Optimizing the Use of Interpolated Tests in Recorded Lectures: The Influence of Interpolated Test Lag

Abstract

The use of recorded lectures is increasing rapidly provided growth in online learning. One technique that can be used to improve learning from recorded lectures is interpolated testing – the presentation of tests throughout the recorded lecture. In the present investigation, we examine a critical question with respect to the implementation of interpolated testing in recorded lectures. Where should the tests be located relative to the tested material? Specifically, we examine the influence of the lag between the presentation of the to-be-remembered material and the interpolated test. Across two experiments we compare an immediate test condition (i.e., a test immediately after the relevant information is presented) and a delayed test condition (i.e., a test is presented about 3 minutes after the relevant information is presented). When no feedback was provided immediate interpolated testing was superior to delayed interpolated testing. There was no difference when feedback was provided. Implications of the present results for implementing interpolated testing in educational contexts are discussed.

Optimizing the Use of Interpolated Tests in Recorded Lectures: The Influence of Interpolated Test Lag

There has been massive growth in online learning in postsecondary education (e.g., Allen & Seaman, 2014). One of the primary pedagogical devices in the online learning domain is a recorded lecture. These lectures present material typically created by an instructor and instructional designer and made available via the Internet. Critically, recorded lectures provide opportunities for implementing teaching strategies that could enhance learning. One such strategy is the use of interpolated tests – tests that are inserted throughout the recorded lecture (Szpunar, Khan, & Schacter, 2013; Szpunar, Jing, & Schacter, 2014). Much research has demonstrated that testing in general can be used to enhance retention (e.g., Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Roediger & Butler, 2011; Roediger & Karpicke, 2006) and Szpunar and colleagues demonstrated this specifically in the context of interpolated tests. However, interpolated tests could be implemented in a number of different ways within a recorded lecture thus raising questions regarding how best to use interpolated tests. We turn to this question in the present investigation. In particular we examine one of the first implementational questions one might encounter: Where in the lecture should the interpolated tests be located with respect to the to-be-tested material?

In the first examination of interpolated testing in the context of lectures, Szpunar et al. (2013) conducted two experiments that involved participants learning statistics from a video lecture (viewed in a laboratory). The lecture was divided into four segments, and participants were either given a test at the end of each segment (interpolated test condition), or an arbitrary arithmetic problem to solve (non-test condition). In the second experiment, they also added a restudy condition. They found that interpolated testing resulted in better performance on the final test than those in the non-testing and restudy conditions. They also found that interpolated testing reduced

mind wandering and increased note taking frequency. Szpunar, et al. (2014), again using a recorded lecture (viewed in a laboratory), replicated and extended this work by demonstrating that interpolated testing helped reduce overconfidence. In both of these studies multiple questions were presented following a lecture segment. As such, the lag between information presentation and the testing of that information was variable. This need not be the case. For example, a test question could be presented immediately after the information is presented or systematically delayed by some amount of time. While no research has directly assessed which method is preferable in terms of retention, two influential ideas lead to two different predictions and relevant research has yielded an unclear picture.

There appears to be a general consensus around the idea that delayed testing benefits retention (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Cepeda et al., 2009; Rawson, Vaughn, & Carpenter, 2015). Delayed testing causes an increase in the difficulty of the test, which in turn results in better long-term retention (i.e., a desirable difficulty). In an extension of the Elaborative Retrieval Hypothesis (Carpenter, 2009; 2011), a popular account of the testing effect, Rawson et al. (2015) suggested that longer lags provide a greater opportunity for elaborative retrieval (i.e., activation of related information) than shorter lags. From this perspective, delayed interpolated tests within a lecture could be expected to be superior to immediate interpolated tests. Alternatively, research on testing has also revealed that retrieval success represents an important factor in successful retention (e.g., Bjork, 1999; Pyc & Rawson, 2009) and immediate testing is likely to lead to greater retrieval success. The influence of retrieval success can also be interpreted within the elaborative retrieval framework where low retrieval success contexts would fail to generate elaborative retrieval (Rawson et al., 2015). From this perspective, immediate interpolated tests within a lecture would be expected to be superior to delayed interpolated tests.

