International **R**eview of **R**esearch in **O**pen and **D**istributed **L**earning
Volume 18, Number 7

November – 2017

Explanations and Interactives Improve Subjective Experiences in Online Courseware

Marshall P. Thomas[1], **Selen Türkay**[2], and **Michael Parker**[1]
[1] Harvard Medical School, [2] Queensland University of Technology

Abstract

As online courses become more common, practitioners are in need of clear guidance on how to translate best educational practices into web-based instruction. Moreover, student engagement is a pressing concern in online courses, which often have high levels of dropout. Our goals in this work were to experimentally study routine instructional design choices and to measure the effects of these choices on students' subjective experiences (engagement, mind wandering, and interest) in addition to objective learning outcomes. Using randomized controlled trials, we studied the effect of varying instructional activities (namely, assessment and a step-through interactive) on participants' learning and subjective experiences in a lesson drawn from an online immunology course. Participants were recruited from Amazon Mechanical Turk. Results showed that participants were more likely to drop out when they were in conditions that included assessment. Moreover, assessment with minimal feedback (correct answers only) led to the lowest subjective ratings of any experimental condition. Some of the negative effects of assessment were mitigated by the addition of assessment explanations or a summary interactive. We found no differences between the experimental conditions in learning outcomes, but we did find differences between groups in the accuracy of score predictions. Finally, prior knowledge and self-rated confusion were predictors of post-test scores. Using student behavior data from the same online immunology course, we corroborated the importance of assessment explanations. Our results have a clear implication for course developers: the addition of explanations to assessment questions is a simple way to improve online courses.

Keywords: assessment, feedback, affect, confusion, online course design

Introduction

Many researchers have evaluated different elements of computerized instruction using experimental and observational methods (e.g., Clark & Mayer, 2011; Szpunar, Khan, & Schacter, 2013; Türkay, 2016). Often these studies were conducted in-person in laboratories rather than online, and few have utilized authentic online course materials. Moreover, the majority of the studies have assessed learning outcomes with an immediate follow-up rather than a more educationally pertinent delayed follow-up test. The current work is guided by experiments performed by Szpunar, Khan, and Schacter (2013) and Szpunar, Jing, and Schacter (2014). The results of these studies support the idea that the interleaving of assessments with short videos enhances learning while reducing mind wandering and overconfidence. In these studies, participants were tested immediately after instruction with the same test items as the study materials. Finally, the assessments in these studies were open response type assessments, which cannot reliably be machine-graded at scale.

To address some of the shortcomings of the studies mentioned above, our goal in this study was to investigate the effectiveness of the types of instructional sequences that are widely available on common online course platforms while utilizing course materials used in real online courses. Moreover, we conducted this study fully online, not in a laboratory setting, to experimentally match the intended delivery modality of these materials. To investigate educationally relevant outcomes, the post-test was administered not as an immediate follow-up, but seven days after instruction. The post-test covered content-matched items but did not directly replicate assessment items that participants encountered in the instruction. We made these design choices in order to study retention and near transfer of the material rather than memorization of the test items and their correct responses. Finally, we varied certain parameters experimentally to examine the influence of instructional choices on students' experiences in and engagement with online courses.

In settings such as massive open online courses (MOOCs) and small private online courses (SPOCs), video is the most common instructional modality (Hansch et al., 2015). Instructors have the option to supplement video instruction with text, assessment, and/or other forms of interaction (simulation, discussion, and interactives, to name a few). These instructional modalities can involve very different production costs, levels of interactivity, and afford distinct opportunities to collect behavior and performance data. We wanted to compare the effectiveness of common modalities that are used in online courses alongside video on learners' affective and cognitive outcomes. For this work, we selected text (which highlighted concepts from videos), assessment (which reinforced concepts from videos), assessment with explanations (which matched the assessment with additional feedback in the form of an explanation of the correct answer), and assessment with a summary interactive (which summarized the content covered in the instructional videos).

In this study, we focused on the impact of routine online course design choices by asking the following questions:

1. How does the addition of multiple choice or short-answer assessments between videos impact learning, persistence, and engagement in a fully online advanced science lesson?

2. How does simple feedback, in the form of explanations to assessment questions, impact learning, persistence, and engagement?

3. How does a summary interactive, which serves as a recap of the content of the lesson, impact learning, persistence, and engagement?
4. What readily measurable variables can be used to predict post-test performance after a learning session in an online course?

Background

The Testing Effect

It is well established that assessment promotes memory and learning (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Pyc & Rawson, 2010; Roediger & Karpicke, 2006). Mechanistically, learning-by-testing is caused by the direct effect of effortful retrieval and diverse indirect or "mediated" effects (for example, improved study behaviors and learning from feedback). The direct effects of testing are strongest with repeated, spaced retrieval of material. In laboratory experiments this has frequently involved repeated study of materials with low educational relevance, such as unrelated word pairs. Although the testing effect has been replicated in laboratory studies with educationally-relevant materials, classroom-based studies produce effect sizes that are typically smaller than those in simple, repeated testing memorization tasks (Gog & Sweller, 2015; Roediger & Karpicke, 2006).

The testing effect has not been extensively studied in online courses. Testing in MOOCs and SPOCs is often limited to simple machine-gradable assessment types (Daradoumis, Bassi, Xhafa, & Caballé, 2013). Understanding the affordances and limitations of these assessment types will guide their usage and help prioritize the development of new assessment types. In the case of formative assessments, the indirect effects of testing may be as important as the direct effects. For example, Agarwal and colleagues (2008) found that testing potentiated the effect of feedback given after the test. In a similar study that included testing or a control (reading content-aligned statements), the effect of testing was not as dramatic (Kang, McDermott, & Roediger, 2007). This suggests that an indirect effect of testing may be targeted re-exposure to the most important content. It is important to note that assessment *format* probably also interacts with the testing effect, as recognition-type testing tasks (multiple choice) are cognitively easier than open-response recall-type tasks (Cabeza et al., 1997). While the learning benefits of testing are well characterized, comparatively few studies have evaluated the impact of assessment in online courses on students' subjective experiences.

Feedback Enhances Learning

The term "feedback" has many connotations and is sometimes taken to refer specifically to personalized, expert-generated feedback (Margaryan, Bianco, & Littlejohn, 2015). In the current study, we adopt a broader definition of feedback that encompasses any information provided to students about their knowledge or their performance. Thus, we adopt the view that feedback can only exist as a response to a student activity, such as interacting with assessment (Hattie & Timperley, 2007). It is well established that many different forms of feedback, including computerized feedback, are effective for promoting learning, but the details do matter. For example, feedback with praise is less effective than feedback without praise (reviewed by Hattie & Timperley, 2007). Feedback can work by providing cues or reinforcement to learners, which could include information about learners' current level of knowledge. In online courses, assessment is one common method used to create opportunities for feedback.

Some researchers have argued that there is a particular advantage to *adaptive* computerized feedback, that is, feedback tailored to students' specific misconceptions or errors (for example, Lütticke, 2004). In a recent study, the authors directly compared a very simple form of feedback (knowledge of the correct response) to adaptive feedback, and found that students preferred the more elaborative adaptive feedback (D'Antoni et al., 2015). However, the authors did not examine the impact of generic yet elaborative feedback (such as written explanations of correct answers) to more simple forms of feedback (binary feedback that simply indicates whether an answer is correct or not). This leaves open the possibility that elaborative but generic feedback, such as question explanations, may provide some of the same benefits as adaptive feedback. Elaborative feedback, while simple to add to assessment, is not particularly common in computerized instruction. In a review of common adaptive assessment systems, Saul and Wuttke (2011) found that most systems only provide students with knowledge of the correct response, without further elaboration. In MOOCs, there can be a tradeoff between the quality and quantity of feedback (Ebben & Murphy, 2014).

