Student-generated pre-exam questions is an effective tool for participatory learning: A case study from Ecology of Waterborne Pathogens course.

Max Teplitski[1], Tracy Irani[3], Cory J. Krediet [1,4], Mariachiara Di Cesare[2], Massimiliano Marvasi [1,2*]

[1] Soil and Water Science Department, University of Florida, Gainesville, Florida, United States of America

[2] Department of Natural Science, Middlesex University, London, United Kingdom.

[3] Department of Family, Youth and Community Sciences, University of Florida, Gainesville, Florida, United States of America

[4] Department of Marine Science, Eckerd College, St. Petersburg, Florida, United States of America

* Corresponding author

E-mail: m.marvasi@mdx.ac.uk

Short title: Learning gains from question design

Conflict of interests: Authors declare not conflict of interests.

Abstract

This multi-year study helps elucidate how the instructional practice of student-generated questions support learning in a blended classroom in STEM subjects. Students designed multiple-choice pre-exam questions aimed at higher levels of learning, according to Bloom's taxonomy. Student-generated questions were edited by the instructor and then discussed by the students in the classroom and in an online forum. We tested the hypothesis that this intervention improves student learning, measured as student achievement on the exam following the intervention, and compared to student achievement on the traditional exam (prior to which a review session focused on instructor-led recitation of the key concepts). Following the intervention in all years, average grade on the post-intervention exam increased by 7.44%. It is important to point out that not all students benefited equally from this activity. Students who were in the 4th quintile (60-80%) based on the results of the first exam demonstrated the highest achievement improving their performance on average by 12.37% percentage points (measured as a score on the second exam). Gains were not observed in the semesters when the intervention was not implemented. In this study we provided students detailed instructions on how to design questions that focus on testing higher levels of learning.

Keywords: undergraduate, collaborative/cooperative learning, assessment, student-centered learning, quizzing

Introduction

In recent years, it has become clear that student-centered and inquiry-driven approaches to learning are successful strategies to significantly obtain improved learning gain in science, technology, engineering and mathematics (STEM) subjects (Knight and Wood 2005; Wood and Knight 2004). This realization has led to the increasing popularity of more interactive teaching approaches, among which “flipped classroom” is one of the available tools. Flipped classroom is an educational approach in which students spend a considerable amount of out-of-class time watching pre-recorded lectures, completing reading assignments and being engaged in various web-based asynchronous activities. In a flipped classroom, face-to-face time is used for individual inquiry and collaborative activities to clarify concepts and contextualize knowledge through application, analysis, planning and researching solutions. During this time, instructors act as learning coaches rather than teachers (O'Flaherty and Phillips 2015). A number of instructors have also adopted a “blended” approach, where elements of the flipped classroom are combined with the more traditional expository educational approaches (Morin 2014; O'Flaherty and Phillips 2015).

Flipped classrooms have been shown to be effective in promoting learning gains, and students generally feel positive about their experiences (Morin 2014; O'Flaherty and Phillips 2015). However, despite their benefits, flipped and also blended classrooms have some logistical drawbacks. Because meaningful pre-class activities are a must (Morin 2014), these approaches require significant investment of the instructor’s time. McLaughlin and others (2014) estimated that instructors spend approximately 30% more of their time to prepare for flipped classes compared to traditional lectures (McLaughlin and others 2014). Instructors need additional time to tape, edit and update pre-recorded lectures, develop asynchronous activities that take place outside the classroom and plan face-to-face exercises (O'Flaherty and Phillips 2015; Spangler

2014). Preparation of quiz test banks is also a major time investment by the instructor (O'Flaherty and Phillips 2015). Several studies identified student-led generation of question banks as an opportunity not only to save the instructor's time and institutional resources, but also to engage students more deeply into the learning process (Bottomley and Denny 2011; Hardy and others 2014; Rhind and Pettigrew 2012). With this study, we aimed to test the effectiveness of student-generated multiple-choice question banks in enhancing learning in a blended classroom.

