The Effect of Embedded Questions in Programming Education

Christopher Mar
School of Computing Informatics and Decision Systems Engineering
Ira. A. Fulton Schools of Engineering
Arizona State University
Mesa, Arizona

Sohum Sohoni
The Polytechnic School
Ira. A. Fulton Schools of Engineering
Arizona State University
Mesa, Arizona

Scotty D. Craig
The Polytechnic School
Ira. A. Fulton Schools of Engineering
Arizona State University
Mesa, Arizona

***Abstract*—This paper investigates the effectiveness of including questions within instructional multimedia content to improve student performance on a related programming assignment. An experiment was conducted where one set of students was provided with an instructional video without any embedded questions and another set of students was provided the same video with embedded questions. The findings of this paper demonstrate that the technique of embedding questions showed promise in improving student performance on a programming assignment.**

Keywords—Education, Programming, Instructional video, Electronic learning, Embedded questions.

I. Introduction

A. Motivation

Multimedia can be used in a variety of learning environments including flipped classrooms, for-credit online university courses and Massive Open Online Courses (MOOCs). In all three environments, multimedia provides a way to deliver educational content without requiring a lecturer or textbook. Popular MOOC providers, such as edX and Coursera, offer a majority of their content through videos, which include embedded questions [1].

This study will look at the effects embedded questions in a multimedia instructional video have on promoting knowledge transfer to a computer programming assignment. The primary motivation behind this study is to determine if adding interactive questions to instructional videos is a worthwhile endeavor for educators who produce their own content.

B. Background and Review of Literature

1) Multimedia Learning

While this paper will not directly compare different techniques used to produce effective multimedia content, these techniques form a foundation for the content used in the experiment. When this paper refers to multimedia it is referring to the type discussed by Mayer [2] in which multiple modes of communication are used. This paper focuses on the combination of auditory communication in the form of narration and visual communication using graphics and text.

Moreno and Mayer [3] highlight the importance of using multiple modes of communication that do not distract from one another. Part of their experiment compared three different ways of presenting information. Two of the groups were shown text and animation without narration. For one group, the text was physically separate from the animation while the other group saw text that was closely integrated with the animation. The third group was shown narrated animation without text. Three measures were used in the experiment. A verbal retention test was administered and scored based on the number of major ideas recalled. A transfer test was used that contained questions requiring participants to apply information presented in the multimedia content to related scenarios. The third measure was a matching test where participants matched labels with elements. In all three of these measures, the highest scoring group was the group given narrated animation. The next highest scores in all measures was the group with text integrated into the animation. The group shown text separated from the animation scored lowest on all three measures.

Recorded multimedia offers a flexible middle ground between books (or text on a webpage) and a real-time instructor lecture [4]. Books are self-paced, but limited to static text and graphics. Real-time lectures, whether they are in-person or watched online, are paced by the instructor and require students to follow along with the instructor with only one order of content presentation. Multimedia provides the self-pacing of a text while allowing for audio present in a lecture and the adding of animation to graphics.

2) The Testing Effect

The potential effectiveness of including questions as part of course content could be explained by the testing effect, which is discussed by Roediger and Karpicke [5]. The testing effect is the effect by which retention of material is generally improved if an individual is tested on it. The theory behind this effect is that the act of retrieval from memory results in improved future recall [5, 6]. If the effect is strong enough it could improve retention of instructional videos that have questions embedded directly in them.

978-1-5090-5920-1/17/$31.00 ©2017 IEEE

3) The Testing Effect with Feedback

Butler and Roediger [7] discuss the advantages and disadvantages of using multiple-choice questions in an educational setting. The most prominent advantage is that they are a practical and reliable way to assess students. The drawback of multiple-choice questions is that they introduce incorrect answer choices, which can result in incorrect knowledge being acquired. In this regard, the testing effect is reinforcing the wrong answer. This is where the authors note that feedback comes into play by both helping to correct errors in knowledge and improving retention of knowledge. They put forth that the minimum feedback message should provide the correct answer. It may also present the original question, which is particularly important if there is a delay between the time the question is presented and when the feedback is presented.

The experiment by Butler and Roediger [7] looked at variances caused by the amount of studying prior to the test (no studying, studying, and restudying), the number of alternative answer choices given (two, four, or six), the type of feedback given (no feedback, immediate feedback, and delayed feedback), and reporting type (free reporting or forced reporting). Of particular interest for this study is the effect of feedback on overall performance. Embedding questions within instructional content creates an opportunity to provide feedback to students on these questions.