Optimizing Engagement and Learning in Online Courses

Observational Studies. There has been a great deal of research into patterns of student engagement and learning in online courses, particularly analyses of big data generated from MOOCs. In MOOCs, student engagement is often equated with retention, which generally drops off over time (Ferguson & Clow, 2015; Kizilcec, Piech, & Schneider, 2013). Although these studies have been useful in characterizing patterns of engagement, they are less informative as to *why* students choose to engage or disengage with online courses. Affective elements such as motivation and intent to complete play a key role in engagement in online courses (Greene, Oswald, & Pomerantz, 2015; Reich, 2014; Wang & Baker, 2015), but course-specific factors can influence engagement. Studies indicate that the modality of online content delivery can influence both learning outcomes and engagement (Türkay, 2016). Moreover, video production style influences students' engagement with the videos (Guo, Kim, & Rubin, 2014), and learning outcomes (Chen et al., 2016). Design of online learning platforms can impact engagement by enabling learners to interact with instructors and other learners. In a qualitative study of MOOC learners, many learners highlighted a sense of community they developed through discussion forums and social media groups (Friedman, Liu, Morrissey, Turkay, & Wong, 2015). It is less clear from these studies how components of online courses other than video and discussions impact student engagement.

Certain student-level priors, including past academic performance, standardized test scores, level of educational experience, and some demographic factors can predict students' persistence and grades in traditional educational settings (Casillas et al., 2012; Geiser & Santelices, 2007). In fact, the entire concept of pre-requisites is based on the premise that prior subject knowledge is important for academic success. In MOOCs, completion rates may differ by students' level of prior educational attainment (Pursel, Zhang, Jablokow, Choi, & Velegol, 2016), but this is not always the case (Goldberg et al., 2015). Currently, there is a great deal of interest in identifying quantifiable predictors of student retention and engagement as a means to intervene early in both traditional and online educational settings.

Methods

Study Administration

Participant Recruitment and Screening. Overall study design is summarized in Figure 1. Study participants were recruited from Amazon Mechanical Turk using TurkPrime (www.turkprime.com) (Litman, Robinson, & Abberbock, 2016). The HITs (human intelligence tasks) associated with this study were only available to US-based workers with a HIT approval rate of 80% or higher to establish equality and quality of participants. First, we screened participants with a 5-question pre-test to assess their knowledge of biology (Appendix A). Workers received $0.25 for taking the test. The pre-test had a 3-minute time limit and was administered with Qualtrics (www.qualtrics.com). Three of the test items were drawn from an introductory biology concept inventory (Shi et al., 2010) and the remaining two were written by the study authors. All pre-test items were reviewed and revised in consultation with two additional PhD-level experts in biology. Participants who answered three or more items correctly on the pre-test were invited into the instructional phase of the study. Throughout the study, Mechanical Turk worker IDs were passed over to Qualtrics to permit data linking.

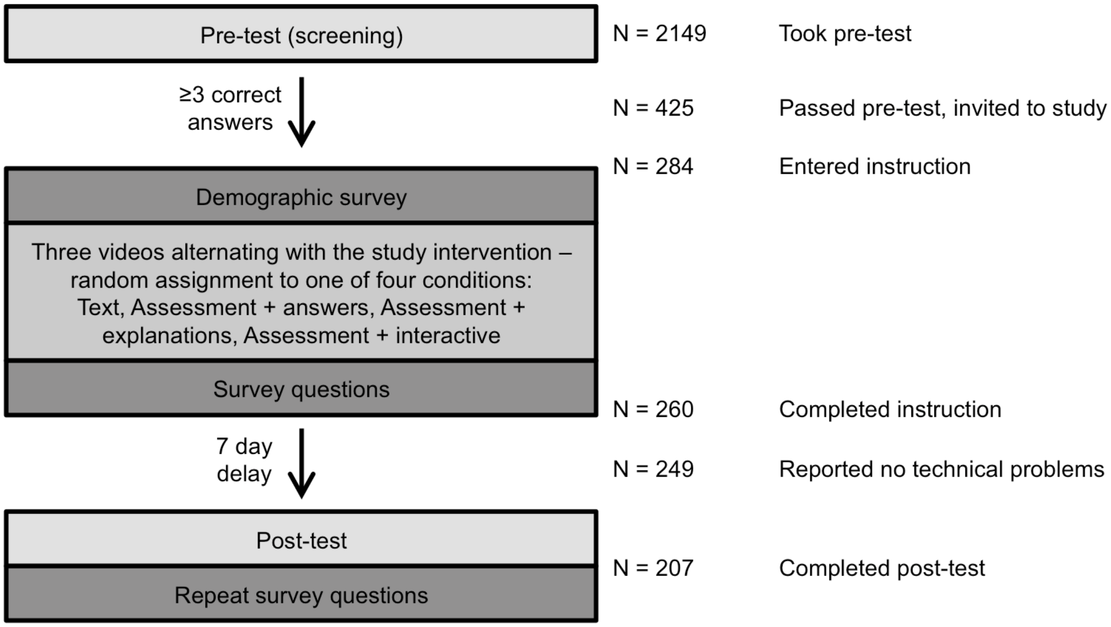

Pre-test (screening)
N = 2149
Took pre-test
≥3 correct answers
N = 425
Passed pre-test, invited to study
N = 284
Entered instruction
Demographic survey
Three videos alternating with the study intervention – random assignment to one of four conditions: Text, Assessment + answers, Assessment + explanations, Assessment + interactive
Survey questions
N = 260
Completed instruction
7 day delay
N = 249
Reported no technical problems
Post-test
Repeat survey questions
N = 207
Completed post-test

Figure 1. Overall study workflow and participant retention in the study. The phases of the study are indicated in the diagram on the left, while the number of participants at each phase are indicated on the right.

Instructional Phase. All instructional materials are part of actual online courses that are provided to students in pre-health care careers. The text and assessment materials were created by a PhD-level expert in immunology and reviewed and revised in collaboration with two MD/PhD immunology experts with decades of relevant teaching experience. The instructional phase of the study was administered in Qualtrics. Participants were paid $5.00 for this phase of the study, which took 32 minutes on average to complete. Before starting the instruction, participants consented to the study and filled out some basic demographic questions (Appendix A). Instruction consisted of three "whiteboard style" videos alternating with different activities (see Appendix B for a screenshot from an instructional video). The videos contained professionally-constructed visual elements representing immunological processes accompanied by a narration and written annotation of these visuals. The videos were planned and produced by a professor with an MD/PhD and decades of experience in teaching immunology, in collaboration with a professional medical illustrator and an MD with decades of medical education experience. The videos auto-played as soon as participants entered a page with the videos; participants could not see the video controls or advance forward until the videos ended. Participants were randomly assigned with equal probability to one of the four different experimental conditions:

1. Text: participants read a series of text statements that corresponded to assessment questions and their answers.

2. Assessment + answers: participants read assessment questions and answered them, then saw the correct answers after answering the questions.

3. Assessment + explanations: participants read assessment questions and answered them, then saw the correct answers with an explanation of the correct answers.

4. Assessment + interactive: participants read assessment questions and answered them, then saw the correct answers after answering the questions. At the end of the instruction, the participants navigated through a review interactive that summarized the steps of the immune process they studied in the lesson. The same team that made the videos produced the interactive.

Immediately after instruction, participants answered a set of survey questions about their study experiences (Appendix A). The survey questions were about the instruction as a whole and not specific to the experimental manipulation (text or assessment questions). We also asked participants to report whether they experienced technical problems during the experiment.

Post-Test and Survey

Seven days after instruction, participants who completed the instructional phase were invited to take a brief follow-up test administered in Qualtrics (Appendix A). The text and assessment materials were created by a PhD-level expert in immunology and reviewed and revised in collaboration with two other PhD-level experts in immunology. The post-test had eight multiple choice items and 10-minute time limit. The total possible score on the post-test was 10 points (two questions were "multiple selection" questions with two correct answers and thus were worth two points each). After the post-test, participants were asked to re-answer the original survey questions regarding their memory of their experiences in the lesson.