It is commonly recognized that the act of creating questions enhances learners' understanding of course materials and promotes deep learning (Draper 2009; Rosenshine and others 1996). Students who wrote questions tended to outperform their peers (Foos 1989). Student-generated questions that involved higher cognitive skills (compared to a simple recall) have been linked to self-directed learning and improved conceptual understanding (Chin and others 2002). However, studies also point to the fact that for this activity to be productive, it has to be properly contextualized (Bottomley and Denny 2011; Hardy and others 2014; Nicol 2007; Nicol and Macfarlane-Dick 2006; Rhind and Pettigrew 2012). For example, in the experiments of Rhind et al (2012) conducted in three courses of a veterinary curriculum over a two-year period, students were directed to develop questions that were submitted into an on-line bank without being evaluated (Rhind and Pettigrew 2012). In only 1 of 6 experiments, the correlation between answering questions submitted by peers and the exam grade was significant ($p<0.05$), despite the fact that over 80% of students agreed that developing original questions and answering questions developed by others helped improve understanding of the material. Students in this study submitted and answered only two multiple-choice questions, and the quality of the student-generated questions was not evaluated. Hardy and others (2014) in a three-school UK study encompassing courses in Physics, Genetics and Chemistry reported only modest, but nevertheless positive, impact of writing and reviewing exam questions on the end-of-term performance (Hardy and others 2014). As was in the study of Rhind (2012), participants in the Hardy (2014) study submitted a limited number of un-moderated questions (between 1 and 2), however, students were required to answer or evaluate between 2 and 20 questions submitted by their peers. Hardy and collaborators (2014) hypothesized that monitoring the quality of the questions that students submit may have improved learning gains. Bottomley and others, 2011 reported that when over one hundred biomedical students were asked to develop own multiple choice questions, over 90% of the questions represented the lowest two of Bloom's taxonomy (Bottomley and Denny 2011). However, if students were specifically instructed to write multiple choice questions representing higher levels of Bloom's taxonomy, they did so (Bates and others 2014). Thus, it appears, that for this exercise to result in learning gains, students have to engage at multiple higher level tasks, such as creation of the more sophisticated questions, engagement with questions developed by their peers and obtaining timely feedback.

It is also reasonable to hypothesize that learning gains resulting from the construction of exam questions could be due to the ability of students to recall significantly more information on the final test if they were subjected to intermediate self-administered tests prior to the final examination (Glover 1989). Both the types of the tests and recall activities have an impact on learning. For information that is not complex (i.e. memorization and recall of word pairs in English and in a foreign language) learning gains are most improved with a repeated retrieval practice (Karpicke and Roediger 2008). For complex information, where the ability to recall certain information requires memorization (or thorough understanding) of the context, retrieval has to modify the information in memory, rather than storing it unchanged (Bjork and others 2015). Furthermore, it involves either an elaboration of the memorized material (Mcdaniel and Fisher 1991) or requires a certain level of "desirable difficulty" (a condition that initially slows down knowledge acquisition, but improves long-term retention and transfer of knowledge), which can be achieved when the recall activities are reasonably complicated (Bjork and others

2015). Therefore, the aim of this study was to test whether student's generation of multiple-choice questions could result in robust learning gains that are similar to those observed in studies of the testing effect and without the undesirable effects of anxiety and biases observed in previous studies (Bjork and others 2015; Kromann and others 2011; Nguyen and McDaniel 2015) This activity required not only a simple recall of information, but a thorough understanding of the tested material and additional information.

In addition to asking students to develop and submit questions on-line, we wanted to further capitalize on this learning opportunity and instructed students to use questions that were developed by their peers and posted within an e-learning platform for self-paced quizzes. The most difficult questions were discussed during two one-hour class periods. Self-paced and un-graded pop quizzes were shown to be effective learning tools, which confer the benefits of the activity without the negative impact of stress and anxiety (Khanna 2015).

Methods

Ethics Statement

All survey questions and the study design were reviewed by the University of Florida Institutional Review Board. All study participants gave written consent to take part in the study. Consent letter was drafted by T.I. (who was not the course instructor), and were administered by a third party. Whether or not students gave their consent to participate in the study was not revealed to the course instructors until after the final grade for the course was finalized and submitted. University of Florida Institutional Review Board reviewed the consent procedure.

Participants

This experiment was conducted in the course "Ecology of Waterborne Pathogens", which is taught once a year. This course is an approved elective for Biology, Microbiology, Soil and Water Science majors. Over the duration of the experiment, 162 students chose to participate in this study from 2009 to 2015 (of them, ~55% where female and ~45% male). Development of the pre-exam questions for the bank and performance on exams were graded activities in the course. Two control cohorts, enrolled in the course in 2009 the year prior to the intervention consisted of 32 students, and the 2015 cohort included 30 students.