The highest proportion of correct responses found in the experiment by Butler and Roediger [7] resulted from providing delayed feedback, followed by immediate feedback and then no feedback. Overall feedback during testing resulted in increased performance during the posttest. The authors note, however, that without a sufficient amount of studying the amount of misinformation acquired by students through a multiple-choice questions test increases. Changing the number of alternative answer choices did not result in significantly lower posttest scores, but scores decreased as the number of answer choices increased.

4) The Testing Effect with Multimedia.

Johnson and Mayer [8] examined the testing effect in a more practical educational setting of a multimedia lesson. They compared students who studied and then took a retention test with students who studied twice (restudy). The experiment also tested knowledge transfer from topics covered in the multimedia lesson to the solving of novel problems. Questions in this study include making predictions about the system or modifying the system to achieve a goal.

The study showed that a retention test did improve retention over restudying when used with multimedia as it had with other non-multimedia materials. It also showed that knowledge transfer with educational material can be improved by the testing effect. From a practical standpoint in education, this means that testing can be used not only as a tool for assessment, but also for teaching. The study demonstrated that such quizzes can even target application of knowledge in related problems by including transfer questions.

While Johnson and Mayer [8] used retention tests after the video, this study focused on how including retention questions embedded within a multimedia lesson would affect knowledge transfer to problem solving. It also varies in that a restudy group is not used and the comparison is between having and not having retention testing within the same multimedia video. The questions also include feedback, which, as previously discussed, provides additional benefits to multiple-choice questions.

5) Deep-Level Reasoning Questions

As important as it is to reinforce a learner's understanding of a topic through testing, testing should also improve a learner's ability to understand and apply new material. A large body of research has demonstrated that problem solving and comprehension can be improved by encouraging students to ask questions [9, 10, 11, 12, 13, 14, 15, 16] or by simply presenting questions to students [17, 18].

Additionally, these questions have been shown to be effective for improving student learning. However, the type of question used matters. Deep-level reasoning questions have been shown to be more effective for learning than various control conditions [19]. Pairing content with deep reasoning questions has even been shown to be more effective than highly dynamic interactive systems [20] and just as effective as teachers [21]. A deep reasoning question requires multiple word answers (e.g., sentences or paragraphs) that involve logical, causal, or goal orientation reasoning processes to answer. These questions typically address causality by asking how or why a given effect was produced [20, 22, 23, 24]. It is possible that these questions would be even more effective if students were encouraged to answer them [25].

C. Problem Statement

The primary question that will be addressed by this study is:

> *Does embedding questions about concepts into instructional videos improve learning outcomes in an applied programming problem where participants write original code?*

Based on this question, the hypothesis proposed is the following:

> *Learners who view an instructional video with questions embedded in it will perform better on a graded programming assignment than those presented a video without embedded questions.*

II. METHODS

A. Design and Variables

This experiment used a pretest-posttest control group design [26]. Dependent variables being measured were results from pretests and posttests. The pretest measured prior knowledge from prerequisite courses. The posttest measured how successfully participants were able to implement concepts from the instructional video. The primary

independent variable was the use of instructional video type, which was either a standard video or an interactive video that included embedded questions about the content being presented. The participants completed the experiment as an individual in-class activity using the Progressive Learning Platform (PLP) Tool [27, 28]. They were shown one of the two versions of the instructional video on PLP and then they were given a programming assignment as a posttest.

This experiment was run with one control group and one treatment group. Each group completed a pretest and posttest. This allowed conclusions to be drawn about whether including embedded questions in the instructional video had an impact on participant's ability to complete the programming assignment.

B. Participants

There were 42 participants in the study. Participants in the study were students in an undergraduate course at a research-intensive PhD granting university in the United States of America. The course was required for students of a BS degree program in Software Engineering. Students were a mixture sophomores, juniors, seniors and graduate students. The control group was composed of 20 participants (17 men and 3 women) and the treatment group was composed of 22 participants (19 men and 3 women). Two outliers, one from each group, were removed due to adherence problems. One participant did not submit a program file and the other submitted an empty program file (the program only contained the project header that is automatically added when a new project is created).

The study was conducted as an in-class activity covering topics that were part of the standard curriculum for this course. This was done to ensure that participants would have similar relevant prior knowledge that would be adequate to prevent floor effects that may have been seen in participants who had no prior domain knowledge.

C. Materials

All aspects of this study, with the exception of the programming assignment attempts, were administered using Blackboard Learn [29]. Video content was embedded within a timed Blackboard assessment with separate timed assessments used for the pretest and posttest. Timing within the assessments was used to ensure all participants adhered to the time limits at each stage of the study.

1) Instructional Video.

