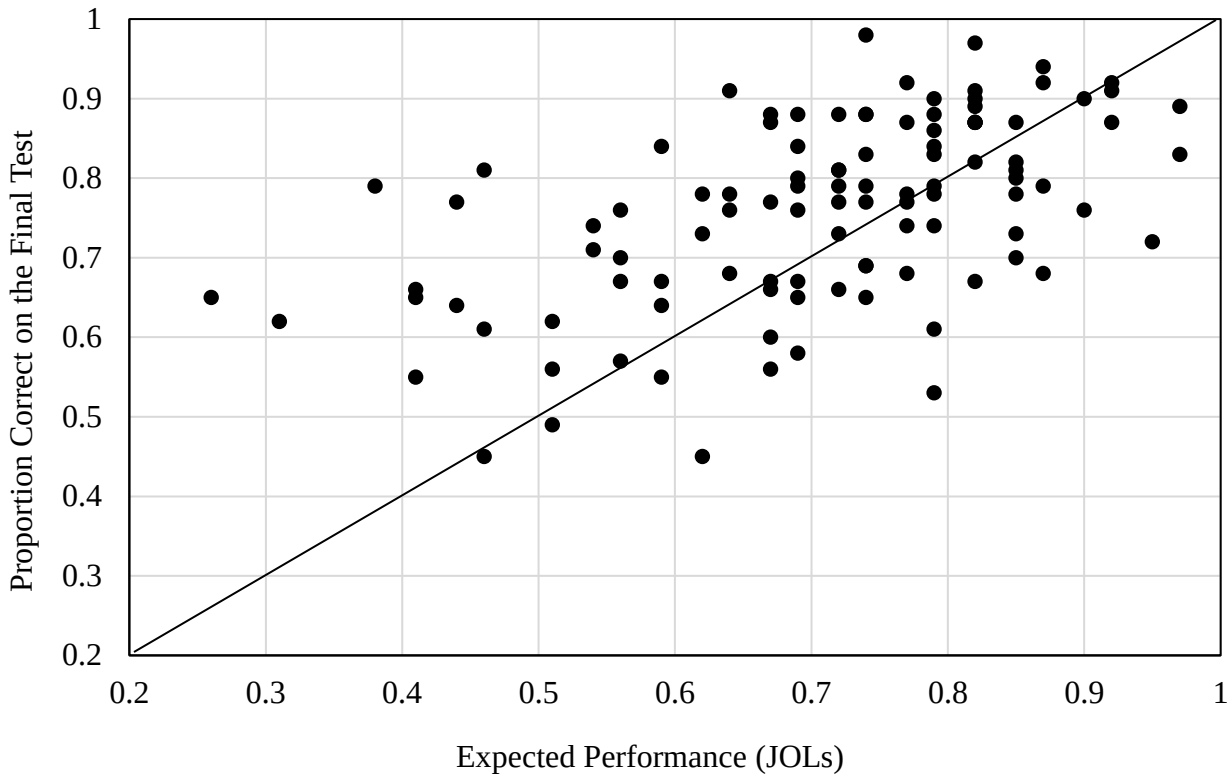

Figure 8. Calibration between participants expected performance (as measured by JOLs) and actual final test performance. JOLs were divided by 10 to maintain the same scale (e.g., a JOL of 5 became .50). The diagonal represents perfect calibration; therefore, data points above the diagonal represent under-confidence, and data points below the diagonal represent over-confidence. Therefore, participants appear to be somewhat under-confident.

1.4 Discussion

1.4.1 The (Null) Effects of Feedback on Final Test Performance

In the present study, we were interested in the interplay between quizzing, restudy behavior, and final test performance, as well as whether this differed depending on the type of feedback one received on the initial quiz (C/I, CA, minimal). Perhaps surprisingly, we found no

effect of feedback type on final test performance, though most of the testing literature manipulating feedback suggests that more elaborative feedback (in this case, CA feedback) leads to better long-term retention. Most of this literature, however, uses highly controlled laboratory research that departs dramatically from testing scenarios in authentic settings. Our study placed high priority on mimicking a realistic scenario, and contained several features that could explain the lack of an effect of feedback.