Course Design

The course included traditional lectures, in-class discussions, pre-recorded on-line instructional videos, weekly on-line discussions, a nine-week long investigation of a simulated outbreak and an activity in which students constructed multiple-choice questions. Approximately two weeks prior to the first class, the instructor emailed students a list of suggested course topics, and students ranked them according to their interest in the subject. Based on the results of the students' votes, more or less time was allocated to certain topics. Some topics (food safety, urban microbiology and diseases of aquatic animals) were not discussed in all years, however, these topics were typically covered in the third module of the course (after the second exam). The first module typically included surveys of outbreaks of water- and foodborne illnesses, environmentally-transmitted pathogens (including emerging and re-emerging pathogens), sample collection and preparation, culture-based methods for pathogen isolation and detection, physiological, immunological and nucleic acid-based detection and characterization, source-tracking of human pathogens, a survey of environments where pathogens survive (soil, air, water, plant-associated environments). Upon completion of the first module, students completed the first exam. During the second module, the following topics were typically covered: biofilms and microbial cell-to-cell signaling, evolution of antibiotic resistance and virulence, pathogens as prey, microbiological components of water quality, indicator organisms, drinking and waste water treatment and disinfection. At the end of the second module students completed the second exam (Figure 1).

Study Design

As mentioned above the course was divided into three modules. The assessment structure is shown in Figure 1. From 2010 to 2012 the intervention was implemented before Exam 2. In 2015, an additional control was introduced: the intervention was implemented before a third exam instead of the Exam 2. For the intervention, students were instructed to design 20 multiple-choice questions. Prior to the assignment, a 15-minute presentation about Bloom's taxonomy of learning was offered in class by one of the instructors. Students also received a handout summary of the presentation, and examples of questions that would not be accepted for credit. Students were instructed that less than 20% of their questions could represent the lower two domains of Bloom's taxonomy and that True/False questions (even when disguised as multiple-choice questions such as: "Which of the following four statements is correct...") would not be accepted for credit. Students were also instructed that questions based on a single power-point slide from lectures would not be accepted for credit, thus forcing them to tap into higher levels of learning. Questions had to be scientifically accurate and grammatically correct, and stems of questions had to be based directly on the topics that were discussed during the second module, although integration of the topics from the first module was also encouraged (see, for example, Questions 1, 6 and 8 in the Table 1). These criteria were used for grading the questions that students submitted. All questions, without the answers, were posted by students in the e-learning system to be accessible to all students. Through the duration of this study, the institution used two different e-learning platforms (Sakai and Canvas). However, because functionalities of the two platforms were similar, this was not considered as a variable that impacted the outcome of the study. Students were offered an opportunity to send some or all of their questions for the instructor's formative feedback prior to posting them (and approximately a third of students took advantage of that opportunity). The instructors provided feedback and graded the posted questions prior to the exam. As the questions were submitted in the e-learning system without the answers, all students were encouraged to use them for self-paced tests (not graded by the instructors), and approximately 80 of the most difficult questions were discussed during two class periods prior to the exam, however review was offered in all sections. None of the questions that were discussed during the review session were included in the actual exam.

Exam and Assessment

All questions from Exam 1 were designed by the instructor. Exam 1 consisted of multiple-choice questions and a list of five questions for short essays from which students selected three to answer. Multiple-choice questions accounted for ~70% of grade, while short essays made up the rest. Approximately 80% of the multiple-choice questions on the first exam were the same for the three semesters in which the intervention took place; there was only ~50% overlap between questions on the first exam in the semesters when the intervention took place and in the two semesters prior to the intervention.