Content for both the control and treatment group was produced from a single instructional video covering the assembly language, PLP, and the software tool used to write and simulate PLP assembly, PLP Tool [27]. The instructional video contained a mixture of narrated screen recordings showing both PowerPoint presentation slides and a screen capture showing PLP Tool being used. The video begins with a screen recording which shows simulated output of the program the participants will write after the video concludes. A high-level overview of the purpose of assembly language is given, followed by an introduction of assembly language instructions necessary to complete the study. Each instruction's meaning is introduced along with example usage to present proper syntax. Two code examples, which combine multiple instructions and concepts, are explained and presented within PLP Tool. This was done to explain the relationship between concepts and to demonstrate features of PLP Tool.

Principles of effective multimedia were used to explain the concept of memory mapped input/output, which instructors and teaching assistants for the course have observed to be a commonly confusing topic for students. A graphic representation of a CPU and memory bus was shown in the instructional video with individual elements of the graphic added one at a time and explained in the video narration. Figure 1 shows the graphic at two different stages of the video.

The control group was presented the instructional video as an embedded YouTube video with no interactive features. The treatment group was presented with the same video through Zaption [30], which allowed embedded activities such as multiple-choice questions to be added to standard YouTube videos. The Zaption service was discontinued shortly after the experiment in this paper was conducted.

Multiple-choice questions were added immediately after each new topic. Figure 2 shows a screenshot of the instructional video as it appears using the Zaption website. In

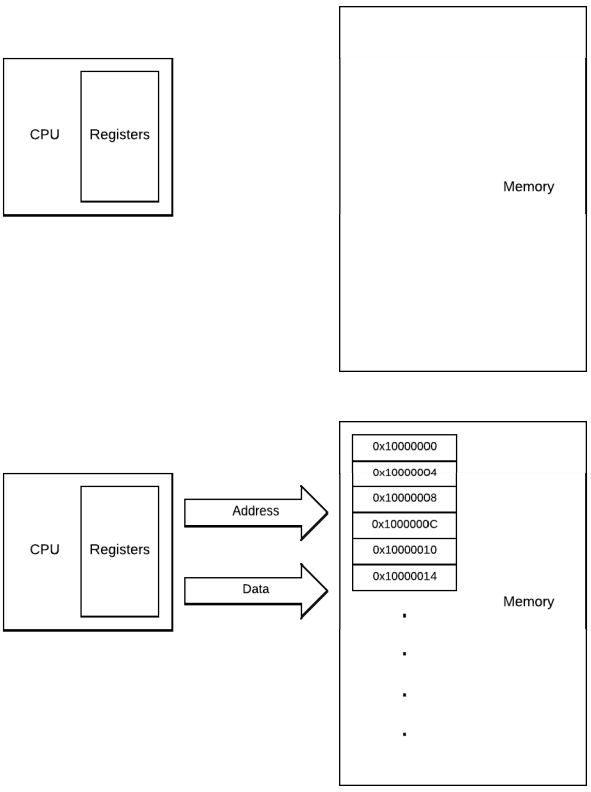

Fig. 1. Instructional video screenshots showing a graphic at two different points during the video as elements are added and explained with narration.

the screenshot, the video has reached a point where a multiple-choice question is being asked. Zaption has several options regarding how the questions are displayed to the user. For consistency, all questions in the instructional video were presented to the right of the video, which is the default display setting. Other options for displaying questions include showing questions to the left of the video and in the center covering the video. Feedback text replaces the question in the region to the right of the video immediately after clicking the "submit" button. Feedback for each question included the correct answer and an explanation of why that answer was correct. If an incorrect answer was given it also explained why the answer selected was incorrect.

2) Programming Assignment Description

A description of the program requirements was created to present to the students on Blackboard during the programming assignment portion of the experiment. This description gave details about how the program should function when complete. It also included two elements meant to mimic a subset of the materials a student might use while working on a similar project during the course. The first element was a reference table containing example usages of the assembly language instructions covered in the video and a description of the meaning of this instruction. This table is similar to a table provided in the documentation for PLP. The second element is a figure taken from the instructional video that covers features of the PLP Tool user interface needed to test the program.

The program participants were asked to implement counts up indefinitely starting from zero using a loop. Each time the counter is increased the value of the counter is represented in binary using an array of LEDs (Light Emitting Diodes).

D. Measures

1) Pretest

A pretest was used to determine successful randomization of participants. The pretest is designed to measure prior knowledge and does not directly test content presented in the study in order to avoid sensitizing participants. It contains four short answer questions and two multiple-choice questions.

The first four questions target knowledge covered in the course participants were recruited from. The first two questions cover number representations and require the application of conceptual knowledge to determine the correct value. The first is a multiple-choice question, but provides a "none of the above option". The second is a short answer question that asks the participant to convert a value from decimal to hexadecimal representation. The next two questions cover Boolean operators and also require participants to apply a procedure.