In research examining expanding retrieval practice, Karpicke and Roediger (2007) asked participants to learn vocabulary word pairs and each pair was presented according to one of several different schedules. Two of these schedules involved single tests (one immediate and one delayed), which to some extent mimics the contrast between immediate and delayed interpolated testing, and in almost all cases a delayed test led to superior retention of information from the lecture. Karpicke and Roediger (2010) examined a similar question using memory for expository texts. In their second experiment, they had participants study nine passages, using various testing schedules and manipulating the presence of feedback. Unlike Karpicke and Roediger (2007), in the single test conditions (one immediate test or one delayed test), overall there was no difference between the immediate and delayed single test condition. However, there was an interaction with lag that appeared to reflect a small benefit for immediate testing (.37 vs. .32) with no feedback and a larger benefit for delayed testing (.46 vs. .34) when feedback was provided. Recent work has also compared retention as a function of whether testing occurred interpolated throughout a study session (i.e., akin to immediate testing) or delayed until the end of the study session (Healy, Jones, Lalchandani, & Tack, 2017; Weinstein, Nunes & Karpicke, 2016; Wissman & Rawson, 2015). For example, Weinstein et al. (2016) compared testing administered after each relevant fact in a slideshow to testing administered at the end of the slideshow. Healy et al. (2017) and Wissman and Rawson (2015) followed similar procedures. Across these studies, performance on final retention tests either showed no difference between conditions or a benefit for interpolated testing over end-of-session testing. Thus, taken together, research on expanding retrieval practice (Karpicke & Roediger, 2007; 2010) and interpolated versus end-of-session testing (Healy et al., in press; Weinstein et al., 2016; Wissman & Rawson, 2015), which are arguably similar to the contrast between immediate and delayed interpolated testing, have yielded a somewhat unclear

picture with respect to whether immediate versus delayed interpolated tests would yield any differential benefits and in which direction.

Present Investigation

None of the research reviewed above compared immediate versus delayed interpolated testing directly. In the experiments reported here, participants were asked to study a recorded lecture online and were tested on the material afterward. In Experiment 1 there were three between-subject conditions represented by the presence or absence of interpolated tests and the delay between the presentation of information in the lecture and the interpolated test on that information. The no-interpolated testing condition was included as a control to assess the benefits of interpolated testing. Based on previous research, interpolated testing should lead to improved performance on the final test. In the interpolated testing conditions, participants were presented with short answer questions either immediately after the information was presented in the lecture or following a 3-minute delay. Performance on the interpolated tests is expected to be greater in the immediate condition than in the delayed condition. The critical question is whether there is also a difference in terms of final test performance. In addition, Experiment 1 provided participants with no feedback whereas in Experiment 2 feedback was presented. Providing feedback is typical in classroom settings and, as it provides a re-exposure to the content, provided us with the opportunity to investigate whether re-exposure would modulate the influence of interpolated test delay. For example, immediate testing is likely to lead to a higher rate of successful recall during the lecture and thus a higher likelihood of re-exposure. Providing feedback could reduce any differential benefit of re-exposure (i.e., both conditions are re-exposed to the answer after the question). Thus, comparing feedback and no feedback conditions provide both practical value (i.e., how do the benefits/costs of immediate versus delayed interpolated tests vary as a function of

pedagogical strategy) and provides some insight into the potential mechanism underlying any differences between immediate and delayed testing.

Experiment 1

Methods

Subjects. We set out to collect 150 participants in the interpolated testing conditions in order to have sufficient power to detect a medium-sized effect (*d = 0.5, power = 0.8*). In Experiment 1 two-hundred-fifty (109 immediate; 108 delayed; 33 no testing control) individuals participated in exchange for $10. All participants took part online through Amazon Mechanical Turk. Participants varied in age, gender, ethnicity, and educational background.

Design. A 3 (Test Type: no interpolated test vs. immediate interpolated test vs. delayed interpolated test) between subject design was used.

Stimuli. The recorded lecture with interpolated testing was created and administered using a software tool developed for this purpose. The video lecture that was used was the second lecture of an Introduction to Biology course taught from Massachusetts Institute of Technology's MIT Open Courseware website (http://ocw.mit.edu/courses/biology/7-012-introduction-to-biology-fall-2004/video-lectures/lecture-2-biochemistry-1/). The topic of this particular lecture is Biochemistry. The first 23 minutes was shown to participants.

A total of 16 short-answer questions were created based on the material in the recorded lecture for use as interpolated tests. Presentation of the question involved pausing the lecture and the question appearing overlaid on the lecture video. Participants could not interact or see the video until an answer was submitted. Video scrubbing was also disabled so participants could not re-watch or skip parts of the lecture. The final retention test consisted of the same sixteen questions

as those used for the interpolated tests but were presented in a multiple-choice format instead of the original short answer format. In order to establish a baseline of performance on this final test, we had thirty-four[1] participants from the same participant population complete the final retention test with no lecture or interpolated questions. The mean test score was 41.5% CI [35.5%, 47.6%].

Procedure. The experiment was conducted online using Amazon's Mechanical Turk system. After accepting our posted task, participants were given a brief overview of the study, as well as a link to a website. When they reached this website, participants were then given the detailed instructions on how to complete the recorded lecture portion of the study. They were told that questions may come up throughout the lecture and that they needed to answer these questions to the best of their ability. Participants were then instructed that they were to watch the lecture. Before doing so, they were asked to complete a demographic questionnaire. The demographic questionnaire asked participants to fill out information regarding their familiarity with the general subject of biochemistry and how many school courses they have taken in both chemistry and biology. Participants then watched the lecture and, depending on condition, either answered interpolated questions throughout (immediate and delayed interpolated testing) or not (control). In the immediate condition, the question appeared directly after the information was presented in the lecture. In the delayed condition, the question appeared approximately three minutes after the information was presented. For those in the control condition, the video played through without any pausing. After the lecture was finished, participants were taken to the final retention test. There

[1]A total of 40 participants were collected in this no lecture baseline data, but 6 had to be excluded from the analysis (because of failure to complete the final test, or duplicate submission).

was no time limit on the interpolated and final test questions. After completing the final test, participants were debriefed.