Data Analysis

Data Processing and Analysis

Data were downloaded from Qualtrics and pre-processed with Python. Participants who participated more than once were excluded at this phase. For statistical analyses, repeated measures of Likert scale data were averaged (1 = strongly agree, 5 = strongly disagree). Only participants who completed all phases of the study and reported no technical problems were included in the analyses of primary outcomes (N = 207). We also analyzed dropout rates. Compiled data were analyzed and plotted in R, and some summary results were exported to Excel for plotting.

Log File Analysis

In addition to the experimental data described above, we utilized log file data from three separate course runs of HMX Fundamentals – Immunology, the SPOC from which the instructional materials were drawn for this study. Log file data were parsed from JSON format and processed in Python; statistical analyses were performed in R. Overall, there were 69,043 unique assessment attempts of over 300 assessment questions. The summary results were plotted in Excel.

Statistical Analyses

All statistical tests were performed in R. The following statistical tests were used: Fisher's exact test (study dropout, gender distributions, MOOC participation), Kruskal-Wallis with post-hoc Mann-Whitney tests (Likert scale results, post-test scores), ANOVA with post-hoc t-tests (time on task), chi-squared test (show answer behavior), and one-sided t-tests and Bartlett's test of variance (post-test score predictions). Ordinal logistic regression was performed using the lrm model in the *rms* R package. The three conditions with assessment grouped together with respect to two primary outcomes (dropout during instruction and time on task), so these conditions were grouped for some analyses (see Appendix C). Because of the difference in dropout, the assessment groups were compared to each other and the text group data are only provided for visual reference.

Results

Text Compared to Grouped Assessment Conditions

Study retention and time on task. Using a pre-test, we screened more than 2,000 individuals; less than 20% passed and were invited into the study. The overall number of participants and retention in the study are summarized in Figure 1. We first compared the dropout between different experimental conditions. Surprisingly, there was a significant between-groups difference in dropout during the instructional phase of the study (*p* = .0046, Fisher's exact test). None of the participants in the text condition dropped out, while there was dropout in all of the other conditions with assessment, so we compared the text condition to the assessment conditions as a group (the "Grouped Assessment conditions"). Participants in the Grouped Assessment conditions had significantly higher dropout than those in the Text condition (*p* = .001, Fisher's exact test). Dropout from the Text condition was also significantly different from each of the assessment conditions individually ($p < .01$ in all cases, Fisher's exact test). However, there was no difference in dropout between the different assessment conditions ($p = .860$, Fisher's exact test). These results are summarized in Appendix C. We analyzed two additional sources of attrition: failure to return to the follow-up test, and exclusion due to reported technical problems. There was no significant difference in attrition between the different conditions due to either of these factors. We next compared time on task in the instruction phase of the study. There was a significant between-groups difference in time on task ($F(3)=14.84$, $p = 8.98\text{E-}09$, one-way ANOVA). Time on task was greater in the Grouped Assessment conditions than in the Text condition ($t(162.22)=-8.00$, $p = 2.20\text{E-}13$, Welch's two-sample t-test), and time on task in the Text condition was significantly different from each of the assessment conditions. However, there was no difference in time on task between the different assessment conditions ($F(2)= 0.012$, $p = .988$). These results are summarized in Appendix C.

Differences in Subjective Experiences

The differences in time on task and dropout between text and all assessment conditions suggest a meaningful difference in participants' experiences between these conditions. Therefore, we compared participants' self-rated subjective experiences in the Text condition with the Grouped Assessment conditions (Assessment + answers, Assessment + explanations, Assessment + interactive). Participants in the Grouped Assessment conditions reported lower levels of mind wandering, greater interest in the lesson, greater effort exerted in the lesson, and more strongly agreed that they would like to learn from similar lessons (Appendix C). There were no significant differences in self-rated difficulty, confusion, enjoyment, or understanding. These results suggest that assessment improved participants' subjective experiences relative to reading text statements, and in particular made the lesson materials more engaging. However, due to the differences in dropout during instruction, we cannot draw causal conclusions about the observed differences in subjective ratings between the Grouped Assessment and Text conditions.

Differences in Post-Test Scores

There was no difference between the Grouped Assessment conditions and the Text condition in the scores of a post-test administered seven days after the instructional phase of the study ($U=4210.5$, $n_1=59$, $n_2=148$, $p = 0.688$). We wanted to ensure that this result was not due to a sensitivity issue with the post-test (although the post-test was content-validated by two expert reviewers). We administered the test to a group of individuals who had passed the pre-test but did not go through instruction (the "instruction-naive" group). Post-test scores were significantly higher in the group of participants exposed to the Text condition ($U= 2305$, $n_1=59$, $n_2=49$, $p = 9.7\text{E-}08$) and Grouped Assessment conditions ($U=5795.5$, $n_1=148$, $n_2=49$, $p = 2.7\text{E-}10$) than the post-test scores of the instruction-naive group. Post-test results are summarized in Appendix C.

Negative and Positive Subjective Elements of Assessment

Differences in students' subjective experiences between assessment conditions. We compared the different conditions with assessment to look for any differences in participants' self-reported prior variables. The Text condition was not included in this analysis because of the significant difference in dropout between this condition and all others. There was no significant difference between any of the groups in any of the self-reported demographic data we collected, including gender and prior experience with MOOCs (Fisher's exact test), time spent weekly on learning activities (ANOVA), attained education level, and prior knowledge in basic biology, advanced biology, and immunology (Kruskal-Wallis test). Moreover, there was no significant difference in pre-test scores between the different groups, nor was there any difference in post-test score between the groups (Kruskal-Wallis test). We next compared the survey responses (Likert scale data) between these groups. The results of this analysis are summarized in Table 1. Significant differences are explored below.

Table 1

Summary of Survey Responses for Assessment Conditions

	Mean Likert Scale Score				
Statement	Assessment + answers *n*=53	Assessment + explanations *n*=46	Assessment + interactive *n*=49	*H*(2)	*p* value
My mind wandered during the lesson	3.44	4.11	3.58	11.544	0.0031
I understood the material in this lesson well	2.80	2.21	2.26	11.335	0.0035
I found this lesson difficult	2.43	3.04	2.88	7.604	0.0223
I found this lesson confusing	3.25	3.84	3.56	5.723	0.0572
I found the lesson interesting	1.74	1.52	1.79	4.694	0.0956
I exerted effort in this lesson	2.04	1.96	2.34	4.150	0.1255
I enjoyed learning from this lesson	1.86	1.58	1.71	3.167	0.2052
I would like to learn from more lessons like this one	1.94	1.70	1.80	1.691	0.4293

**Note*. Mean Likert scale scores are given (1 = Strongly Agree, 5 = Strongly Disagree). Between-groups comparisons were done with Kruskal-Wallis tests; test statistics and *p* values are reported in the right-hand columns.

Assessment explanations reduce perceived difficulty and mind wandering. We were surprised to find the largest average difference in subjective responses was found in participants' perceived level of mind wandering. Participants in both the Assessment + answers and the Assessment + interactive condition reported significantly greater levels of mind wandering than participants in the Assessment + explanation condition (Figure 2A). We hypothesize that explanations provided immediately after the questions is a form of feedback that increases the perception of interactivity. Explanations also affected perceptions of difficulty. The addition of explanations and an interactive reduced the average reported difficulty of the lesson, compared with assessment answers only (Figure 2B). However, the only significant difference in reported difficulty was between the Assessment + answers and Assessment + explanations conditions. This suggests that immediate feedback decreases the perception that material is challenging.