Unlike most laboratory studies, we not only allowed participants a restudy opportunity, but they were additionally allowed to look at their quiz during this period. Rather than using word-lists or foreign language vocabulary (as is typical in most testing studies), we used excerpts from a published textbook on research methods in psychology. Whereas application questions are frequently neglected (e.g., in the Rowland, 2014 meta-analysis 81% of studies were limited to either word list recall or paired associates—situations in which application questions are impossible), we tested participants on both definition and application questions; further, our final test contained both repeated and non-repeated questions. Finally, we varied the quiz and testing format, such that participants were quizzed on multiple-choice questions, but were tested with short-answer questions (in the Rowland, 2014 meta-analysis 65% used identical test formats). Some combination of these factors could have reduced performance in our C/I and CA feedback conditions (i.e., the overlap between feedback on the quiz and correct final test answers was far less than in most testing studies) and/or boosted performance in the minimal feedback condition (the ability to look previous quiz questions and the confidence in their answers appears to have been enough information to successfully guide restudy).

To be more specific, we may have obtained an effect of feedback type if we had lessened the environmental support in the minimal feedback condition or boosted the overlap between the

quiz and final test questions (which should maximize the benefits of feedback). Examining the minimal feedback condition first, it is useful to imagine what the learning experience may have been like for these participants. First, they answer quiz questions and rate their confidence in each response. When they are provided minimal feedback, it is apparent to the learner whether or not they performed well overall (e.g., they were shown their overall score on the quiz; 34/40). Perhaps the learner then looked through the quiz to determine which items they were unsure about (and they can easily assess this because their confidence ratings are available). It is likely that many of their low-confidence answers were incorrect, and they can further begin to look at the multiple-choice response options to hypothesis test (i.e., seek out their own feedback by reading relevant sections in the textbook) which option may have been the correct one. Potentially, this was enough information to guide restudy and display equivalent final test performance to the other feedback conditions.

Conversely, the C/I and CA feedback conditions may not have benefitted to the same degree as would be expected because we did not exclusively use repeated questions. A participant in either feedback condition, for example, might look at the questions they answered incorrectly and be able to glean from the feedback and response options alone what they believe the correct answer should have been (without returning to the text). This would have helped them answer an identical question on the final test, but their understanding of the underlying concepts may not have been complete. Therefore, when a final test question was different in both format and question stem (e.g., they were asked to provide a short-answer to an entirely different question targeting the same concept) the participant may not have been up to the task. We can partially address this claim by looking at differences in performance based on question stem. Although we found no interaction with condition, participants were better able to answer final

test questions containing the same stem as the previous quiz question than those containing different stems (see Figure 1). If the above claim is true, there should have been a significant interaction such that the C/I and CA conditions performed worse on different-stem questions than the minimal feedback condition. It could be that we simply did not have the power to detect such an effect, but in fact we obtained .94 power to detect a medium-sized within/between interaction. Therefore, perhaps the best interpretation is that we boosted performance in the minimal feedback condition, rather than hurt the C/I and CA conditions.

Another possible reason why we did not find an effect of feedback type is that quiz performance was near-ceiling (84% correct), and prior research shows that feedback is especially helpful for items initially answered incorrectly (Bangert-Drowns et al., 1991; Butler & Roediger, 2007; Pashler et al., 2005). Why we obtained higher performance than McDaniel et al. (2015), though we used the same subject pool and materials, can be explained by our retention interval. McDaniel and colleagues used a 5 day delay between their restudy session and the final test, whereas we only used a 2 day delay. Another possibility is that our selection methods allowed people with more prior knowledge to participate. Though we excluded participants who admitted taking a previous class on research methods, participants may have had higher levels of prior knowledge than desired. However, we did ask participants to gauge their prior knowledge of the material as part of a post-experimental questionnaire (on a scale of 1-10), and including this factor as a covariate in our ANOVA analyses did not affect the results (average prior knowledge $M = 4.84$).