Results

In Experiment 1 fifty-seven participants were excluded due to technical reasons (e.g., duplicate participation). After these exclusions, there were 88 participants in the immediate test type condition, 74 in the delayed test type condition, and 30 in the control condition. Interpolated tests were graded by two markers, one graded all the answers and the other graded 25% of the answers. The raw percentage agreement (on answers marked by both coders) in their marking was 93.3%, with a Kappa of 0.85. When a Levene's test revealed a violation of the homogeneity of variance, a correction was applied and did not alter the results qualitatively. The original results are reported. There was no difference across conditions in terms of familiarity with the general subject of biochemistry (8 participants did not complete this information), $F(2,181) = 1.0$, $p = .35$, $\eta^2_G = 0.01$, or how many school courses they had taken in chemistry, $F(2,189) = 0.56$, $p = .57$, $\eta^2_G = 0.01$, or biology, $F(2,189) = 1.38$, $p = .25$, $\eta^2_G = 0.01$. Data are presented in Figure 1.

Interpolated Testing Scores. A 2 (immediate vs. delayed) one-way between subject ANOVA was conducted on interpolated test scores. There was a significant effect of interpolated test delay, $F(1,160) = 185.43$, $p < 0.001$, $\eta^2_G = 0.54$, $d = 2.14$, such that participants in the immediate condition (M = 63.9%, CI [61.3%, 66.4%]) scored significantly higher than those in the delayed condition (M = 34.0%, CI [30.3%, 37.7%]).

Final Retention Scores. A 3 (test type: immediate vs. delayed vs. control) one-way between subject ANOVA on final retention test scores was conducted. There was a significant effect of test type, $F(2,189) = 13.62$, $p < 0.001$, $\eta^2_G = 0.13$, such that participants in the immediate interpolated

test condition (M = 75.0%, CI [71.6%, 78.4%]) performed better on the final test than those in the delayed interpolated test condition (M = 67.8%, CI [63.8%, 71.8%]), $t(160) = 2.74$, $p = 0.007$, d = 0.43. Participants in the control condition (M = 56.7%, CI [49.7%, 63.7%]) performed worse than individuals in the immediate test condition, $t(116) = 5.21$, $p < 0.001$, $d = 1.06$, and delayed test condition, $t(102) = 2.90$, $p = 0.005$, $d = 0.62$.

Question Response Time. While not the focus of the present research, we also collected response times to each interpolated and final test question. We report an exploratory analysis here. With respect to the interpolated test questions, twelve participants were removed because the response time data was not collected. In addition, of the remaining participants, 1.1% of observations were removed for the same reason. An outlier analysis was conducted that excluded responses times greater than 2.5 standard deviations from the participant's mean in the interpolated condition resulting in the removal of 2.8% of the data. A 2 (immediate vs. delayed) one-way between subject ANOVA was conducted on interpolated test response times. There was a significant effect of interpolated test delay, $F(1,145) = 24.19$, $p < 0.001$, $\eta^2_G = 0.14$, $d = 0.88$, such that participants in the delayed condition (M = 53103 ms, CI [44103, 62103]) responded slower than those in the immediate condition (M = 29182 ms, CI [25452, 32912]). With respect to the final test questions, twelve participants were removed because the response time data was not collected. The same outlier analysis was conducted and resulted in the removal of 3.9% of the data. A 2 (immediate vs. delayed) one-way between subject ANOVA was conducted on final test response times. There was no significant effect of interpolated test delay, $F(1,145) = 0.5$, $p = 0.49$, $\eta^2_G = 0.0$, $d = .11$. Participants in the delayed condition (M = 12349 ms, CI [11224, 13474]) and the immediate condition (M = 13055 ms, CI [11327, 14783]) took approximately the same amount of time to respond to questions.

Discussion

The results of Experiment 1 demonstrate that in the context of a recorded lecture with no feedback immediate interpolated tests lead to greater gains in retention than a delayed (3 minute) interpolated test. Both forms of interpolated tests outperformed a no-test control demonstrating again that interpolated testing in the context of recorded lectures can benefit retention. A detailed discussion will follow Experiment 2.