The second exam consisted of 20 questions, ~75% of them were drawn from over 600 questions submitted by the students to the question bank during that semester, and ~25% of the questions were designed by the instructor. Questions developed by students were edited by the instructor for clarity and grammar. Multiple-choice questions offered by the instructors in the semesters when the intervention took place were essentially the same. In the year prior to the intervention, all questions were designed by the course instructors. To accommodate a variety of learning styles, students had a choice to complete multiple-choice or short essay questions on the second exam; students self-selected which version of the exam to take. Essay exams included a “menu” of 15-20 questions, from which students selected 10-12 to answer. Each question required a clear understanding of at least 2 different concepts discussed throughout the course. For example, “We have discussed several examples of water quality indicators. Can indicator organisms be used in programs aimed at controlling accidental or opportunistic waterborne pathogens (such as *Aeromonas, Vibrio cholera, Vibrio vulnificus*)? Why or why not?”; “How would you use multiplex PCR to identify subspecies of *Salmonella*? Design primers for this experiment, include picture of the gel that you expect will result from this PCR experiment. Make sure to include all appropriate positive and negative controls. In addition to multiplex PCR targeting genes you identified as important, what 3 other genetic techniques could be used to distinguish between subspecies of *Salmonella*? Why? What would the data look like (provide examples of gels)?”, “Resistance to antibiotics is often associated with acquisition of the new genes and functions. Please provide 3 hypothetical scenarios in which loss of a gene or a mutation of a gene within the core genome would increase resistance of a pathogen to antimicrobials.” Mastery of the same concepts was required to complete multiple choice (MC) exams. Over the duration of this study, approximately 40% of students chose to take the essay version of the exam. All exams were graded and moderated by the two instructors in the course to ensure consistency (with the exception of one semester prior to the intervention). Essays were graded based on whether students demonstrated understanding of individual concepts, ability to integrate them and to provide clear, concise and accurate answers. When achievements of students who took the essay or multiple-choice exam were compared, no statistically significant differences were observed. Students’ achievement on the second exam was compared with their achievement on the first exam. As a control, there was a formative pre-exam review session prior to the first exam, in which the instructor led an in-class discussion of the key concepts covered during the first module.

Coding of student questions according to Blooms’s taxonomy. Pre-exam questions developed by students were considered to be recall-type questions if they required a restatement of a specific fact or a definition and were based on a single power point slide;

comprehension questions required basic understanding of at least two different concepts, and were often exemplified by simple analogies; application questions involved basic case study analyses, data interpretation or calculations; analysis and synthesis questions required original analysis of complex data.

Activity satisfaction questionnaire. Upon completion of the semester, we administered a questionnaire, which aimed to determine which factors (such as clarity of the assignments, recognition of students' contributions, appropriateness of guidance by the course instructors, relevance of the assignment) correlated with the overall satisfaction with the activity. Students answered questions by assigning either 1 (strongly disagree) or 5 (strongly agree) to each question. Participation in the questionnaire at the end of the course was voluntary, and students received extra credit points for filling it out. Eighty seven students completed this post-activity questionnaire consisted of 11 questions. In order to measure factors associated with the student satisfaction with the intervention, linear regressions of each question correlated with the overall satisfaction was performed. Statistical analysis was carried out with JMP (SAS) software package.

Statistical analysis to assess the impact of the pre-exam activity on student achievement. Individual student grades were converted into percentages and three tests were performed: 1) A one-way ANOVA across cohorts was inferred to determine differences in mean grades on Exam 1, the pre-exam and Exam 2, and in changes between Exams 1 and 2. 2) A quintile regression analysis to identify differential effects of the pre-test and Exam 1 performance among different groups of grades in Exam 2. In other worlds, the quintile analysis allowed us to identify to what extent students grouping into different grades (0-20%, 20-40%, 40-60%, 60-80% and 80-100%) were affected by the implementation of the activity. To test for differences in the performance of cohorts in Experiment 1 on each exam (Exam 1, pre-exam, and Exam 2) a one-way ANOVA was performed on the mean of the grades and on the differences between grades on Exam 2 and grades on Exam 1. To check for the effects of the pre-test among different groups of grades, a quintile regression has been performed including both pre-test and grades on the Exam 2. Statistical analyses were carried out in R and JMP (SAS) software package. While preliminary analyses of the data were carried out by the course instructors, final analyses presented in this paper were conducted by the individuals who were not associated with the study design or course instruction.

Results

Quality of student-generated pre-exam questions. Prior to completing the assignment, students were introduced to Bloom's Taxonomy of Learning and were given specific instructions on what pre-exam questions would be accepted for full credit. Only approximately 15% of the questions submitted by students were simple recall-type questions. The majority of the questions engaged higher levels of learning. Approximately 35% of the questions were scored as comprehension-type questions, 28% were scored as application, 20% were analysis and the rest were synthesis-type questions, according to Bloom's Taxonomy. For example, Q1 (Table 1) required that students were able to recall definition of "fecal coliform", analyze composition of the xylose lysine deoxycholate (XLD) medium and deduce how fecal coliforms would behave when plated on XLD agar. Construction of Q2 and Q3 required comprehension of the topics discussed under Fundamentals of Microbial Evolution, using examples from the instructor's presentations on common food- and waterborne pathogens as well as indicators of water quality. To design Q4, students had to recall that *S. bongori* carries *Salmonella* Pathogenicity Island 1 (discussed in Evolution of Enteric Pathogens) and synthesize it with the information discussed under Virulence Mechanisms. Virulence of *S. bongori* was not an explicit topic that was discussed in class.