1.4.2 Metacognitive Influences on the Interplay between Quizzing and Restudy

Though we did not obtain a significant effect of feedback type, our design enabled us to examine the item-level influences of quizzing and restudy on final test performance. In model 2 (see Table 2 and Figure 3) we showed that participants were more likely to answer a final test question correctly if they got it right on the quiz, that restudying the text led to greater final test performance, and that highlighting was only beneficial if they initially answered the quiz question incorrectly. Further, we found that these effects did not differ depending on the type of feedback participants received on the quiz. This is important for several reasons: First, people seemed able to identify which questions they answered correctly or incorrectly on the quiz, even in the minimal feedback condition. That is, participants displayed accurate metamemory even without much feedback, and were able to isolate questions they were uncertain about. Second, people were able to selectively target items for restudy that were objectively more difficult (i.e., concepts with lower average scores on the final test), as well as items that were individually more difficult (those they answered incorrectly on the quiz). Again, these effects did not differ by condition, suggesting that participants' metacognitive monitoring in the minimal feedback condition was as good, or nearly as good, as those given item-level feedback. Finally, by selecting initially incorrect items to restudy, people improved their performance on the final test relative to those items they did not restudy.

These results are in direct opposition to Rawson and Dunlosky (2007; Dunlosky et al., 2005), who found that participants were over-confident in their self-scored responses even when CA feedback was provided, and that when no feedback was provided, participants were especially bad at self-scoring their responses (i.e., they displayed poor monitoring ability). Further, it is not immediately clear why this divergence occurred. Both their study and ours used

authentic materials (they used excerpts from textbook chapters), our study also contained definition questions as theirs did, and both studies had conditions which provided CA feedback. The only notable differences were that we provided minimal feedback, whereas they provided no feedback, and we quizzed participants using a multiple-choice format rather than cued-recall as they did. As mentioned previously, perhaps a combination of knowing generally how well one performed (e.g., 34/40 correct) when combined with an ability to simultaneously compare multiple-choice options was enough information to accurately judge the correctness of each response. Further, our participants were able to see the confidence ratings they gave for each question, and may have been able to effectively eliminate high-confidence responses from their considerations.

Targeting the metacognitive influences within a quizzing and restudy scenario more specifically, we were able to test a variety of effects on restudy behavior and final test performance by obtaining confidence ratings after each quiz response. Before continuing, however, we should highlight our departure from using a common measure of metacognitive resolution (Dunlosky & Metcalfe, 2008). Resolution is the within-person correlation between confidence and accuracy (or restudy), and is typically calculated using the Goodman-Kruskal gamma coefficient for each participant. Gamma assesses the correlation between an ordinal factor and a nominal outcome. For example, a positive gamma coefficient (range, -1:1) between confidence ratings (1-10) and quiz accuracy (0, 1) would indicate that a participant was more likely to be correct (1) at higher levels of the confidence scale (e.g., 8, 9, 10), more likely to be incorrect (0) at lower levels of the confidence scale (1, 2, 3), and more likely to display variability on quiz performance in the middle of the scale (4, 5, 6, 7). Thus, a positive gamma

would indicate that a participant was accurately able to predict their performance. However, the gamma correlation runs into problems when participants do not use the full range of the scale.