In Experiment 2, we wanted to test one potential explanation of the benefit of immediate over delayed interpolated testing on the final test. Specifically, in the immediate condition, the associated benefit on the final test could reflect the higher likelihood of being re-exposed to the answer to the questions during the interpolated testing part of the experiment provided the large difference in successful retrieval across conditions. To test this account, we replicated the immediate and delayed interpolated conditions from Experiment 1 but added feedback. If the benefit of immediate interpolated testing is a product of the increased likelihood of re-exposure, then the advantage of immediate over delayed interpolated testing should be reduced/eliminated when feedback is provided. In this case, feedback can be conceptualized as reducing the putative difference between the immediate and delayed conditions in terms of re-exposure afforded by the difference in successful retrieval across conditions (i.e., all participants are re-exposed to the answers after each interpolated test). From a practical perspective, different instructional contexts may make the provision of feedback more or less desirable and as such understanding how implementational decisions might be impacted this variable is valuable.

Experiment 2

Methods

Subjects. In Experiment 2, one-hundred-and-ninety-three Amazon Mechanical Turk workers took part. Participants were rewarded with $10.

Design, *Stimuli & Procedure.* The Design, Stimuli & Procedure were the same as in Experiment 1 except for the provision of feedback in the feedback condition. Feedback consisted of the correct answer being provided to each interpolated testing question after they submitted their answer. Provided the questions were short answer, feedback consisted of a one word or short sentence providing the correct answer. As the questions were largely fact-based, no further feedback was provided (i.e., why this is the correct answer). In addition, feedback always occurred immediately after the participant entered their response to the interpolated question.

Results

We did not remove any participants in Experiment 2. The raw percentage agreement in the two markers' grading of the short answer questions for all participants was 94% with a Kappa of 0.87. There was no difference across conditions in terms of familiarity with the general subject of biochemistry, $F(1,191) = 0.04$, $p = .84$, $\eta^2_G = 0.00$, or how many school courses they had taken in chemistry, $F(1,191) = 0.69$, $p = .41$, $\eta^2_G = 0.00$, or biology (1 participant did not complete this information), $F(1,190) = 0.76$, $p = .39$, $\eta^2_G = 0.00$. Data are presented in Figure 2.

Interpolated Testing Scores. A 2 (immediate vs. delayed) one-way between subject ANOVA was conducted on interpolated test scores. There was a significant effect of interpolated test delay, $F(1,191) = 170.1$, $p < 0.001$, $\eta^2_G = 0.47$, $d = 1.89$, such that participants in the immediate condition (M = 58.3%, CI [54.6%, 62.0%]) scored significantly higher than those in the delayed condition (M = 27.1%, CI [24.1%, 30.1%]). We also conducted a 2 (Experiment: 1 vs. 2) x 2 (test type: immediate vs. delayed) 2-way between subjects ANOVA on interpolated test scores. There

was a significant effect of interpolated test type, $F(1,351) = 342.75$, $p < .001$, $\eta^2_G = 0.49$, such that participants performed better in the immediate condition and a significant effect of Experiment, $F(1,351) = 14.35$, $p < .001$, $\eta^2_G = 0.04$, such that interpolated test scores were lower in Experiment 2 (with feedback; M = 42.7%) than in Experiment 1 (no feedback; M = 48.9%). The interaction between test type and experiment was not significant, $F(1,351) = .19$, $p = 0.67$, $\eta^2_G < 0.01$.

Final Retention Scores. A 2 (test type: immediate vs. delayed) one-way between subject ANOVA on final retention test scores was conducted. There was no significant effect of test type, $F(1,191) = .94$, $p = .33$, $\eta^2_G = 0.01$, $d = .14$. Participants' performance in the immediate interpolated test condition (M = 83.4%, CI [80.2%, 86.6%]) and the delayed interpolated test condition (M = 81.2%, CI [78.1%, 84.3%]) were similar. We also conducted a 2 (Experiment: 1 vs. 2) x 2 (test type: immediate vs. delayed) two-way between subjects ANOVA on final test scores. There was a significant effect of interpolated test type, $F(1,351) = 7.43$, $p = 0.007$, $\eta^2_G = 0.02$, such that participants performed better in the immediate condition and a significant effect of Experiment, $F(1,351) = 39.97$, $p < .001$, $\eta^2_G = 0.10$, such that final test scores were higher in Experiment 2 (with feedback; M = 82.3%) than in Experiment 1 (no feedback; M = 71.4). The interaction between test type and experiment was not significant, $F(1,351) = 2.10$, $p = 0.148$, $\eta^2_G = 0.01$.