Assessment without additional feedback reduces score predictions and perceived understanding. We also observed a significant between-groups difference in self-rated understanding of the lesson material. Participants in the Assessment + answers condition reported the lowest levels of understanding of the lesson material; this effect was abrogated by the addition of explanations or a summary interactive (Figure 2C). This could indicate that explanations increase students' feelings of fluency. At the end of the instruction phase of the study, participants were asked to predict their post-test scores. We measured the accuracy of these predictions by comparing predicted scores to actual post-test scores (accuracy = Predicted Score – Actual Score) (Figure 2D). There was no difference in the variance of accuracy between any of the groups (*T*(3)=0.894, *p* = 0.827, Bartlett's test). On average, score predictions were low by about one point out of ten in the Assessment + answers group, whereas they were quite accurate in the other groups. The Assessment + answers group was the only group that did not accurately predict their scores, as measured by a deviation from a mean accuracy of zero. It may be the case that participants who completed assessment with only assessment answers as feedback had lower feelings of self-efficacy than participants in the other groups, as manifested by lower ratings of understanding and inaccurate test score predictions. The results suggest that summary interactives and explanations can mitigate a perceived lack of familiarity with instructional material.

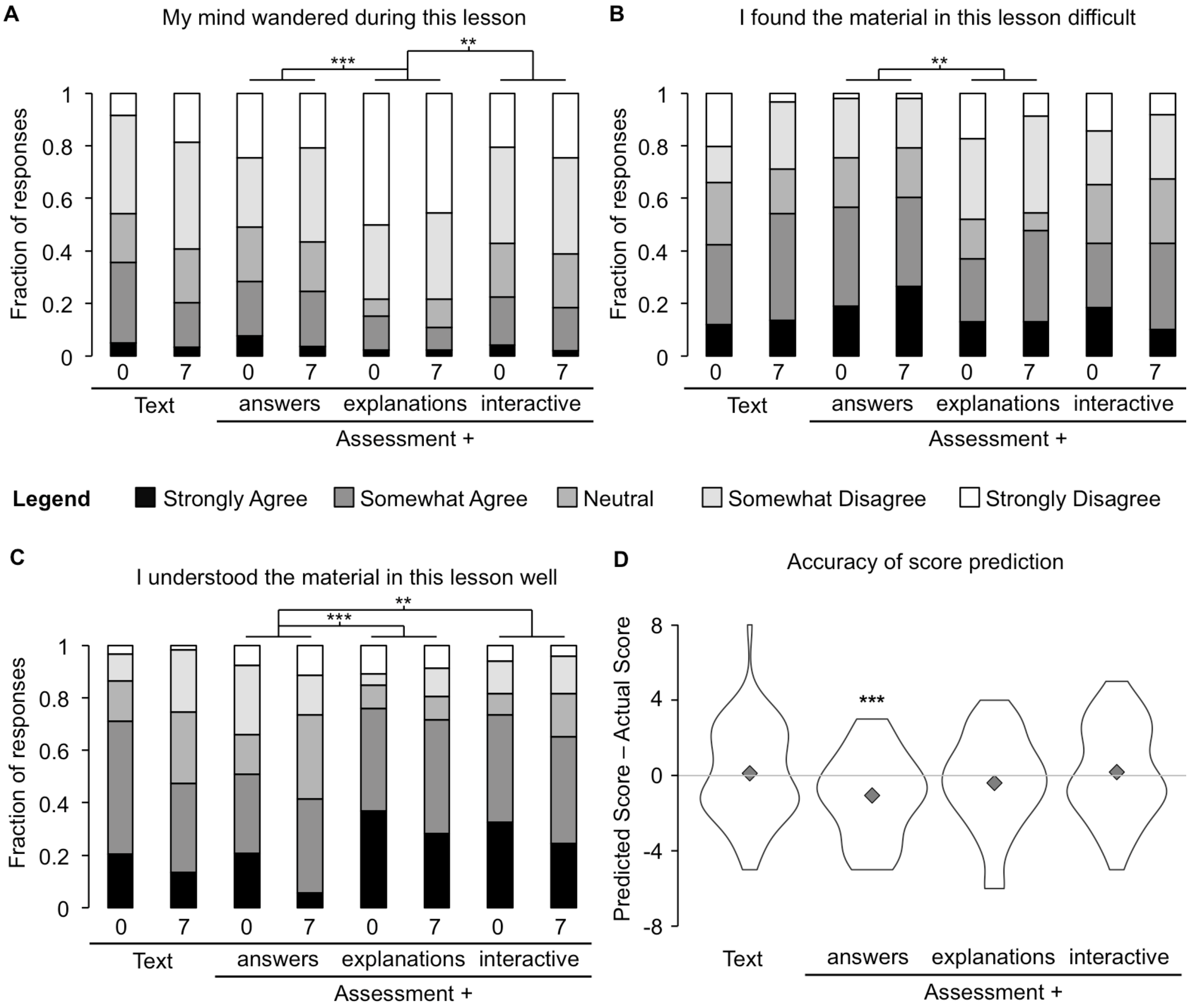


Figure 2. Significant differences in survey responses between the Assessment conditions.

Results from the Text condition are shown for visual comparison, but because of the significant differences in dropout, this condition was not statistically compared to the Assessment conditions. For panels A-C, all P values were computed with post-hoc Mann-Whitney tests. Sample sizes for each group are given in Appendix C. Frequency of responses the day of instruction (0) and after a seven-day follow-up (7) are shown. A – Participants in the Assessment + explanations condition reported substantially lower levels of mind wandering than participants in either the Assessment + answers (*U*=779, *p*=0.002) or Assessment + interactive (*U*=1488, *p*=0.006) conditions. B – Participants in the Assessment + explanations condition perceived that the lessons were less difficult than participants in the Assessment + answers condition (*U*=842.5, *p*=0.008). C – Compared to the Assessment + answers condition, participants reported greater understanding in the Assessment + explanations (*U*=1640, *p*=0.003) and Assessment + interactive (*U*=1701.5, *p*=0.006) conditions. D – Participants in the Assessment + answers condition made the least accurate score predictions; this group is the only group for which the average accuracy (Predicted – Actual Score) was not equal to zero (*t*(52)=-3.4612, *p*=0.0011). P values on graphs: ** < 0.01; *** < 0.005.

Evidence from an authentic context: Students choose to view assessment explanations in online courses. To test whether our experimental results hold true in a real-world setting, we analyzed assessment interaction behavior in three separate runs of HMX Fundamentals – Immunology, the SPOC from which we drew instructional materials for this study. The course platform provides students with a "Show Answer" button that appears after students have completed a question (meaning the students have answered correctly or used up all assessment attempts). Students were allowed two attempts for most assessments in the course. If students do answer correctly, they already have knowledge of the correct response and the only additional information revealed by the "Show Answer" functionality is an explanation. We analyzed whether students showed the answer and explanation, broken down by the number of answer attempts and whether or not their responses were correct (Figure 3). When students answered correctly on the first attempt, they viewed the explanation almost half of the time (48.9%). This increased if students were incorrect on the first attempt but correct on a later attempt (62.5%). Finally, if students never answered correctly, they almost always viewed the answer and explanation (95.0%). Overall, these results suggest that, even with knowledge of the correct response, students often choose to view assessment explanations.

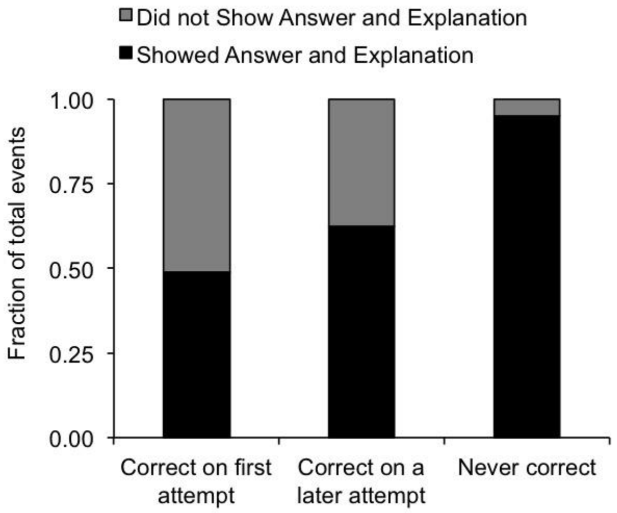


Figure 3. Students opt in to reading explanations in online courses.