Take, as an extreme example, a participant who gave half of their quiz answers a confidence rating of 9, and the other half a confidence rating of 10. Because of the reduction in scale, a gamma correlation would only be perfect (1) in this scenario if every answer given a 9 was incorrect, and every answer given a 10 was correct. Therefore, if our theoretical participant was perfectly accurate with confidence ratings of only 9s and 10s, they would have a gamma of 0, which would indicate no resolution at all (though most people would agree that this participant had excellent resolution). Our data displayed extreme negative skew in confidence ratings (see Figure 9), presumably due to the high quiz performance. In fact, many participants had such good quiz performance and resolution that they did not use the full range of the scale. Several real participants, for example, only made confidence ratings between 6-10, and scored very highly on the quiz. Thus, their gamma correlations were actually *negative*. Our alternative approach, using multi-level logistic regression, accounts for the within-subject variance in confidence ratings (e.g., individuals may attribute different memory signals the same confidence rating) and does not encounter the same problems as gamma does with varying ordinal scales. Therefore, studies that have similar negative skew in confidence ratings, and participants who do not use the full range of the scale, may want to be wary when applying the gamma coefficient to determine metacognitive resolution.

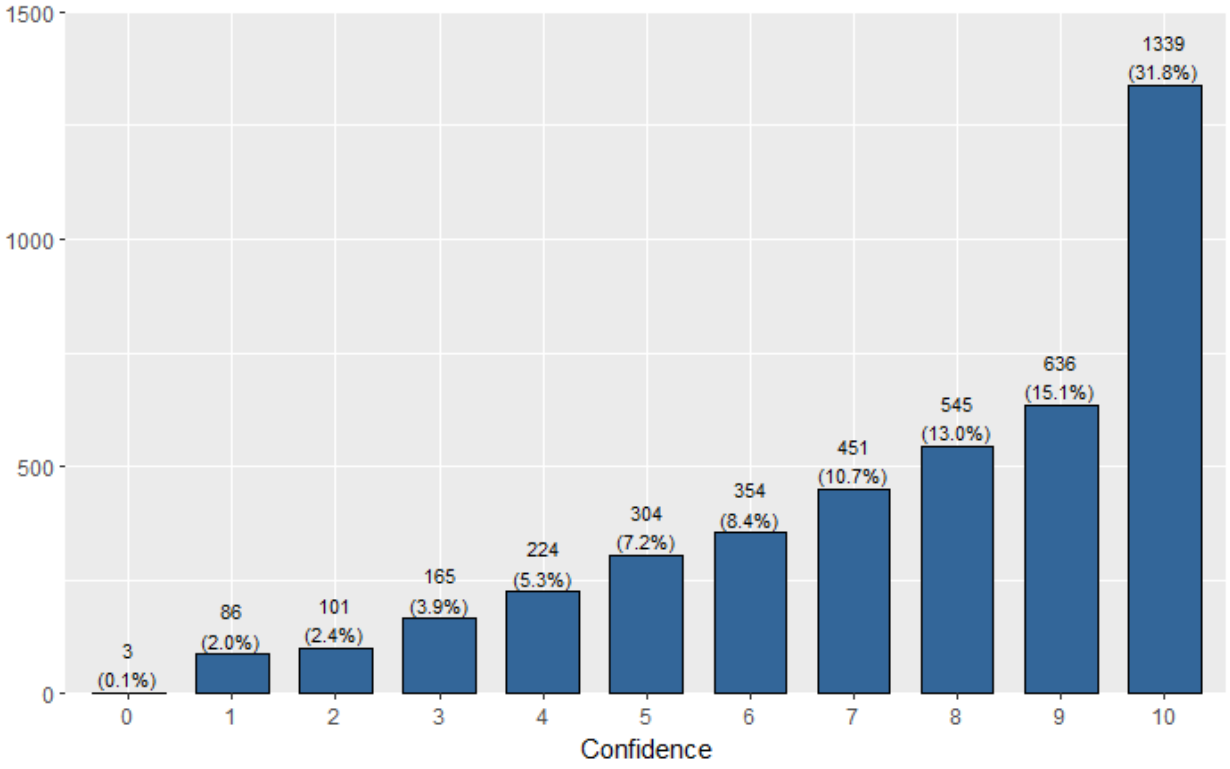

Figure 9. Histogram displaying the frequency with which each confidence rating was given (0-10) on the initial quiz.