Because majority of students in this course were on the pre-med track, they used this as an opportunity to apply information they learned throughout the course to what was discussed in this course and in other classes they have taken (see, for example Q5 and Q6). Construction of these questions required comprehension of the mechanisms of action of antibiotics (discussed in this course under Antibiotics: Mechanisms of Action and Resistance) as well as the instructor's presentation Pathogens As Prey and a number of concepts discussed under Quorum Sensing, and a presentation on Detection of Waterborne Pathogens. Students were specifically instructed that stems of questions had to be based on the material covered in the second module of the course, and questions that could have been answered without taking the course were not accepted for credit.

We note that analogies represented a significant number of questions designed by students (over 22% overall). Analogy questions designed by students clearly represented higher levels of learning (for example, Q7 and Q8). Of these two questions, Q7 was not accepted for full credit because it was based on the same power point slide in the instructor's presentation on Quorum Sensing. Q8 demonstrated synthesis of topics discussed under Pathogens As Prey and introductory presentations on the life cycle of waterborne human pathogens.

Impact of the pre-exam activity on student achievement.

The first hypothesis tested with this study was that student achievement on an exam would be improved by constructing pre-exam questions before Exam 2. Student achievement on this pre-exam assignment (construction of exam questions) averaged 89.56±14.16%, 98.24±4.12%, and 96.60±7.37%, for 2010, 2011 and 2012 respectively (Table 2). No significant differences in student achievement on the pre-exam were observed over the three years ($F = 2.12$, $p=0.125$).

Following the intervention from 2010 to 2012, average grade on the post-intervention exam increased by 7.44%.

On Exam 1, which was administered following an in-class discussion of topics covered prior to the exam, but without the intervention, the four-year average was 78.70% (Table 2, Figure 2A). No differences in students' performance were observed for Exam 1 ($F=0.67$, $p=0.575$) in the 4 years of observations (Figure 2 and Table 2).

The three-year average on the second exam, administered following the intervention, which involved development, submission and discussion of the student-generated questions, was 86.58±12.86% (Figure 2 B). There were no statistically significant differences in student achievement on the second exams over the three years when the intervention was implemented (86.46±10.38%, 86.42±11.66%, and 86.86±16.56% in 2010, 2011 and 2012, respectively). Student achievement on the same exam in 2009 (when students did not submit pre-exam questions prior to the second exam) was significantly lower than in the years when the intervention was implemented (Figure 2 B).

When comparing students' performance on Exam 2, the average grade in 2010-2012 (when the pre-exam was carried out) was statistically higher than the average grade in 2009 (when no pre-exam was performed) ($F=4.84$, $p=0.003$) (Figure 2, B). The change in performance between Exam 1 and Exam 2 (see Exam2-Exam1 column, Table 2) also shows a significantly greater improvement for the 2010-2012 cohorts than for the 2009 cohort ($F=4.90$, $p=0.003$).

An additional control experiment was carried out in 2015. Students were instructed to design multiple choice questions as in the Experiment 1 prior to the third (final, and not the second) exam. This was done to address the possibility that the differences in learning gains observed on the second exam during previous years were due not to an intervention but due to the inherent relative "easiness" of the material covered during the second module or due to some other factors that may be associated with the particular timeframe during the spring semester. The grades for the second exam were used as a control for all the years, in which the students were asked to write the pre-exam questions before the second exam. When grades from Exams 2 where compared along the years, the two years (2009 and 2015) in which the pre-exam was not implemented were significantly lower (p=0.001) when compared with the years in which the intervention was implemented (Figure 2 B), thus indicating that learning gains on the second exam were most likely due to the intervention, and not to other factors.

Despite the fact that the intervention (construction and discussion of the pre-exam questions) appears to have been associated with a significant increase in student learning gains, there was no correlation between points earned for the design of the pre-exam questions and their grade on the second exam (p=0.198).