The last two questions are targeted at identifying potential outliers who may have already had experience with topics related to those in this study such as computer microarchitecture. The first of these questions is a short answer question where the concept being tested is whether the participant understands the limitations of a processor in handling numbers that are larger than its registers. The last question is a multiple-choice question to determine if the participant is already familiar with memory mapped input/output, which is an underlying principle in the programming assignment.

2) Posttest

Participant performance was measured using a programming assignment that combines all of the instructions and concepts covered in the instructional video. Programming assignments were scored based on whether they met functional requirements given in the assignment description.

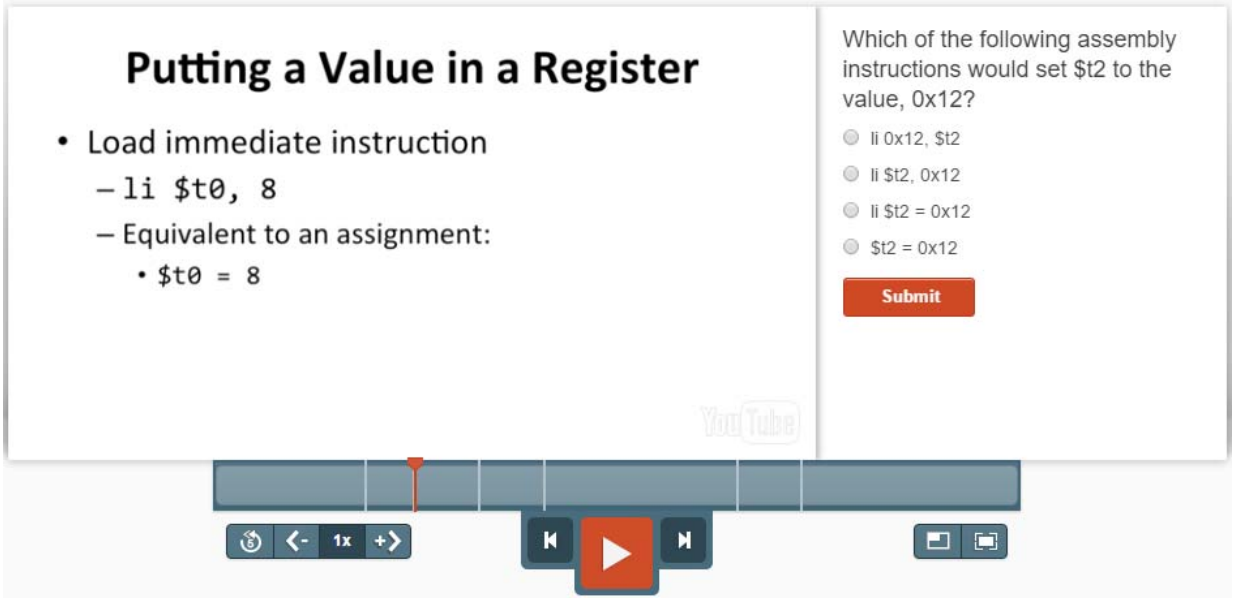

Fig. 2. Zaption video screenshot with an embedded question displayed.

3) Procedure

Participants first completed a 10-minute pretest in which concepts important to understand content in the study were tested, followed by a 20-minute instructional video. Next each participant was allowed 25 minutes to make an attempt at the programming assignment.

III. Pilot Studies

Two pilot studies were conducted primarily as usability studies. The main concerns before running the final experiment were if participants would be able complete all the tasks in the study during the given amount of time and if all content critical to the completion of the programming assignment was adequately addressed.

A. Participants

Both pilot studies were conducted with participants that were taking online coureses from the same university. These courses were prerequisites for the course from which material in this study was drawn. These online courses were chosen in order to have sample populations with experience and knowledge similar to participants in the final experiment.

B. Materials

The only changes in material between the pilot studies and the final experiment were the use of Qualtrics [31] for administration instead of Blackboard and the modifications made to the instructional video to match adjustments made to the programming assignment. Qualtrics is a professional online data collection platform commonly used to conduct research and customer surveys. Changes to the programming assignment between the two pilot studies and rationale for these changes is detailed below.

1) First Pilot Study

While designing instructional content for this study, there was some uncertainty about the concepts and scope of the programming assignment because programs are typically assigned as projects that span multiple days or weeks rather than a single class period. The results from the first pilot study showed that the programming assignment was too challenging to complete in the given amount of time for the target participants. Considerable floor effects were seen in programs submitted. Only half of the 28 participants in the first pilot study submitted assembly code, and none of the participants submitted a fully functional solution.