Question Response Time. As in Experiment 1, we also collected response times to each interpolated and final test question. We report an exploratory analysis here. No participants were removed. In the interpolated test, an outlier analysis was conducted that excluded response times greater than 2.5 standard deviations from the participant's mean resulting in the removal of 2.2% of the data. A 2 (immediate vs. delayed) one-way between subject ANOVA was conducted on interpolated test response times. There was a significant effect of interpolated test delay, $F(1,191)$

= 6.8, p = .01, η^2_G = .03, d = .38, such that participants in the delayed condition (M = 41350 ms, CI [36702, 45999]) responded slower than those in the immediate condition (M = 33333 ms, CI [29384, 37282]). With respect to the final test questions, the outlier analysis resulted in the removal of 3.5% of the data. A 2 (immediate vs. delayed) one-way between subject ANOVA was conducted on final test response times. There was a marginally significant effect of interpolated test delay, $F(1,191) = 4.03$, p = .05, η^2_G = .02, d = .31, such that participants in the delayed condition (M = 9875 ms, CI [8977, 10773]) responded faster than participants in the immediate condition (M = 12187 ms, CI [10095, 14279]).

Discussion

Experiment 2 replicated the large effect of test type on performance on the interpolated tests. Participants performed much better when probed immediately after information was presented than approximately three minutes after. In addition, there was a clear positive effect of feedback on final test scores. Unexpectedly, there was also a negative effect of feedback on interpolated tests scores (we return to this effect in the General Discussion). Critically, unlike Experiment 1 there was no significant difference between immediate and delayed interpolated test conditions on final test scores. Thus, when feedback is provided, the benefit of immediate over delayed interpolated testing was eliminated. This is consistent with the benefit of immediate interpolated testing reflecting at least in part an increase in the likelihood of successful retrieval and the concomitant re-exposure to the answer. Importantly, the provision of feedback did not lead to a benefit for delayed over immediate interpolated testing. In addition, it is important to note the lack of an interaction between test type and experiment revealed in the combined analysis. This likely reflects a lack of power provided the effect in Experiment 1 (d = .43) decreased from a small to medium sized effect to a small effect in Experiment 2 (d = .14).

General Discussion

Implementing interpolated testing in recorded lectures raises a number of practical questions. In the present experiments, we examined the influence of test lag in the context of interpolated testing in recorded lectures. Specifically, we compared the relative effectiveness of immediate versus delayed interpolated tests both in a context with feedback provided and feedback not provided. Experiment 1 revealed a benefit of immediate and delayed interpolated testing over a no interpolated testing control, thus replicating the beneficial effect of interpolated testing (Szpunar et al., 2013). Unsurprisingly, performance on the interpolated tests was superior when those tests were immediate than delayed. On the final test performance was greater in the immediate interpolated condition than in the delayed interpolated testing condition when no feedback was provided (Experiment 1) but performance did not differ significantly across the two conditions when feedback was provided (Experiment 2). In neither experiment was there any evidence that delayed interpolated testing was superior to immediate interpolated testing. As reviewed in the introduction, recent research relevant to the present work was mixed with respect to whether immediate or delayed interpolated testing would be superior. The present research provides some clarity, at least within the contexts evaluated. Namely, immediate interpolated testing appears to be superior or at least equivalent to delayed interpolated testing and both are superior to a no interpolated testing condition.

One critical question that emerges from this work is why delayed testing is not superior to immediate testing (Karpicke & Roediger, 2007). One salient difference between conditions is the large difference in retrieval success across the immediate and delayed interpolated test conditions (in favor of the immediate condition). This can have at least two consequences. First, the higher likelihood of retrieval success in the immediate condition increases the likelihood of re-exposure

in that condition (relative to the delayed condition). Re-exposure can be beneficial (e.g., Rawson, Dunlosky, & Thiede, 2000; Roediger & Karpicke, 2006; though not always, see Callender & McDaniel, 2009; Phillips, Mills, D'Mello & Risko, 2016; Martin, Mills, D'Mello & Risko, in press). Thus, in Experiment 1 participants were more likely to be re-exposed to the material than individuals in the delayed condition. That this factor (i.e., differential rates of re-exposure) might be operative in the benefit of immediate interpolated testing, relative to delayed interpolated testing, draws support from Experiment 2. In Experiment 2 participants were all provided feedback and, thus, were all re-exposed to the material following each interpolated test. In this sense, feedback could be construed as leveling the playing field in terms of re-exposure across the immediate and delayed interpolated testing conditions. If the higher likelihood of re-exposure to the material in the immediate condition (relative to the delayed condition) during the interpolated tests was partly responsible for the benefit on the final test, then equating re-exposure via feedback should have reduced/eliminated the benefit of immediate interpolated testing. This is what we observed in Experiment 2.