The statistical approach favored in this study therefore gives an overall measure of metacognitive accuracy in terms of odds ratios, but simultaneously accounts for within-person resolution by letting intercepts vary between participants. Using this method, we found support for the discrepancy reduction hypothesis (Dunlosky & Hertzog, 1998; Kulhavy & Stock, 1989) in that participants were more likely to restudy items they answered incorrectly on the quiz. Additionally, we found that items with the highest discrepancy (high-confidence incorrect) were more likely to be restudied than those with the lowest discrepancy (high-confidence correct). This indicates that people selectively focused their restudy behavior on items they thought they knew, but objectively did not. Further, as we showed in model 2, this restudy behavior was beneficial for final test performance. Finally, it is important to note that discrepancy reduction

effects were not dependent on the presence of feedback, and were obtained in our minimal feedback condition as well.

Our results additionally broaden the influence of the discrepancy reduction framework by extending these findings to the use of naturalistic materials in authentic settings. The extensive review of study-time allocation by Son and Metcalfe (2000) shows that the vast majority of studies have been conducted using relatively impoverished materials, of which paired associates appear to be the most common. Our findings provide an important extension of the discrepancy reduction framework to additionally include complex conceptual information. Further, we found that following a discrepancy reduction approach boosted performance on the final test, even when that final test contained both definition and application short-answer questions.

We did not find support for the region of proximal learning hypothesis (Son & Metcalfe, 2000; Metcalfe, 2009), however, despite the fact that restudy time was experimenter-imposed (15 min) and presentation of the material was simultaneous (because it was a scrollable PDF of the book excerpts). Metcalfe (2009) has argued that in situations in which students are under time pressure (e.g., the night before an exam) and when you have the opportunity to simultaneously compare which items you could choose to restudy, students are especially likely to study in their region of proximal learning. Although our study contained these features, participants did not study in their own proximal region, but instead focused most of their restudying on difficult items. Much of the research showing proximal learning effects uses less naturalistic materials (e.g., cued-recall of foreign language vocabulary pairs), and this could be a possible reason for the departure of our results from prior studies. Additionally, it could be the case that participants judged 15 min to be plenty of time to study the more difficult concepts (i.e., they did not feel as though they were pressured for time).

Moving on to effects of quiz accuracy and confidence on final test performance, we examined whether participants hypercorrected their high-confidence errors and whether their low-confidence successes persevered to the final test (Butterfield & Metcalfe, 2001; Butler et al., 2008). Both of these metacognitive effects have been shown to contribute to the benefits of feedback on final test performance. We found significant hypercorrection effects in that participants were more likely to answer final test questions correctly for high-confidence errors than for low-confidence errors. We did not, however, replicate the Butler et al. (2008) findings: Low-confidence successes did not translate into higher performance than low-confidence errors. Just as in the case of our inability to find an effect of feedback type, it is possible that our use of authentic materials, combined with the generally high quiz performance, resulted in our inability to replicate the Butler et al. findings. However, in authentic settings, it may be that feedback simply benefits high-confidence errors more so than low-confidence successes (but also note that feedback was not necessary for hypercorrection effects, as they were still obtained in our minimal feedback condition).

Although we cannot tease apart the differential effects of quiz performance and restudy behavior on JOLs because they were obtained following restudy, we were able to examine whether participants in a naturalistic study/test/restudy scenario made accurate predictions of final test performance. As seen in Figure 8, participants seemed to make fairly accurate predictions of final test performance based on their average JOL, though they were somewhat under-confident. These results suggest that some combination of quizzing, feedback, and restudy led students to be well-calibrated in their predictions. Many studies have found that immediate JOLs are far less accurate than delayed JOLs (a finding termed the delayed-JOL effect, Nelson & Dunlosky, 1991; Dunlosky & Nelson, 1992; 1994), which could also explain the high correlation

we obtained. For example, an additional possible explanation of our findings is that participants were basing predictions of their performance on only long-term memory, rather than now-unaccessible short-term memory (after a delay), reducing the likelihood of being over-confident. However, when we look at participants' calibration between immediate confidence ratings and their accuracy on the quiz (see Figure 10), participants still appeared to be under-confident, rather than over-confident.