Results are from three course runs of a small private online immunology course, comprising 69,043 total attempted questions. Students had the option to view answers and explanations with a “Show Answer” button. In any case where students answered correctly, they already had knowledge of the correct response before selecting “Show Answer”. Most questions in this course allowed the students to attempt the question twice. Approximately half of the time when students had the correct answer on a first attempt, they still viewed the explanation. In cases where the students were incorrect on the first attempt, but correct on a later attempt, students more frequently opted to see the explanation. Finally, when students were incorrect, they viewed the explanation (and correct answer) over 95% of the time. Differences in show answer behavior were statistically significant (*H(2)*=5540.2*, p*<2.2E-16, chi-squared test).

Predicting Post-Test Scores

Prior knowledge and reported confusion are predictive of test scores. Numerous studies have demonstrated a relationship between prior variables, including student demographics, and learning in traditional and online courses (Casillas et al., 2012; Geiser & Santelices, 2007; Pursel et al., 2016). We used ordinal logistic regression to test whether participants’ self-reported background information (gender, education, biology background, experience with online courses, and time spent on various educational activities) was predictive of post-test scores. These variables were not significantly predictive of post-test scores. We generated a separate ordinal logistic regression model to test whether participants’ pre-test scores or survey responses immediately after instruction were predictive of post-test scores. As a precaution, we excluded the results from the follow-up survey (administered after the post-test) because presumably a participant’s perception of her performance on the post-test could influence her survey responses. The results of this analysis are shown in Table 2. The only statistically significant predictors of post-test scores were participants’ pre-test scores and perceived confusion with the lesson. Specifically, higher pre-test scores and stronger disagreement with the statement “I found this lesson confusing” were predictive of higher post-test scores. Surprisingly, participants’ *predicted* post-test scores were *not* significantly predictive of their actual post-test scores in this model (although there was a positive correlation between these variables). It is less surprising that pre-test performance is one of the best predictors of eventual post-test performance, because the pre-test has the highest level of task similarity to the post-test. However, the finding that confusion is uniquely predictive of final test scores suggests that students' self-reported confusion may be more useful than other subjective ratings for self-evaluation purposes.

Table 2

Predicting Post-Test Scores with Ordinal Logistic Regression

Variable	Coefficient	*p* value	Likert data?
Pre-test score	0.581	0.004	N
I found this lesson confusing	0.496	0.011	Y
I found the lesson interesting	-0.397	0.051	Y
I enjoyed learning from this lesson	0.462	0.062	Y
My mind wandered during the lesson	-0.238	0.070	Y
Predicted score	0.111	0.160	N
I found this lesson difficult	0.191	0.220	Y
I understood the material in this lesson well	-0.237	0.270	Y
I would like to learn from more lessons like this one	-0.114	0.633	Y
I exerted effort in this lesson	-0.022	0.882	Y

**Note*. For Likert data, higher numbers indicate stronger disagreement, thus a positive coefficient indicates that disagreement with a statement was positively predictive of test scores.

Discussion

Practical Applications and Theoretical Implications

Our goal in this study was to test educationally-relevant instructional design choices that instructors and course creators must make when building online courses. From the current study, we can conclude some simple rules to inform the design of online courses. If an instructor chooses to include multiple choice or short answer formative assessments, she should add text explanations of the correct answers. This will improve students' subjective experiences and help to mitigate potential downsides of assessment. If she has the resources to add a summary interactive, this will also improve students' experiences. If instructors do make instructional changes based on these conclusions, students stand to directly benefit from more interesting and engaging online courses.

Simple Feedback Improves Automated Formative Assessment

Although the term "feedback" often connotes personalization and direct instructor involvement, the results of this work suggest that some of the benefits of feedback may accrue from universal feedback, such as written explanations of assessment questions. This feedback increases students' perceptions of engagement and understanding and reduces mind wandering relative to assessment with scoring only. In our experience, elaborative explanations of correct answers are inexpensive to rapidly produce, which suggests that adding answer explanations may be a particularly cost-effective means of improving online courses. Moreover, single explanations are much simpler to implement technically than the multiple different explanations that would be necessary for personalized feedback. According to the view that feedback can only exist in response to a student activity, explanations that provide information in addition to the accuracy of an answer are a way to provide *more feedback*. While instructors in traditional (face-to-face) courses often grade formative assessments, it is less common to provide explanations for all formative assessment questions. We hypothesize that assessment explanations would improve students' experiences in traditional courses.

Finding Better Measurements of Metacognition

There is some evidence that formative assessment improves the accuracy of score predictions on follow-up tests (Szpunar et al., 2014). In this work, we found that participants in the Assessment + answers condition made more inaccurate predictions than participants in other conditions, including a condition lacking any assessment (the Text condition). However, if the follow-up test were more difficult, the results could have been just the opposite. Almost any measurement of metacognitive accuracy based upon score predictions can suffer from the same problem. However, it is interesting to note that self-rated *confusion* with the material was uniquely predictive of low post-test scores, while other ratings, including score predictions, understanding, and difficulty, added no significant predictive value over pre-test scores and confusion. Paradoxically, this could indicate that the feeling of confusion is a stronger indicator of actual comprehension than feelings of understanding or difficulty, but this is an area that merits further investigation.

Limitations of this Study

Lack of a Testing Effect

We were surprised to find that the addition of assessments to a sequence of materials did not lead to an improvement in post-test scores. This may seem to contradict the testing effect, a widely replicated finding that testing of material increases retention of that material. It is important to note that in this study we controlled for some variables that could explain a lack of a testing effect. The test itself was written by one content expert and verified by two others outside of the study team, and we did find that a matched group of instruction-naive participants had much lower performance on the test than those who completed the instruction. If the test was not sensitive to learning of the instructional content, or participants did not learn at all, we would not observe such a difference.

In our view, our study differed from other studies that have demonstrated strong testing effects in several key ways that we summarize here. First, the testing effect is the strongest with repeated re-testing of study material. Repeated testing is thought to maximize the direct effects of testing (effects due to effortful retrieval of material). In this study, formative assessment was only delivered in one study session, so re-exposure to the material was minimal. Second, many studies utilize the same test items for study and for evaluation. While this may maximize the measurable effects of testing, it is also less educationally relevant, because the test items can be memorized and it requires no transfer of the material to a different context. In support of this, Gog and Sweller (2015) found that the testing effect disappears as the complexity of the material increases. In this study, we did not re-use assessment items, so no participant was exposed to an assessment item more than once. Finally, the direct effects of testing are likely related to the difficulty of effortful retrieval of study material. In this study, we deliberately utilized simple machine-gradable assessment types widely used in online courses, which are probably less challenging than free-response assessments. These differences may explain why the direct effects of testing were not observed in this study. There are also indirect effects of testing. In particular, testing helps students to identify areas for improvement and increases cognitive engagement during a study session. Testing may also serve to highlight the most important material. It may be the case that reading interleaved text statements (the Text condition in this study) activates some or all of the same indirect effects. Text statements could highlight the most relevant material or help students to evaluate their own understanding. It may be the case that, in a setting of computerized instruction, simply varying the multimedia modality increases engagement.