One critical question that emerges from this work is why delayed testing is not superior to immediate testing (Karpicke & Roediger, 2007). One salient difference between conditions is the large difference in retrieval success across the immediate and delayed interpolated test conditions (in favor of the immediate condition). This can have at least two consequences. First, the higher likelihood of retrieval success in the immediate condition increases the likelihood of re-exposure in that condition (relative to the delayed condition). Re-exposure can be beneficial (e.g., Rawson, Dunlosky, & Thiede, 2000; Roediger & Karpicke, 2006; though not always, see Callender & McDaniel, 2009; Phillips, Mills, D'Mello & Risko, 2016; Martin, Mills, D'Mello & Risko, in press). Thus, in Experiment 1 participants were more likely to be re-exposed to the material than

individuals in the delayed condition. That this factor (i.e., differential rates of re-exposure) might be operative in the benefit of immediate interpolated testing, relative to delayed interpolated testing, draws support from Experiment 2. In Experiment 2 participants were all provided feedback and, thus, were all re-exposed to the material following each interpolated test. In this sense, feedback could be construed as leveling the playing field in terms of re-exposure across the immediate and delayed interpolated testing conditions. If the higher likelihood of re-exposure to the material in the immediate condition (relative to the delayed condition) during the interpolated tests was partly responsible for the benefit on the final test, then equating re-exposure via feedback should have reduced/eliminated the benefit of immediate interpolated testing. This is what we observed in Experiment 2.

While the differential likelihood of re-exposure could explain the lack of a benefit for the delayed interpolated testing condition relative to the immediate interpolated testing condition in Experiment 1, this seems a somewhat unsatisfactory account of why there was no benefit for delayed testing in Experiment 2 (where re-exposure was more similar). As noted above, in Karpicke and Roediger (2010), provision of feedback led to a benefit for delayed testing over immediate. Such a reversal might have been expected if providing feedback controlled the likelihood of re-exposure across conditions leaving only the relative benefits of elaborative retrieval (i.e., activation of related information in the course of retrieving the information) which would presumably benefit the delayed condition. There are several potential explanations for the observation of no difference in the feedback condition, which are worth considering. The first focuses on a second consequence of the large difference in retrieval success. While feedback reduces the asymmetry in terms of re-exposure to the answer, there, nevertheless, still exists a large gap in successful retrieval prior to the provision of that feedback and successful retrieval yields

more than just re-exposure. That is, successful retrieval also leads to (or at least the potential for) elaboration and this elaboration can benefit memory (Carpenter, 2009; 2011). While successful delayed retrieval might involve more elaboration than successful immediate retrieval, the latter likely involves more elaboration than no retrieval (or incorrect retrieval). Thus, the immediate interpolated test condition can be seen as having a greater number of less elaborative retrieval experiences relative to the lower number of more elaborative retrieval experiences in the delayed retrieval condition. From this theoretical perspective, the higher rate of successful retrievals in the immediate interpolated test condition is cancelling out the benefit of greater elaboration in the delayed interpolated test condition. It is also important to note that while the testing that occurred in the immediate condition was much easier than delayed, it was still difficult (i.e., around 60-65% correct) relative to contexts that found a benefit of delayed over immediate testing (e.g., Karpicke & Roediger, 2007).

A second explanation emerges from the unexpected difference in interpolated test performance across Experiments 1 and 2. Participants performed better on the interpolated tests overall when they did not receive feedback on those tests. This difference could simply reflect a difference in samples as Experiments 1 and 2 were conducted at different times. However, this difference might reflect a genuine negative effect of feedback. Indeed, recent work in an anagram task has demonstrated that providing answers reduces persistence (i.e., time spent trying to solve the anagram; Risko, Huh, McLean & Ferguson, 2017). If participants put less effort into retrieval of answers to the interpolated tests because the answers are available whether retrieval is successful or not, this would likely reduce any potential benefits of elaborative rehearsal and reduce any effect across conditions. The exploratory analysis of question response time does not appear to bear this prediction out. A last potential explanation, and likely the least interesting, is that feedback simply

increases performance to such an extent that it reduces the opportunity for genuine differences across conditions to emerge. For example, in Experiment 2 (with feedback) 18% of participants scored 100% on the final test while only 2% of participants did so in Experiment 1 (no feedback). Future work exploring these different explanations for the lack of a difference between immediate and delayed interpolated testing when feedback is provided would be valuable.

Beyond Video Lectures

The focus of the present research has been on the burgeoning use of video lectures and, in particular, an attempt to help inform the use of interpolated testing in that context. That said, interpolated testing is certainly not restricted to recorded lectures. Interpolated questions are also used in live lectures using technologies like clickers and their close cousins (e.g., McDaniel, Agarwal, Huelser, McDermott & Roediger, 2011; Morling, McAuliffe, Cohen & DiLorenzo, 2008). Like interpolated testing in a recorded lecture, testing in this context appears to have benefits for retention of lecture material. Live lecture teachers are also presented with a choice as to when to present quiz questions. Assuming the results here generalize to this live lecture context, the present results could be used to inform these decisions for those using "clicker" technology to quiz in class in more traditional brick-and-mortar pedagogical environments.