A reference table was provided during the programming assignment with example usage of the assembly instructions needed to complete the assignment, given in the same order they were presented in the instructional video. Example usages of the assembly instructions were arbitrary and would not produce the required behavior without modification as described in the instructional video. Despite this, it was common to see programs composed of lines that had been copied and pasted from the reference table examples with no modification. In some of these instances the order of the instructions in a submitted program was also the same as the ordering of the examples in the reference table.

2) Second Pilot Study

A simpler programming assignment was designed for the second pilot study, which did not require the use of one assembly instruction from the previous assignment. This modification had the desired result of reducing the floor effect and allowing more variability in programming assignment scores. This simplified programming assignment and the instructional video corresponding to this change were used for the final experiment in this study.

C. Measures

The first pilot study programming assignment was scored using a rubric similar to the rubric used in the final experiment, but with one more grading criterion for the additional required assembly instruction. The second pilot study, which had the same programming assignment as the final experiment, used the same rubric as the final experiment.

IV. Results

A. Posttest Results

Table 1 shows the mean score and standard deviation for each of the four grading criteria. On the left are the results for participants in the control group and on the right are the results for participants in the treatment group. The final row shows the mean and standard deviation for the overall score on the posttest.

TABLE I. Posttest Results

Grading Criteria	Treatment Condition			
	Control Group		*Treatment Group*	
	Mean	*Std. Dev.*	*Mean*	*Std. Dev.*
Load Immediate	.90	.13	.93	.11
Store Word	.78	.38	.67	.42
Control Flow	.90	.26	1.00	.00
Addition	.95	.15	.97	.12
Overall Score	3.53	0.58	3.57	0.55

B. Quantitiative Analysis

Statistical analysis was performed to compare differences between each of the groups for each grading criterion and for their overall posttest scores. The experimental design had two groups into which participants were randomized. Based on this design, independent samples t-tests were conducted on the dependent measures. Because the study is testing a directional hypothesis that embedded questions would have a positive effect, one-tailed t-tests were performed. If a standard t-distribution table is consulted [32, pp. 693], it will be noted that the .05 α level for a one-tailed t-test is identical to the .10 α level of a two-tailed t-test. The t-statistic is conducted the

same regardless of the directionality of the hypothesis. Because the methods are the same, statistical software packages calculate two-tailed t-tests. When a two-tailed t-test is conducted, an α level of .1 can be used instead of the traditional .05 to equate for the one-tailed t-test. Thus, an α level of .1 is used in the analyses to determine significance in order to account for the one-tailed (directional) hypothesis and predictions [32]. Analysis also includes Cohen's d measurements of effect size. Effect sizes indicates the strength of an effect with larger numbers indicating stronger effects. In general, a Cohen's d effect size of .2 is considered small, .5 is a medium, and .8 and above is a large [33]. It should be noted that effect sizes are sensitive to the variability in the data, and increases in environmental noise (uncontrollable events) can reduce the effect size [26]. The natural classroom setting used in this experiment will, as a result, have a tendency towards smaller observed effect sizes.

1) Load Immediate

A t-test was conducted on the average scores for the load immediate grading criterion to determine if there was a difference in scores between participants who viewed the standard video and the interactive video. The test did not reveal a statistically significant difference, but there was a small effect size in the predicted direction ($t(40) = 0.861$, $p = .4$; $d = 0.25$). The observed mean was M=.90; SD=.13 for control group and M=.93; SD=.11 for the treatment group.

2) Store Word

A t-test was conducted on the average scores for the store word grading criterion to determine if there was a difference in scores between participants who viewed the standard video and the interactive video. The test did not reveal a statistically significant difference, but there was a small effect size in the opposite direction from what was predicted resulting in a negative effect size ($t(40) = 0.845$, $p = .4$; $d = -0.27$). The observed mean was M=.78; SD=.38 for control group and M=.67; SD=.42 for the treatment group.

3) Control Flow

A t-test was conducted on the average scores for the control flow grading criterion to determine if there was a difference in scores between participants who viewed the standard video and the interactive video. The test revealed a statistically significant difference with a medium effect size in the predicted direction ($t(19) = 1.710$, $p = .1$; $d = 0.54$). The observed mean was M=.90; SD=.26 for control group and M=1.00; SD=.00 for the treatment group. The significance most likely would have been higher had it not been for a ceiling effect. Every participant in the treatment group earned full points on this criterion.

4) Addition

A t-test was conducted on the average scores for the addition grading criterion to determine if there was a difference in scores between participants who viewed the standard video and the interactive video. The test did not reveal a statistically significant difference and the effect size did not reach a small level, but the effect was in the predicted direction ($t(40) = 0.379$, $p = .7$; $d = 0.15$). The observed mean was M=.95; SD=.15 for control group and M=.97; SD=.12 for the treatment group. Again, a ceiling effect is observed, with near perfect scores from both groups.