Use of Mechanical Turk for Study Recruitment

Amazon Mechanical Turk is increasingly utilized for education and psychology research studies, but concerns have been raised about sample representativeness and screening methods used in Mechanical Turk Studies (Paolacci & Chandler, 2014). We were careful to screen for participants with a baseline level of subject knowledge, reading comprehension, and test taking skills using a pre-test, and we did not exclude participants after the fact unless they had a technical problem. We believe that the direct financial motivations of the study participants may differ from the motivations of a learner who has enrolled in a program of study or online course. In this study, none of the participants in the text condition dropped out, but there was dropout in all of the assessment conditions. Moreover, time on task was significantly greater in the assessment conditions. It may be the case that participants made an economic choice with respect to how much time and effort they were willing to invest in exchange for study compensation. There is some evidence that forms of extrinsic reward can reduce task performance and undermine intrinsic motivation (reviewed by Hattie and Timperley, 2007). Theories of self-regulated learning posit that tangible rewards can abrogate an individual’s sense of responsibility for her own learning. Although motivation is a concern for the study population, it is important to note that we did collect strong evidence that the individuals who finished instruction did learn. Moreover, financial compensation is commonplace in laboratory studies of learning, so some of the same concerns about motivation probably hold true for other sample populations.

Conclusions

Our results support the concept that there are negative and positive impacts of formative assessment on students’ subjective experiences. In this study, formative assessment was directly linked to attrition - a negative outcome that instructors want to avoid. However, incorporating feedback into assessment, even simple written explanations, reduced mind wandering, enhanced the perception of understanding, and increased predicted performance. This aligns with our observational findings from analyzing student behavior in online courses – students often opt in to viewing assessment explanations, even when they already have knowledge of the correct answer. Adding explanations is a simple change that should enhance the subjective benefits of formative assessment. To maximize the direct effects of testing, instructors should consider utilizing more challenging recall-type testing tasks and increasing the repetition of formative assessment, although the potential learning benefits of each of these modifications must be weighed against potential impacts on students' experiences. It is important to note that open response (recall-type) assessment is less common in online courses in part because it is more difficult to grade, but there are spaced repetition testing applications that can be incorporated into many major learning management systems.

We also noted some challenges with using Amazon Mechanical Turk workers for education research. This study establishes a simple pre-screening procedure (a subject-aligned pre-test) that likely screens for multiple desirable traits and behaviors in study participants: reading comprehension, test-taking proficiency, attention, and subject-matter knowledge. We would recommend that other investigators use this approach in the future. Finally, we found that prior test performance and the feeling of confusion are predictors of future test performance. We hypothesize that the feeling of confusion may provide students with a more accurate metacognitive barometer than feelings of effort or understanding. This is a hypothesis that merits further exploration in the future.

Acknowledgements

This study was supported by a Spark grant from the Harvard Initiative for Learning and Teaching (HILT). We are grateful to Kevin Bonham and Bradley Coleman for their feedback on the study instruments. We would like to thank the HMX team, as well as Andrew Lichtman and Shiv Pillai for their contributions to instructional materials. Finally, we would like to thank Johanna Gutlerner and the HMS Curriculum Fellows Program for support and mentorship. This study was approved by the Harvard Medical School Institutional Review Board.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*(7), 861–876. doi: https://doi.org/10.1002/acp.1391

Cabeza, R., Kapur, S., Craik, F. I. M., McIntosh, A. R., Houle, S., & Tulving, E. (1997). Functional neuroanatomy of recall and recognition: A PET study of episodic memory. *Journal of Cognitive Neuroscience*, *9*(2), 254–265. doi: https://doi.org/10.1162/jocn.1997.9.2.254

Casillas, A., Robbins, S., Allen, J., Kuo, Y.-L., Hanson, M. A., & Schmeiser, C. (2012). Predicting early academic failure in high school from prior academic achievement, psychosocial characteristics, and behavior. *Journal of Educational Psychology*, *104*(2), 407–420. doi: https://doi.org/10.1037/a0027180

Chen, Z., Chudzicki, C., Palumbo, D., Alexandron, G., Choi, Y.-J., Zhou, Q., & Pritchard, D. E. (2016). Researching for better instructional methods using AB experiments in MOOCs: results and challenges. *Research and Practice in Technology Enhanced Learning*, *11*(1). doi: https://doi.org/10.1186/s41039-016-0034-4

Clark, R. C., & Mayer, R. E. (2011). *e-Learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*. San Francisco, CA: John Wiley & Sons.

D'Antoni, L., Kini, D., Alur, R., Gulwani, S., Viswanathan, M., & Hartmann, B. (2015). How can automatic feedback help students construct automata? *ACM Transactions on Computer-Human Interaction*, *22*(2). doi: https://doi.org/10.1145/2723163

Daradoumis, T., Bassi, R., Xhafa, F., & Caballé, S. (2013). A review on massive e-learning (MOOC) design, delivery and assessment. In *2013 Eighth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC)* (pp. 208–213). doi: https://doi.org/10.1109/3PGCIC.2013.37

Ebben, M., & Murphy, J. S. (2014). Unpacking MOOC scholarly discourse: A review of nascent MOOC scholarship. *Learning, Media and Technology*, *39*(3), 328–345. doi: https://doi.org/10.1080/17439884.2013.878352

Ferguson, R., & Clow, D. (2015). Examining engagement: Analysing learner subpopulations in massive open online courses (MOOCs). In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge* (pp. 51–58). New York, NY, USA: ACM. doi: https://doi.org/10.1145/2723576.2723606

Friedman, M., Liu, J., Morrissey, M., Turkay, S., & Wong, T. (2015). ChinaX course report [PDF]. Retrieved from http://harvardx.harvard.edu/files/harvardx/files/chinax_course_report.pdf

Geiser, S., & Santelices, M. V. (2007). Validity of high-school grades in predicting student success beyond the freshman year: high-school record vs. standardized tests as indicators of four-year college outcomes [PDF]. Retrieved from http://eric.ed.gov/?id=ED502858

Gog, T. van, & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review*, *27*(2), 247–264. doi: https://doi.org/10.1007/s10648-015-9310-x

Goldberg, L. R., Bell, E., King, C., O'Mara, C., McInerney, F., Robinson, A., & Vickers, J. (2015). Relationship between participants' level of education and engagement in their completion of the Understanding Dementia Massive Open Online Course. *BMC Medical Education*, *15*. doi: https://doi.org/10.1186/s12909-015-0344-z

Greene, J. A., Oswald, C. A., & Pomerantz, J. (2015). Predictors of retention and achievement in a massive open online course. *American Educational Research Journal, 52*(5), 925-955. doi: https://doi.org/10.3102/0002831215584621

Guo, P. J., Kim, J., & Rubin, R. (2014). How video production affects student engagement: An empirical study of MOOC videos. In *Proceedings of the First ACM Conference on Learning @ Scale*

Conference (pp. 41–50). New York, NY, USA: ACM. doi: https://doi.org/10.1145/2556325.2566239

Hansch, A., Hillers, L., McConachie, K., Newman, C., Schildhauer, T., & Schmidt, P. (2015). *Video and online learning: Critical reflections and findings from the field* (SSRN Scholarly Paper No. ID 2577882). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com.ezp-prod1.hul.harvard.edu/abstract=2577882

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. doi: https://doi.org/10.3102/003465430298487

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4-5), 528–558. doi: https://doi.org/10.1080/09541440601056620

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 170–179). New York, NY, USA: ACM. doi: https://doi.org/10.1145/2460296.2460330

Litman, L., Robinson, J., & Abberbock, T. (2016). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*(433), 1–10. doi: https://doi.org/10.3758/s13428-016-0727-z

Lütticke, R. (2004). Problem solving with adaptive feedback. In P. M. E. D. Bra & W. Nejdl (Eds.), *Adaptive hypermedia and adaptive web-based systems* (pp. 417–420). Berlin, , Springer-Verlag. Retrieved from http://link.springer.com.ezp-prod1.hul.harvard.edu/chapter/10.1007/978-3-540-27780-4_64