Limitations

Research in psychology is often criticized for reliance on WEIRD samples (i.e., White, Educated, Industrialized, Rich, Democratic; Henrich, Heine, & Norenzayan, 2010) typically drawn from University courses. The sample used here was drawn from a different and more diverse sample (i.e., Mechanical Turk; Paolacci & Chandler, 2014). While this would usually be considered a strength, if one were concerned only with University samples it might reasonably be

considered a limitation. That said, online learning, where video-based lectures are often used, is a format that has the capacity to be accessed by a much broader population than traditional University samples. Indeed, this is one of the most attractive aspects of online education, namely, the asynchronicity affords different populations access to learning opportunities (e.g., Breslow, Pritchard, DeBoer, Stump, Ho, & Seaton, 2013) not typically afforded by traditional brick-and-mortar colleges and universities. Nevertheless, future research examining whether testing effects vary as a function of sample would be valuable as these methods are increasingly used in live classrooms. In a similar vein, future work focused on different lecture topics would also be of immense value.

Educational Implications

The present research provides important insights with respect to the use of interpolated testing as a pedagogical device. First, we have provided an independent replication of the beneficial effect of interpolated testing in the context of learning complex material. Beyond providing further evidence for the benefit of interpolated testing, the present investigation also provides novel insights into where to place those interpolated tests. Specifically, at least when there is no feedback to be provided, it seems best to place the interpolated tests near to the to-be-tested material. When feedback is provided the decision appears less important. This advice seemingly goes counter to the typical practice of only presenting tests at the end of the lecture and appears consistent with recent research on that general topic (Healy et al., in press; Weinstein et al., 2016; Wissman & Rawson, 2015). It is important to note, however, that much work stands to be done in determining "best practices" for the use of interpolated testing. As noted above, this will include examining potential moderation by population type or learner characteristics in general (e.g., knowledge level) and information type. In addition, there are valuable questions yet

to be understood regarding the potential benefit of interpolated testing for long term retention (e.g., months, years). Lastly, while the content of the lecture used here was not directly on the topic of psychology there is good reason to expect that the results would generalize. For example, the original work investigating interpolated testing used a statistics lecture (e.g., relation between sample and population; Szpunar et al., 2014; 2014) and here we used a lecture on biology, both topics not to distantly related to what one would expect in a psychology course (e.g., research methods, physiological psychology). In addition, the testing effect in general have been found across a wide swath of topics. All that said, replicating these results with a traditional psychology lecture would certainly have value in considering the utility of the present work to a particular topic.

Conclusion

In the present investigation, we demonstrated that interpolated testing in a recorded lecture provides significant benefits to information retention over no testing. We also showed that when no feedback is presented, using interpolated test questions that are presented immediately after information needed to answer the question is given yields further improvements to retention over the alternative of having a delay between this information presentation and interpolated testing. When interpolated test feedback is provided, we found no difference between immediate and delayed interpolated tests. These results have both theoretical and practical implications in that it furthers our understanding of testing effects and distributed practice as well as provides guidance for educators in their instructional planning.

References

Allen, I. E., & Seaman, J. (2014). Grade change: Tracking online education in the United States. *Babson Survey Research Group and Ouahog Research Group*.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher A. Koriat (Ed.), (pp. 435-459). Cambridge, MA, US: The MIT Press.

Breslow, L., Pritchard, D., DeBoer, J., Stump, G., Ho, A., & Seaton, D. (2013). Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research to Practice in Assessment, 8*, 13–25.

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.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1563-1569.

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1547-1552.

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology, 56*, 236-246.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*, 354-380.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*, 4-58.

Healy, A. F., Jones, M., Lalchandani, L., & Tack, L. A. (2017). Timing of Quizzes During Learning: Effects on Motivation and Retention. *Journal of Experimental Psychology. Applied*, *23*, 128-137.

Karpicke, J. D., & Roediger III, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 704-719.

Karpicke, J. D., & Roediger III, H. L. (2010). Is expanding retrieval a superior method for learning text materials? *Memory & Cognition, 38*, 116-124.

Martin, L., Mills, C., D'Mello, S., & Risko, E. F. (in press). Re-watching lectures as a study strategy and its effect on mind wandering. *Experimental Psychology*.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger III, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*, 399-414.

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

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. C*urrent Directions in Psychological Science, 23*, 184-188.

Phillips, N. E., Mills, C., D'Mello, S., & Risko, E. F. (2016). On the influence of re-reading on mind wandering. *The Quarterly Journal of Experimental Psychology*, e-pub Ahead of Print.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*, 437-447.

Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition, 28*, 1004–1010.

Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis. *Memory & Cognition*, *43*, 619-633.

Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis. *Memory & Cognition*, *43*, 619-633.

Risko, E. F., Huh, M., McLean, D., & Ferguson, A. (2017). On the prospect of knowing: Providing solutions can reduce persistence. *Journal of Experimental Psychology: General, 146,* 1677-1693.

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*, 20-27.

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

Szpunar, K. K., Jing, H. G., & Schacter, D. L. (2014). Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education. *Journal of Applied Research in Memory and Cognition, 3*, 161-164.

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences of the United States of America, 110*, 6313-6317.