Quintile regression was also performed in order to measure to what extent the implementation of the pre-exam affected each quintile of the grades (0 to 100%). In the quintile regression for each unit, an increase in student performance on the second exam varied by quintile. Interestingly when the pre-test was used, a significant improvement in student achievement on Exam 2 was observed.

It is important to point out that not all students benefited equally from this activity (Table 3). Based on the student achievement on the first exam, ninety-nine students participating in the study were divided into five quintiles to track their improvement. Following the intervention, six out of nine students in the second quintile (20-40%) improved their scores on the second exam on an average of 6.88% percentage points. Students who were in the fourth quintile based on the results of the first exam demonstrated the highest achievement (measured as a score on the second exam). Students in the 4^{th} quintile (60% to 80% of correct answers), twenty-nine out of thirty-eight students improved their performance on an average by 12.37% percentage points. Finally, of the students who were within the top (fifth) quintile (from 80% to 100% of correct answers on the first exam), thirty-one students out of fifty-two improved their performance by an average of 10.88% percentage points (Table 3).

Factors associated with the student satisfaction with the intervention. In order to assess the impacts of the intervention (construction and discussion of the multiple-choice pre-exam questions), students were surveyed on how they felt their contributions were received. Students were given a questionnaire consisting of 11 questions (Table 4) upon completion of the course. We wanted to determine whether students felt that the assignment was clear, challenging and relevant, whether they have received the appropriate level of guidance, whether their contribution was properly recognized. Overall, study participants felt that construction of the pre-exam questions was a rewarding experience (3.9/5.0, Table 4), despite the fact that they were generally unsure that this activity would be relevant in their future careers (Table 4).

Whether or not rules of the assignment were clear, appropriateness of instructor's guidance, proportionality of earned points to the investment of energy, relevance of the experience to the future career, recognition of the student's investment by the instructor all had a statistically significant correlation with the overall satisfaction with the assignment (Table 4). Nevertheless, the r^2 values for all of these correlations were low, with the relevance to the future career, instructor recognition and the proportionality of the investment to the overall grade having the strongest impact on the overall satisfaction with the assignment (Table 4).

Discussion

As instructors retreat from relying on lectures as a main instructional tool to adopt more participatory strategies, student-centered educational models are becoming more wide-spread. In a classroom, whether traditional or flipped, testing remains an important component of the educational process and institutional assessment. Beneficial effects of testing on long-term retention of knowledge have been well documented (Brame and Biel 2015; Jacoby and others 2010; Kang and others 2007; Kromann and others 2011; McDaniel and others 2007; van Gog and Sweller 2015). Learning gains are observed even when students attempt to answer questions, but either fail to answer them correctly or do not receive timely feedback (Kornell 2014; Richland and others 2009). It is hypothesized that attempting to answer questions activates cognitive networks, potentially allowing retrieval of related content and identifying the need for related information (Richland and others 2009).

Exams and quizzes are stress- and anxiety-inducing activities (Nguyen and McDaniel 2015) and references therein), and benefits of test-enhanced learning maybe gender-biased (Kromann and others 2011). Furthermore, poorly constructed multiple-choice pre-exams with implausible alternative answers do not result in meaningful learning gains (Bjork and others 2015). Frequent quizzes may also be seen by students as an attempt to assert (or even usurp) authority by the instructor, and thus undermine development of the student-centered participatory classroom. To reduce the levels of stress and anxiety associated with exam-taking, instructors have experimented with ungraded pop quizzes (Khanna 2015), peer-graded quizzes (Coppola and Pontrello 2014) and also “flipped exams”, in which students work collaboratively to solve exam questions (Lujan and DiCarlo 2014). Clearly, engaging students into an active learning process can take different forms.

When the benefits of testing were further dissected, several hypotheses (e.g., elaborative retrieval and transfer-appropriate processing theories, retrieval induced facilitation, unspecific-goal perspective) attempted to explain its cognitive benefits (Carpenter 2009; Carpenter and

DeLosh 2006). According to the elaborative retrieval hypothesis, activation of the elaborative semantic networks (especially when presented with weak cues) during the retrieval process improves long-term retention. Thus, the elaborative retrieval hypothesis postulates that the intensity of mental effort invested during the intervention phase accounts for the gains of testing (Carpenter 2009; Carpenter and DeLosh 2006). Furthermore, follow-up studies showed that it was the mental effort *per se* involved in the retrieval of the information during the intervention (intermediate test) that accounted for much of the gains (Endres and Renkl 2015). Bjork and others (2015) agree and highlight the need for a certain level of “desirable difficulty” for the intermediate testing to result in increased recall (Bjork and others 2015). Therefore, it is reasonable to hypothesize that benefits of the testing may be obtained from any activity requiring high mental effort. Therefore, with this study, we tested to what extent student learning outcomes are improved by engaging students in developing pre-exam questions.