Margaryan, A., Bianco, M., & Littlejohn, A. (2015). Instructional quality of massive open online courses (MOOCs). *Computers & Education*, *80*, 77–83. doi: https://doi.org/10.1016/j.compedu.2014.08.005

Paolacci, G., & Chandler, J. (2014). Inside the Turk understanding mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188. doi: https://doi.org/10.1177/0963721414531598

Pursel, B. k., Zhang, L., Jablokow, K. W., Choi, G. W., & Velegol, D. (2016). Understanding MOOC students: Motivations and behaviours indicative of MOOC completion. *Journal of Computer Assisted Learning*, *32*(3), 202–217. doi: https://doi.org/10.1111/jcal.12131

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*(6002), 335–335. doi: https://doi.org/10.1126/science.1191465

Reich, J. (2014, December). MOOC completion and retention in the context of student intent. *Educause Review*. Retrieved from http://er.educause.edu/articles/2014/12/mooc-completion-and-retention-in-the-context-of-student-intent

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. doi: https://doi.org/10.1111/j.1745-6916.2006.00012.x

Saul, C., & Wuttke, H.-D. (2011). Feedback personalization as prerequisite for assessing higher-order thinking skills. *European Journal of Open, Distance and E-Learning*, *14*(2). Retrieved from http://www.eurodl.org/?p=special&sp=articles&inum=3&abstract=442&article=445

Shi, J., Wood, W. B., Martin, J. M., Guild, N. A., Vicens, Q., & Knight, J. K. (2010). A diagnostic assessment for introductory molecular and cell biology. *CBE Life Sciences Education*, *9*(4), 453–461. doi: https://doi.org/10.1187/cbe.10-04-0055

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*(3), 161–164. doi: https://doi.org/10.1016/j.jarmac.2014.02.001

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*, *110*(16), 6313–6317. doi: https://doi.org/10.1073/pnas.1221764110

Türkay, S. (2016). The effects of whiteboard animations on retention and subjective experiences when learning advanced physics topics. *Computers & Education*, *98*, 102–114. doi: https://doi.org/10.1016/j.compedu.2016.03.004

Wang, Y., & Baker, R. (2015). Content or platform: Why do students complete MOOCs? *Journal of Online Learning and Teaching*, *11*(1).

Appendix A

Study Materials

Pre-test

This pre-test was administered before the instructional phase of the study as a separate task. Only participants who answered three or more questions correctly were invited into the study.

------Pre-test text begins here-----

Question 1

This question was drawn from the IMCA exam, an introductory biology concept inventory (Shi et al., 2010). **Question 2** from the IMCA exam was used here.

Question 2

This question was drawn from the IMCA exam, an introductory biology concept inventory (Shi et al., 2010). **Question 9** from the IMCA exam was used here.

Question 3

This question was drawn from the IMCA exam, an introductory biology concept inventory (Shi et al., 2010). **Question 22** from the IMCA exam was used here.

Question 4
Endocytosis is best described as a process of cells
A releasing substances through holes in the cell membrane.
B taking up substances through holes in the cell membrane.
C releasing substances in vesicles.
D taking up substances in vesicles.

Answer: D

Question 5
Which of the following does NOT describe an important cellular function of proteins?
A catalysts of biochemical reactions
B information storage molecules
C structural components of cells
D signaling molecules

Answer: B

Reference for questions 1, 2, and 3:
Shi, J., Wood, W. B., Martin, J. M., Guild, N. A., Vicens, Q., & Knight, J. K. (2010). A diagnostic assessment for introductory molecular and cell biology. *CBE Life Sciences Education*, *9*(4), 453–461. http://doi.org/10.1187/cbe.10-04-0055

Demographic Questions

This survey was administered at the beginning of the instructional phase of the study.

------Survey text begins here-----

What is your gender?

❍ Male
❍ Female

What is the highest level of education you have completed?

❍ Did not complete high school
❍ High school or equivalent
❍ Some college
❍ Associate degree
❍ Bachelor's degree
❍ Master's degree
❍ Doctoral or professional degree

How many hours per week do you spend on the following activities?

______ Watching instructional videos
______ Listening to instructional podcasts
______ Reading instructional text online
______ Taking online courses

Have you ever taken an online course, such as a massive open online course (MOOC)?

❍ Yes
❍ No

How familiar are you with each of the following topics?

	Not at all	Slightly	Somewhat	Very	Extremely
Basic biology	❍	❍	❍	❍	❍
Advanced biology (such as biochemistry, cell biology, or molecular biology)	❍	❍	❍	❍	❍
Immunology	❍	❍	❍	❍	❍

Survey questions

This survey was administered at the end of the instructional phase of the study.

------Survey text begins here-----

Now please reflect on your experience taking this online lesson. Please answer each of these questions truthfully.

------New Survey Webpage-----

Please indicate your level of agreement with the following statements

	Strongly Agree	Somewhat Agree	Neutral	Somewhat Disagree	Strongly Disagree
I would like to learn from more lessons like this	❍	❍	❍	❍	❍
I enjoyed learning from this lesson	❍	❍	❍	❍	❍
I understood the material in this lesson well	❍	❍	❍	❍	❍
I found this lesson difficult	❍	❍	❍	❍	❍
My mind wandered during the lesson	❍	❍	❍	❍	❍
I found this lesson interesting	❍	❍	❍	❍	❍
I exerted a large amount of effort in this lesson	❍	❍	❍	❍	❍
I found this lesson confusing	❍	❍	❍	❍	❍

------New Survey Webpage-----

In one week, you will have an opportunity to take a 10-point multiple choice quiz. All of the questions on the quiz were addressed in this instructional material. Please predict your score out of 10.

❍ 0 (lowest score)
❍ 1
❍ 2
❍ 3
❍ 4
❍ 5
❍ 6
❍ 7
❍ 8
❍ 9
❍ 10 (highest score)

------New Survey Webpage-----

Did you experience any technical problems taking this study?
❍ Yes
❍ No

Please describe the technical problems in detail here.

------New Survey Webpage-----

Was there anything confusing about this study to you?
❍ Yes
❍ No

Please describe confusing aspects of the study here.

Post-test

This test was administered one week after the instructional phase of the study.

------Post-test text begins here-----

Question 1
Which of the cells below are tissue-resident sentinel cells? **(select two answers)**
A dendritic cells
B lymphocytes
C neutrophils
D monocytes
E mast cells

Answer: A, E

Question 2
A ____________ is a cell that uses innate immune receptors to recognize and phagocytose microbes; these cells have a short life span within tissue and often rapidly die by apoptosis.
A dendritic cell
B macrophage
C neutrophil
D monocyte
E mast cell

Answer: C

Question 3
A ____________ is a cell that uses innate immune receptors to recognize and phagocytose microbes. It also will phagocytose and digest apoptotic cells.

A macrophage
B lymphocyte
C monocyte
D mast cell

Answer: A

Question 4
Which of the cells below are circulating blood cells that will migrate into tissue in response to inflammation? **(select two answers)**
A red blood cells
B dendritic cells
C neutrophils
D monoctyes
E mast cells

Answer: C,D

Question 5
Pro-inflammatory cytokines and mediators bind to receptors on ____________ cells, which respond by undergoing changes that will promote the recruitment of circulating leukocytes from the blood into the tissue.
A endothelial cells
B red blood cells
C epithelial cells
D macrophages

Answer: A

Question 6
Leukocyte adhesion deficiency (LAD) is a genetic defect that leads to recurrent infections in the tissue and severe problems with wound healing. LAD patients also develop gingivitis (infections and inflammation of the gums). In LAD, leukocyte migration into tissues is severely impaired. All of these problems can be traced back to a genetic defect. Of the genetic defects listed below, which is the most likely cause of LAD?
A A mutation that impacts blood cell development, leading to below-normal numbers of monocytes, but normal numbers of other blood cells.
B A mutation that introduces a stop codon into a gene that encodes part of the LFA-1 molecule (leading to a truncated protein).
C A mutation that increases the stability of the E-selectin ligand protein without affecting its other functions.
D A mutation that leads to high pro-inflammatory cytokine expression in the tissue.