Weinberg, R. A. (2004). Biochemistry 1. Retrieved from http://ocw.mit.edu/courses/biology/7-012-introduction-to-biology-fall-2004/video-lectures/lecture-2-biochemistry-1/

Weinstein, Y., Nunes, L., & Karpicke, J. D. (2016). On the placement of practice questions during study. *Journal of Experimental Psychology: Applied, 22*, 72-84.

Wissman, K. T., & Rawson, K. A. (2015). Grain size of recall practice for lengthy material: Fragile and mysterious effects on memory. *Journal of Experimental Psychology: Learning Memory and Cognition, 41*, 439-455.

Figure 1

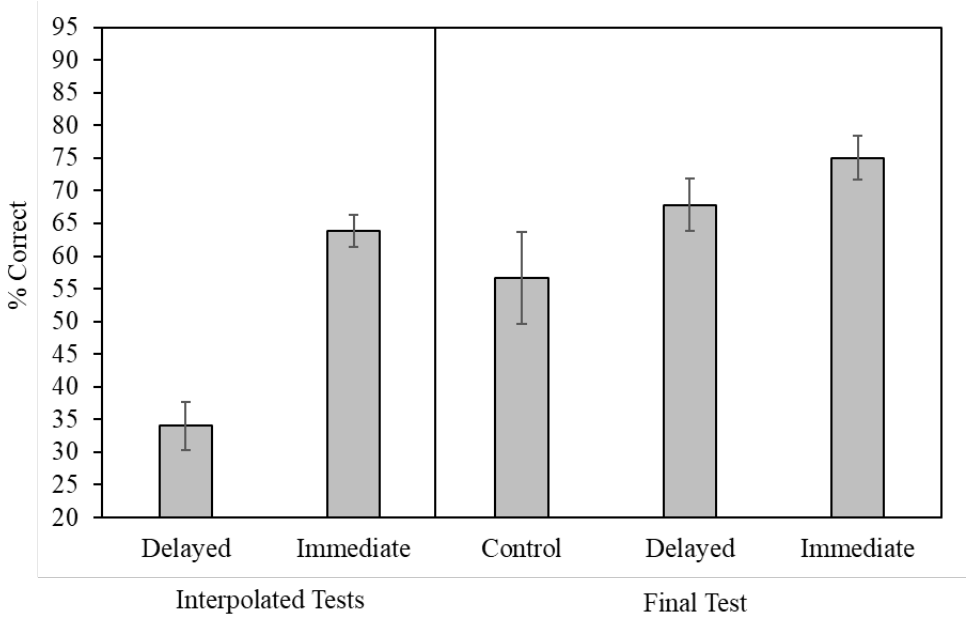
% Correct
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
Delayed
Immediate
Control
Delayed
Immediate
Interpolated Tests
Final Test

Figure 2

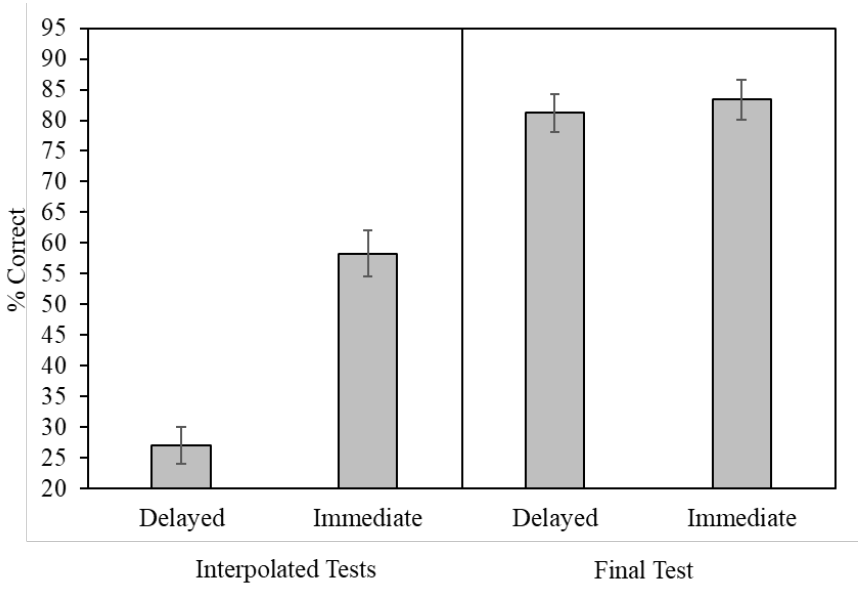
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
% Correct
Delayed
Immediate
Delayed
Immediate
Interpolated Tests
Final Test

Figure Captions

Figure 1. Performance on the interpolated tests (left panel) and final test (right panel) questions as a function of condition. Error bars represent 95% confidence intervals.

Figure 2. Performance on the interpolated test questions (left panel) and the final test (right panel) as a function of test type and feedback. Error bars indicate the 95% confidence intervals.