In this study, the best learning gains were obtained by students in the 4^{th} quintile (Table 3, who earned between 60% and 80% on the first exam, prior to the intervention). This cohort consisted of students who improved their grade by 12.37% compared with the year when the intervention was not implemented. While encouraging, we note that other studies report that a variety of customized educational experiences tend to benefit weakest students (Nalliah and Allareddy 2014).

Engaging students with developing multiple-choice questions is not entirely novel. The benefits of the activity obtained in this study are significantly higher than those reported previously. In previous studies carried out in several institutions and in courses covering diverse STEM disciplines, student learning gains from the question development activities were modest and/or inconsistent (Hardy and others 2014; Rhind and Pettigrew 2012). In analyzing designs of these studies (Hardy and others 2014; Rhind and Pettigrew 2012), several commonalities became apparent: students were asked to develop a limited number of questions (up to 5), quality of the submitted questions was not monitored and, when, evaluated, student-generated questions represented lower levels of Bloom's Taxonomy of Learning. Therefore, in this study, we provided detailed instructions on how to design questions that focus on testing higher levels of learning. We agree with the authors who suggest that the act of developing multiple-choice questions represents a higher order learning activity regardless of the complexity of the question (Chin and others 2002). It is also likely that multiple-choice questions that are more complex may lead to higher learning gains (Hardy and others 2014), since development of these questions requires a more significant mental effort ("desirable difficulty" (Bjork and others 2015). Therefore, it is reasonable to conclude that our requirement for students to develop multiple-choice questions targeting higher levels of learning is one of the main differences responsible for greater learning gains in this study compared to those of others (Hardy and others 2014; Rhind and Pettigrew 2012). The requirement that students design 20 questions spanning all topics covered during the period leading up to the test may have also ensured that a significant mental effort was involved in preparation for the exam.

At least three other possibilities may account for the students' learning gains. First, Linton and others (2014) reported that writing exercises that focused on student comprehension of the material had a positive effect on student achievement on multiple-choice exams (Linton and others 2014). Second, approximately 75% of the questions on the actual exam were derived from the questions submitted by the students. Considering that all questions submitted by students were posted on-line prior to the exam, one may suspect that students simply memorized all the questions and correct answers. We do not feel that this is likely given that the student pre-exam questions were posted without answers, and there were at least 600 questions in the test bank. It is doubtful that students memorized that many questions and answers. Another possible explanation for the increased learning gains is the notion of transfer-appropriate processing (Morris and others 1977). According to it, similarities between intervention (e.g., pre-exam questions) and the tests drive the testing effect. While this was not specifically tested in this study, others found no experimental support for the transfer-appropriate processing (Carpenter and DeLosh 2006; Endres and Renkl 2015).

We note that similar learning gains were not observed in the two semesters in the years without the intervention. In fact, in the semesters without the intervention, students grades on the first exam was higher than on the second exam. Even though the course and module content were very similar over these years, the questions used on the first exam overlapped by only 50%, and there was negligible overlap in the questions used on the second exam over these years. Given differences in the exam content and presentation, we conclude that "vertical" comparisons (between the cohorts pre- and post-intervention) are less informative than "horizontal" comparisons (within the cohorts that either experienced or did not experience the intervention). Intervention in 2015 prior to the final, and not the second exam was pivotal in determining that the learning gains on the second exam following the intervention were due to the intervention itself, and not a host of other possibly confounding factors. Once the pre-exam activity was removed in 2015, learning gains on the second exam were lost, and the grades were similar to those achieved in 2009 when no intervention took place. The effect size between the control and treatments can be described as medium to large on the bases of calculated Cohen's *d* statistics.

Concluding, to being a valuable learning tool, design of the pre-exam questions was an activity that students enjoyed. We can easily envision this activity being incorporated in a variety of STEM courses.

Acknowledgements

This research was support by USDA NIFA CRIS project 1007300 (FLA-SWS-005474).