Answer: B

Question 7
Some of the steps of an acute inflammatory response are listed below. Which of these steps would occur **first** in a given episode of inflammation?
A Tissue-resident sentinel cells release inflammatory mediators.
B Microbial molecules bind to innate immune receptors.
C Endothelial adhesion molecule expression increases.
D Circulating neutrophils migrate into the tissue.

Answer: B

Question 8
Psoriasis is an inflammatory disease that impacts the skin. It most commonly manifests as scaly, raised, red or white areas on the skin caused by local inflammation. The inflammation leads to overgrowth of skin cells called keratinocytes. The triggers that lead to psoriasis are largely unknown, but microscopic examination of skin biopsies from psoriasis patients reveals massive infiltration of leukocytes into the tissue with no evidence of bacterial or viral infection.

Which drug or therapy would you expect to be **LEAST** effective in treating psoriasis?

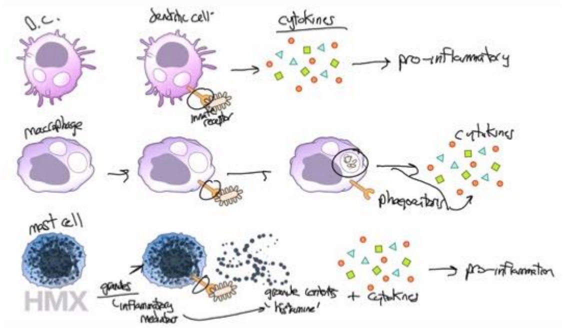
A A treatment that blocks the release of pro-inflammatory cytokines.
B A treatment that blocks the removal of apoptotic neutrophils.
C A treatment that prevents the binding of LFA-1 to ICAM-1.
D A treatment that kills leukocytes that migrate into tissue.

Answer: B

Appendix B

Representative Image of a “Whiteboard” Video

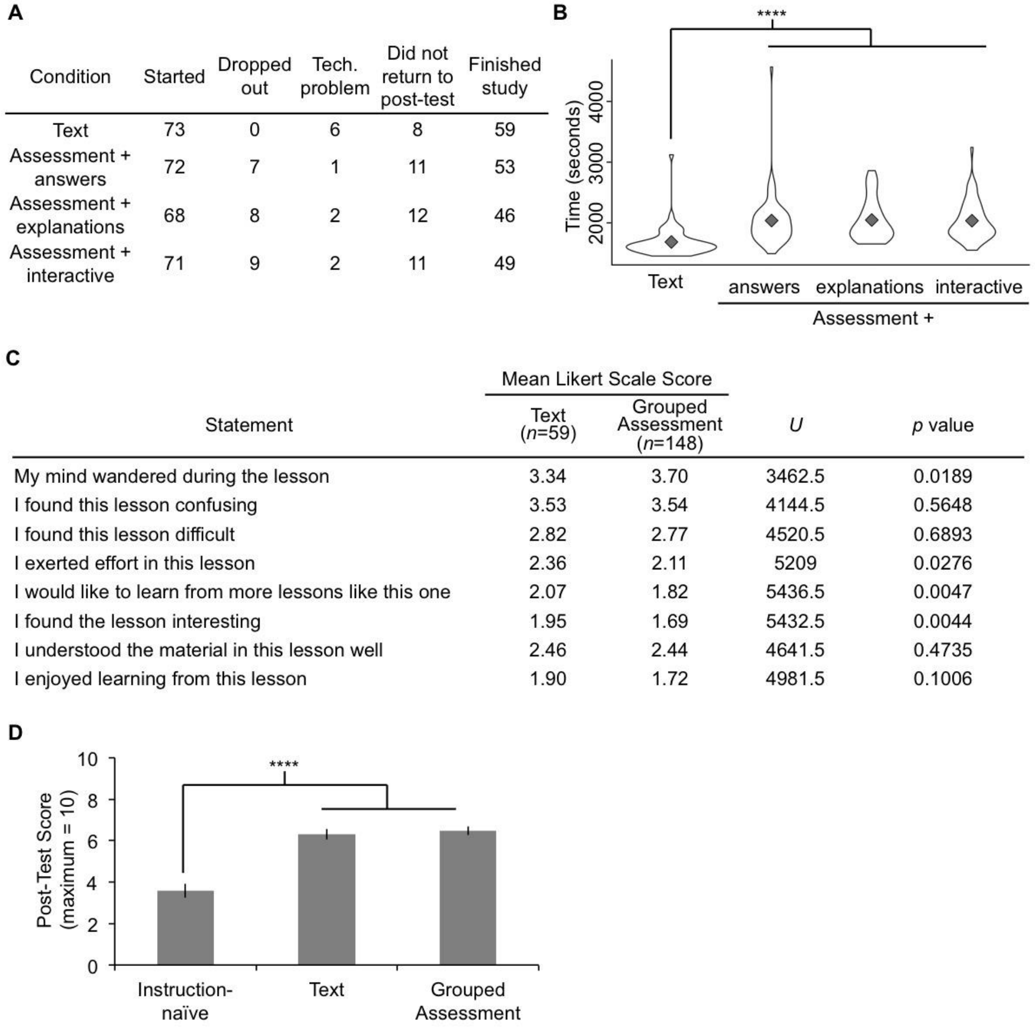
A representative screen-shot from a “whiteboard-style” teaching video that was used in the instructional phase of this study is shown here.



Appendix C

Differences in Dropout and Survey Responses Between Text and Assessment Conditions

A – Differences in attrition between the different instruction conditions. “Started” refers to participants who entered instruction, “Dropped out” refers to participants who dropped out of instruction, “Tech. problem” refers to participants who reported a technical problem, “Did not return to post-test” refers to participants who did not take the 1-week follow-up test, and “Finished study” refers to participants who completed all phases of the study and were included in this analysis. There was a significant difference between groups in dropout during instruction, and post-hoc tests showed a difference between the Text and Assessment conditions, but no difference between the different Assessment conditions (see text for significance tests). **B** – Differences in time on task between the different instruction conditions. **C** – Summary of differences between Text and Grouped Assessment conditions. Mean Likert Score results are reported (1 – Strongly Agree; 5 – Strongly Disagree). Due to differences in dropout, it is not possible to conclude that the more favorable ratings are due to better overall experiences in the assessment conditions. **D** – No significant difference in post-test scores between the groups, although all groups had much higher scores than an instruction-naïve group. P values: **** < 0.001.



A

Condition	Started	Dropped out	Tech. problem	Did not return to post-test	Finished study
Text	73	0	6	8	59
Assessment + answers	72	7	1	11	53
Assessment + explanations	68	8	2	12	46
Assessment + interactive	71	9	2	11	49

C

Statement	Mean Likert Scale Score: Text (*n*=59)	Mean Likert Scale Score: Grouped Assessment (*n*=148)	*U*	*p* value
My mind wandered during the lesson	3.34	3.70	3462.5	0.0189
I found this lesson confusing	3.53	3.54	4144.5	0.5648
I found this lesson difficult	2.82	2.77	4520.5	0.6893
I exerted effort in this lesson	2.36	2.11	5209	0.0276
I would like to learn from more lessons like this one	2.07	1.82	5436.5	0.0047
I found the lesson interesting	1.95	1.69	5432.5	0.0044
I understood the material in this lesson well	2.46	2.44	4641.5	0.4735
I enjoyed learning from this lesson	1.90	1.72	4981.5	0.1006