5) Overall

A t-test was conducted on the average total posttest scores to determine if there was a difference in scores between participants who viewed the standard video and the interactive video. The test did not reveal a statistically significant difference, but there was a small effect size in the predicted direction ($t(40) = 0.246$, $p = .81$; $d = 0.23$). The observed mean was M=3.53; SD=0.58 for control group and M=3.57; SD=0.55 for the treatment group.

V. Discussion

In most cases, with the exception being the store word instruction, the data shows that embedding questions in instructional videos results in small improvements in learner performance compared to the presentation of the same video without embedded questions. The effect sizes were not as strong as they might have been in a laboratory setting, which is to be expected, but they were present in the predicted direction.

As illustrated in Figure 3, the current findings could be impacted by sensitivity issues in the materials or tests on three of the four topics. In the case of load immediate, control flow instructions, and addition, there was a ceiling effect with all participants earning the maximum possible score or close to it. The store word topic was the only topic that did not show potential ceiling effects. Since the result was seen on store word topic, but others exhibited ceiling effect and did not show differences, this points to potential problems with difficulty levels in the materials being taught or the assessments. It is possible that the assessments were not sensitive enough to show results in these cases. While the overall results were not conclusive, they do warrant further research.

While reviewing the mean scores for the grading criteria a plausible explanation was observed for the relatively low mean score by both groups for the store word grading criterion. According to the serial position effect, subjects are most likely to remember information mentioned earlier on due to the primacy effect and information mentioned towards the end due to the recency effect [33]. This results in a U-shaped curve with information closest to the beginning or end having the highest probability of being recalled and information in the middle having a lower probability of being recalled.

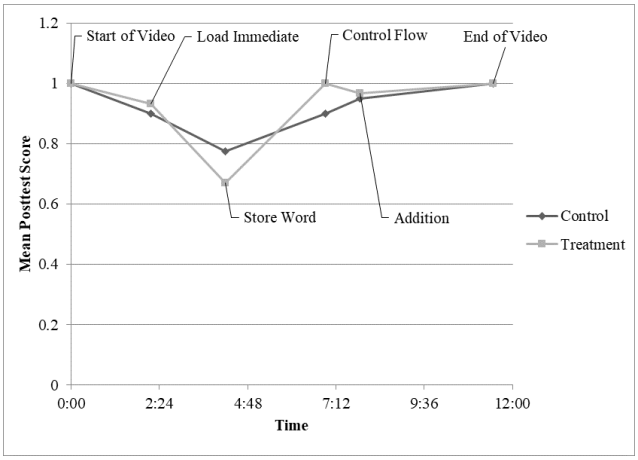

Fig. 3. Posttest scores vs. time of subject introduction

Figure 3 shows the mean score for both groups at the time in the instructional video when the presentation of information related to each grading criterion begins. The start and end times of the video have been plotted with a mean score of 1 to show the general trend and its similarity to the serial position effect, however no posttest measurements were taken at these points in time. From this figure, it can be seen that material related to the store word grading criterion began toward the middle of the instructional video, which is where mean scores are also the lowest compared to other criteria.

Topics related to the store word criterion are some of the more challenging topics covered in the instructional video, in part because they build on previous topics. Content in the instructional video was arranged in this order because it made the most sense to build concepts on top of one another, but this could have led to the unintended consequence of placing this information where it was least likely to be recalled. Small consideration like this demonstrate how complex and challenging it can be to generate effective instructional content based on available research findings. These challenges are well documented and many researchers have put forth suggestions to help bridge the gap between evidence-based practices and teaching methods used in engineering classrooms [35, 36, 37, 38, 39, 40].

Mayer's [41] work on cognitive theory in multimedia suggests that one challenge students may face when presented questions embedded within a video is the added load to working memory. Quickly switching from watching and listening to a video to reading and thinking about a question may overload a student's working memory. From this perspective, the questions would serve as a distraction, much as separated text and graphics were distracting in the work done by Moreno and Mayer [3]. Alternatively, the switch from watching and listening to interacting could be preventing the overloading of working memory by reducing the amount of new information presented to the student at one time and requiring the retrieval of memory (i.e. the testing effect).

Instructional videos with embedded questions have an added benefit for instructors of assessing instructional content during consumption. Student engagement is difficult to measure with standard instructional videos that do not contain interactive content, but with embedded questions (and analytics, such as those provide by Zaption) instructors can see viewing patterns and answer choices for individual students. Such information is useful for identification of students who may need additional attention and areas of instructional content that may need to be improved.