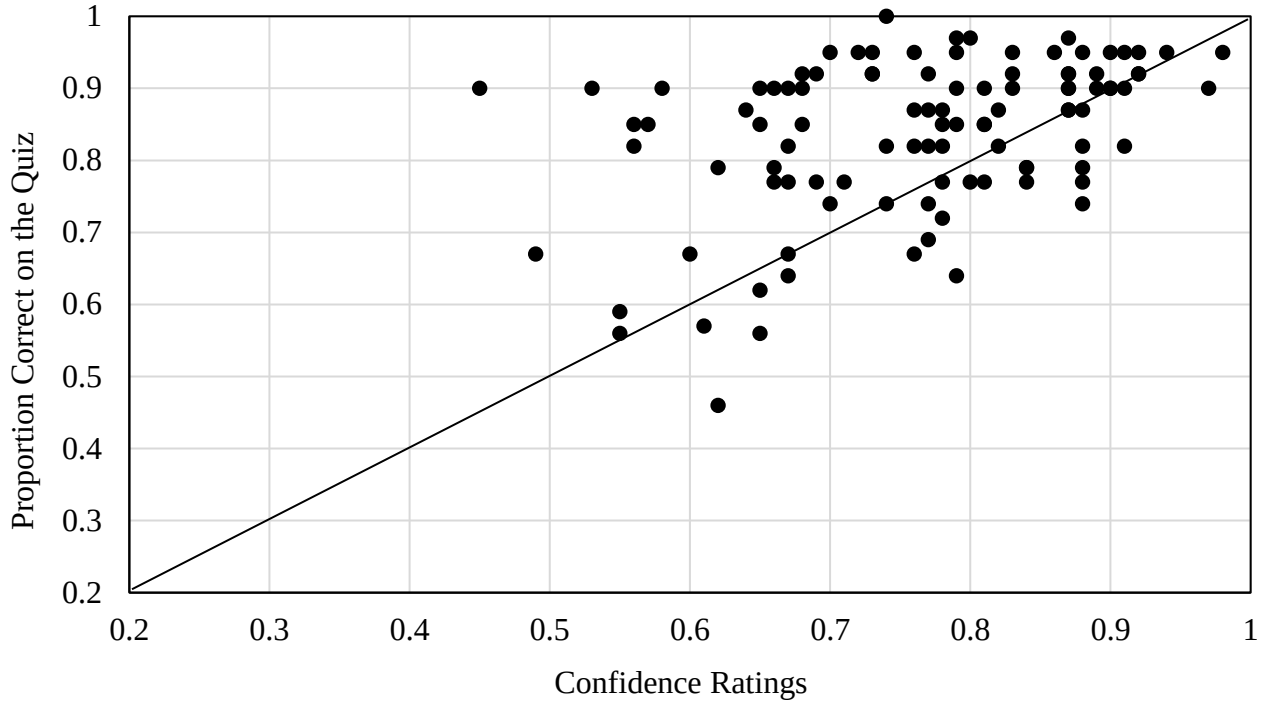

Figure 10. Calibration between participants' average confidence ratings and quiz performance. Confidence ratings were divided by 10 to maintain the same scale (e.g., a confidence rating of 5 became .50). The diagonal represents perfect calibration; therefore, data points above the diagonal represent under-confidence, and data points below the diagonal represent over-confidence. Therefore, participants appear to be under-confident.