VI. Conclusion

This paper compares the performance between learners in two groups. One group watched an instructional video without any interactive content and the other group watched the same instructional video, which had been augmented with embedded questions. The results warrant further research after findings indicate that some improvements can be seen from learners when questions with feedback are embedded within instructional content. This finding means it may be worthwhile for instructional content creators, particularly those in the domains that teach computer programming, to include embedded questions and feedback in their instructional videos.

VII. Future Work

In this study only quantitative data was collected and analyzed. In a future study, qualitative data could be collected including process data and student preferences. This data could reveal what parts of an instructional video work and what happens when participants were struggling.

Alternative services to Zaption could also be explored. Other services that allow for the creation of interactive instructional content may provide different tools and methods for presenting questions and feedback. These variations could potentially offer improved effectiveness of instructional videos with embedded questions. In their announcement that they were ending their service, Zaption listed HapYak and HP5 as potential replacements. In addition to these services, creating a course through a MOOC service such as edX or Coursera is also an option provided that making course content publically available through one of these services is allowed by the researcher or course creator's institution.

References

[1] N. Mamgain, A. Sharma and P. Goyal, "Learner's perspective on video-viewing features offered by MOOC providers: Coursera and edX," in *MOOC, Innovation and Technology in Education (MITE)*, 2014.

[2] R. E. Mayer, "Multimedia learning: are we asking the right questions," *Educational Psychologist,* vol. 32, no. 1, pp. 1-19, 1997.

[3] R. Moreno and R. E. Mayer, "Cognitive principles of multimedia learning: the role of modality and contiguity," *Journal of Educational Psychology,* vol. 91, no. 2, pp. 358-368, 1999.

[4] R. C. Clark and R. E. Mayer, E-learning and the science of instruction: proven guidelines for consumers and designers of multimedia learning, John Wiley & Sons, 2016.

[5] H. L. Roediger and J. D. Karpicke, "The power of testing memory: basic research and implications for educational practice," *Perspectives on Psychological Science,* vol. 1, no. 3, pp. 181-210, 2006.

[6] C. Brochok, C. Mar and S. D. Craig, "Is free recall active: the testing effect through the ICAP lens," *Journal of Interactive Learning Research,* vol. 28, no. 2, pp. 127-148, 2017.

[7] A. C. Butler and H. L. Roediger, "Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing," *Memory & Cognition,* vol. 36, no. 3, pp. 604-616, 2008.

[8] C. I. Johnson and R. E. Mayer, "A testing effect with multimedia learning," *Journal of Educational Psychology,* vol. 101, no. 3, pp. 621-629, 2009.

[9] S. D. Craig, B. Gholson, M. Ventura, A. C. Graesser and The Tutoring Research Group, "Overhearing dialogues and monologues in virtual tutoring sessions: effects on questioning and vicarious learning," *International Journal of Artificial Intelligence in Education,* vol. 11, pp. 242-253, 2000.

[10] B. Davey and S. McBride, "Effects of question generating training on reading comprehension," *Journal of Educational Psychology,* vol. 78, pp. 256-262, 1986.

[11] J. R. Gavelek and T. E. Raphael, "Metacognition, instruction, and the role of questioning activities," *Metacognition, Cognition, and Human Performance,* vol. 2, pp. 103-136, 1985.

[12] A. C. Graessar, W. Baggett and K. Williams, "Question-driven explanatory reasoning," *Applied Cognitive Psychology,* vol. 10, pp. S17-S32, 1996.

[13] A. King, "Effects of self-questioning training on college students' comprehension of lectures," *Contemporary Educational Psychology,* vol. 14, pp. 366-381, 1989.

[14] A. King, "Guiding knowledge construction in the classroom: effect of teaching childern how to question and explain," *American Educational Research Journal,* vol. 31, pp. 338-368, 1994.

[15] A. King, A. Staffieri and A. Adelgais, "Mutual peer tutoring: effects of structoring tutorial interaction to scaffold peer learning," *Journal of Educational Psychology,* vol. 90, pp. 134-152, 1998.

[16] A. S. Palincsar and A. L. Brown, "Reciprocal teaching of comprehension fostering and comprehension monitoring activities," *Cognition and Instruction,* vol. 1, pp. 117-175, 1984.

[17] S. D. Craig, B. Gholson, J. K. Brittingham, J. Williams and K. T. Shubeck, "Promoting vicarious learning of physics using deep questions with explanations," *Computers & Education,* vol. 58, pp. 1042-1048, 2012.

[18] B. Gholson, R. Coles and S. D. Craig, "Features of computerized multimedia environments that support vicarious learning process," *New Science of Learning,* pp. 53-77, 2010.

[19] D. Driscoll, S. D. Craig, B. Gholson, M. Ventura, X. Hu and A. Graesser, "Vicarious learning: effects of overhearing dialog and monolog-like discourse in a virtual tutoring session," *Journal of Educational Computing Research,* vol. 29, pp. 431-450, 2003.

[20] S. D. Craig, J. Sullins, A. Witherspoon and B. Gholson, "Deep-level reasoning questions effect: the role of dialog and deep-level reasoning questions during vicarious learning," *Cognition and Instruction,* vol. 24, no. 4, p. 565–591, 2006.

[21] S. D. Craig, A. Graesser, J. Brittingham, J. Williams, T. Martindale, G. Williams, R. Gary, A. Darby and B. Gholson, "An implementation of vicarious learning environments in middle school classrooms," in *The Proceedings of the 19th International Conference for the Society for Information Technology & Teacher Education*, Chesapeake, VA, 2008.

[22] S. D. Craig, "Questioning," in *International Guide to Student Achievement*, London, Routledge, 2012, pp. 414-415.

[23] A. C. Graessar and N. K. Person, "Question asking during tutoring," *American Educational Research Journal,* vol. 31, no. 1, pp. 104-137, 1994.

[24] A. C. Graesser, N. K. Person and J. P. Magliano, "Collaborative dialogue patterns in naturalistic one-to-one tutoring," *Applied cognitive psychology,* vol. 9, no. 6, pp. 495-522, 1995.

[25] H. Garcia-Rodicio, "Questioning as an instructional strategy in multimedia environments: does having to answer make a difference?," *Journal of Educational Computing Research,* vol. 52, no. 3, pp. 365-380, 2015.

[26] W. R. Shadish, T. D. Cook and D. T. Campbell, Experimental and quasi-experimental designs for generalized causal inference, Wadsworth Cengage Learning, 2002.

[27] "PLP Tool Manual," 2017. [Online]. Available: http://progressive-learning-platform.github.io/home.html. [Accessed 2017].

[28] W. D. Mulia, D. J. Fritz, S. Sohoni, K. Kearney, and M. Mwavita, "PLP: A Community Driven Open Source Platform for Computer Engineering Education" *International Journal of Engineering Education*, vol 29 no. 1, pp. 215-229, 2013.

[29] "Blackboard Learn," 2017. [Online]. Available: http://www.blackboard.com/learning-management-system/blackboard-learn.aspx. [Accessed 2017].

[30] "Zaption," Zaption, 2016. [Online]. Available: www.zaption.com. [Accessed 2016].

[31] "Qualtrics," Qualtrics, 2017. [Online]. Available: www.qualtrics.com. [Accessed 2017].

[32] F. J. Gravetter and L. B. Wallnau, Statistics for the behavioral sciences, Cengage Learning, 2016.

[33] J. Cohen, Statistical power analysis for the behavioral sciences, Hillsdale, NJ: Lawrence Eribaum Associates, 1988.

[34] D. Rundus, "Analysis of rehearsal processes in free recall," *Journal of Experimental Psychology,* vol. 89, no. 1, pp. 63-77, 1971.

[35] M. Besterfield-Sacre, M. F. Cox, M. Borrego, K. Beddoes and J. Zhu, "Changing engineering education: views of U.S. faculty, chairs, and deans," *Journal of Engineering Education,* vol. 103, no. 2, pp. 193-219, 2014.

[36] J. Feser, M. Borrego, R. Pimmel and C. Della-Piana, "Results from a survey of National Science Foundation Transforming Undergraduate Education in STEM (TUES) program reviewers," in *ASEE Annual Conference & Exposition*, San Antonio, TX, 2012.

[37] C. Henderson and M. H. Dancy, "Barriers to the use of research-based instructional strategies: the influence of both individual and situational characteristics," *Physical Review Special Topics-Physics Education Research,* vol. 3, no. 2, 2007.

[38] C. Henderson, N. Finkelstein and A. Beach, "Beyond dissemination in college science teaching: an introduction to four core change strategies," *Journal of College Science Teaching,* vol. 39, no. 5, pp. 18-25, 2010.

[39] S. Sohum and S. D. Craig, "Making the case for adopting and evaluating innovative pedagogical techniques in engineering classrooms," in *ASEE Annual Conference and Expo*, New Orleans, LA, 2016.

[40] S. Sohoni, S. D. Craig and K. Vedula, "A blueprint for an ecosystem for supporting high quality education for engineering," *Journal of Engineering Education Transformation.*, in press.

[41] R. E. Mayer, "Cognitive theory and the design of multimedia instruction: an example of the two□way street between cognition and instruction," *New directions for teaching and learning,* vol. 2002, no. 89, pp. 55-71, 2002.