In sum, our findings help further the understanding of students' metacognitive processes (especially in terms of study allocation policies) in a naturalistic educational setting. Although we believe the current methodology to be typical in terms of what is expected of most students, it should also be noted that there are significant departures from a completely authentic setting. To name a few, our participants do not have to juggle time pressures from other classes, we force participants to restudy even in situations that they may choose not to (e.g., with high quiz performance), and even when students do choose to restudy, their instructor may not provide them with a copy of their quiz. We believe that in most classrooms, however, students are asked to read a textbook or primary sources on unlearned information, come to class to hear a lecture on the topic, are perhaps quizzed as part of that class, and must go home to restudy before a final, summative test. Even if quizzes are not part of the classroom experience, most teachers do ask students to recall information from the beginning of the semester on cumulative final exams. Therefore, it is of critical importance to understand the metacognitive monitoring processes students use to base their restudy decisions on.

Our study highlights several important factors within this scenario; most notably that type of feedback does not appear to have much influence on students' restudy decisions or final test performance. However, we do find that students systematically choose to restudy questions they answered incorrectly, especially when they were highly confident in their responses. Further, this behavior appears to be adaptive, in that those items they chose to restudy were more likely to be answered correctly on the final test. Therefore, we end on the striking note that students seem to "know what they are doing"; and that as long as we structure our classes in such a way that there are plenty of opportunities for quizzing and restudy, they appear able to engage appropriate control processes and perform well in the class.

1.5 References

Andre, T., & Thieman, A. (1988). Level of adjunct question, type of feedback, and learning concepts by reading. *Contemporary Educational Psychology*, *13*, 296-307.

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

Bates, D., Machler, M., Ben, B., & Steve, W. (2015). Fitting linear mided-effects models using *lme4*. *Journal of Statistical Software, 67*, 1-48.

Birenbaum, M., & Tatsuoka, K. K. (1987). Effects of "on-line" test feedback on the seriousness of subsequent errors. *Journal of Educational Measurement*, *24*, 145-155.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (185-206). MIT press.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417-444.

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

Butler, A. C., Karpicke, J. D., & Roediger III, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 918.

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

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1491.

Dunlosky, J. & Ariel, R. (2011). Self-regulated learning and the allocation of study time. In B. H. Ross (Eds.), *The psychology of learning and motivation: Advances in research and theory* (103-136). Academic Press.

Dunlosky, J. & Hertzog, C. (1998). Training programs to improve learning in later adulthood: Helping older adults educate themselves. In D. J Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (249-276). Routledge.

Dunlosky, J., & Metcalfe, J. (2008). *Metacognition*. Sage Publications.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, *20*, 374-380.

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur?. *Journal of Memory and Language*, *33*, 545-565.

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language*, *52*, 551-565.

Dunlosky, J., & Thiede, K. W. (1998). What makes people study more? An evaluation of factors that affect self-paced study. *Acta Psychologica*, *98*, 37-56.

Fazio, L. K., Huelser, B. J., Johnson, A., & Marsh, E. J. (2010). Receiving right/wrong feedback: Consequences for learning. *Memory*, *18*, 335-350.

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

Hanna, G. S. (1976). Effects of total and partial feedback in multiple-choice testing upon learning. *The Journal of Educational Research, 69*, 202-205.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112.

Heiman, G. W. (2002). *Research methods in psychology (3rd ed.)*. Boston, MA: Houghton Mifflin.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 609.

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

Kulhavy, R. W., White, M. T., Topp, B. W., Chan, A. L., & Adams, J. (1985). Feedback complexity and corrective efficiency. *Contemporary Educational Psychology, 10*, 285-291.

Kulhavy, R. W., Yekovich, F. R., & Dyer, J. W. (1976). Feedback and response confidence. *Journal of Educational Psychology, 68*, 522.

Kulhavy, R. W., Yekovich, F. R., & Dyer, J. W. (1979). Feedback and content review in programmed instruction. *Contemporary Educational Psychology, 4*, 91-98.

Lee, O. M. (1985). The effect of type of feedback on rule learning in computer based instruction (Doctoral dissertation, Florida State University, 1985). *Dissertation Abstracts International, 46*, 955A.

Lüdecke, D. (2017). Data visualization for statistics in social science. Retrieved from https://CRAN.R-project.org/package=sjPlot.

Mazzoni, G., & Cornoldi, C. (1993). Strategies in study time allocation: Why is study time sometimes not effective?. *Journal of Experimental Psychology: General*, *122*, 47.

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation?. *Memory & Cognition*, *18*, 196-204.

McDaniel, M. A., Bugg, J. M., Liu, Y., & Brick, J. (2015). When does the test-study-test sequence optimize learning and retention?. *Journal of Experimental Psychology: Applied*, *21*, 370.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, *16*, 192-201.

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

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger III, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, *20*, 3.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science*, *18*, 159-163.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language*, *52*, 463-477.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The “delayed-JOL effect”. *Psychological Science, 2,* 267-271.

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science, 5,* 207-213.

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect". *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 676.

Nelson, T. O. & Narens, L. (1994). Why investigate metacognition. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (1-25). MIT press.

Noonan, J. V. (1984). *Feedback procedures in computer-assisted instruction: Knowledge-of-results, knowledge-of-correct-response, process explanations, and second attempts after errors* (Doctoral dissertation, University of Illinois at Urbana-Champaign).

Rawson, K. A., & Dunlosky, J. (2007). Improving students’ self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology, 19,* 559-579.

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

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

Roper, W. J. (1977). Feedback in computer assisted instruction. *Programmed Learning and Educational Technology, 14,* 43-49.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*, 1432-1463.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words?. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 3.

Pyc, M. A., & Rawson, K. A. (2011). Costs and benefits of dropout schedules of test–restudy practice: Implications for student learning. *Applied Cognitive Psychology, 25*, 87-95.

Sassenrath, J. M., & Garverick, C. M. (1965). Effects of differential feedback from examinations on retention and transfer. *Journal of Educational Psychology, 56*, 259.

Son, L. K. & Kornell, N. (2008). Research on the allocation of study time: Key studies from 1890 to the present (and beyond). In Dunlosky, J. & Bjork, R. A. (Eds.), *Handbook of metamemory and memory* (333-351). Psychology Press.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning Memory and Cognition, 26*, 204-221.

Thiede, K. W. (1999). The importance of monitoring and self-regulation during multitrial learning. *Psychonomic Bulletin & Review, 6*, 662-667.

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1024.

Travers, R. M. W., Van Wagenen, K. R., & McCormick, M. (1964). Learning as a consequence of the learner's task involvement under different conditions of feedback. *Journal of Educational Psychology, 55*, 167-173.

Wentling, T. L. (1973). Mastery versus nonmastery instruction with varying test item feedback treatments. *Journal of Educational Psychology, 65*, 50.

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition, 3*, 214-221.

Zacks, R. T. (1969). Invariance of total learning time under different conditions of practice. *Journal of Experimental Psychology*, *82*, 441.

1.6 Appendices

1.6.1 Instructions

Participants were initially told they would be reading a chapter on research methods in psychology, would take a quiz on this information, and have an opportunity to restudy before returning 2 days later to take the final test. They were instructed to read through the chapter at first, rather than study as they went, because it was long, and we wanted them to get through the whole thing. They were also told that it was alright if they did not finish, and that if they did they could study sections as they wished.

For the quiz, participants were told there were 40 multiple-choice options, and that they would have as much time as they needed, though it takes most people 15-20 min. Additionally, they were asked to indicate their confidence after each answer, on a scale of 1-10. We also encouraged them to use the full range of the scale, rather than simply putting 1s and 10s. After completion, they were told they would be playing Tetris for 5 min.

During restudy, participants were told they would have 15 min, and were asked to make a mark next to each paragraph they restudied. Rather than highlighting each line of the document, they were shown how to just make a small mark with the highlight tool next to whole paragraphs. They were also told that they could use their quiz during restudy. That is, they were told if they wanted to they could look at the quiz to help guide restudy.

For the final test, participants were simply told to answer to the best of their ability, but that they did not need to write an essay for each short answer question. They had unlimited time, but were informed that most people took approximately 30-45 min.