CAPTIONS

Table 1. Examples of Multiple-choice questions designed by students.

Table 2. Mean (standard deviation) grades (expressed as percentage) in 2009-2012 and Anova results.

Table 3. Quintile regression. Effect of implementation of the activity of on Exam 2.

Table 4. Student evaluation of the construction of pre-exam questions.

Figure 1. Assessments structure of the module during different years.

Figure 2. Comparison of student performance on Exam 1 (control) and Exam 2 (following intervention). (A) Overall student achievement on the exam 1. Box plots include the lower and upper quartiles, lines within the box are the medians and whiskers indicate the degree of dispersion of the data. (B) Distribution of points earned by the students on exam 2. Box plots include the lower and upper quartiles, lines within the box are the medians and whiskers indicate the degree of dispersion of the data. Dots represent data outliers. Arrows show when the activity has been implemented or removed.

References

Bates SP, Galloway RK, Riise J, Homer D. 2014. Assessing the quality of a student-generated question repository. PRPER 10(2) doi:ARTN 020105

Bjork EL, Soderstrom NC, Little JL. 2015. Can multiple-choice testing induce desirable difficulties? Evidence from the Laboratory and the Classroom. Am J Psychol 128(2):229-239

Bottomley S, Denny P. 2011. A participatory learning approach to biochemistry using student authored and evaluated multiple-choice questions. Biochem Mol Biol Edu 39(5):352-361

Brame CJ, Biel R. 2015. Test-enhanced learning: the potential for testing to promote greater learning in undergraduate science courses. Cbe-Life Sci Edu 14(2)

Carpenter SK. 2009. Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. J Exp Psychol Learn Mem Cogn 35(6):1563-9

Carpenter SK, DeLosh EL. 2006. Impoverished cue support enhances subsequent retention: support for the elaborative retrieval explanation of the testing effect. Mem Cognit 34(2):268-76

Chin C, Brown DE, Bruce BC. 2002. Student-generated questions: a meaningful aspect of learning in science. Int J Sci Edu 24(5):521-549

Coppola BP, Pontrello JK (2014) Using errors to teach through a two-staged, structured review: peer-reviewed quizzes and "What's Wrong With Me?". J Chem Edu 91(12):2148-2154

Draper SW. 2009. Catalytic assessment: understanding how MCQs and EVS can foster deep learning. B J Edu Technol 40(2):285-293

Endres T, Renkl A. 2015. Mechanisms behind the testing effect: an empirical investigation of retrieval practice in meaningful learning. Front Psychol 6:1054

Foos PW. 1989. Effects of student-written questions on student test performance. Teach Psychol 16(2):77-78

Glover JA. 1989. The testing phenomenon - not gone but nearly forgotten. Journal of Edu Psychol 81(3):392-399

Hardy J, Bates SP, Casey MM, Galloway KW, Galloway RK, Kay AE, Kirsop P, McQueen HA. 2014. Student-generated content: enhancing learning through sharing multiple-choice questions. Internat J Sci Edu 36(13):2180-2194

Jacoby LL, Wahlheim CN, Coane JH. 2010. Test-Enhanced Learning of Natural Concepts: Effects on Recognition Memory, Classification, and Metacognition. J Exp Psychol Learn Mem Cogn 36(6):1441-1451

Kang SHK, McDermott KB, Roediger HL. 2007. Test format and corrective feedback modify the effect of testing on long-term retention. E J Cogn Psychol 19(4-5):528-558

Karpicke JD, Roediger HL. 2008. The critical importance of retrieval for learning. Science 319:966-968

Khanna MM. 2015. Ungraded pop quizzes: test-enhanced learning without all the anxiety. Teach Psychol 42(2):174-178

Knight JK, Wood WB. 2005. Teaching more by lecturing less. Cell Biol Educ 4(4):298-310

Kornell N. 2014. Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. J Exp Psychol Learn Mem Cogn 40(1):106-114

Kromann CB, Jensen ML, Ringsted C. 2011. Test-enhanced learning may be a gender-related phenomenon explained by changes in cortisol level. Med Edu 45(2):192-199

Linton DL, Pangle WM, Wyatt KH, Powell KN, Sherwood RE. 2014. Identifying key features of effective active learning: the effects of writing and peer discussion. Cbe-Life Sci Edu 13(3):469-477

Lujan HL, DiCarlo SE. 2014. The flipped exam: creating an environment in which students discover for themselves the concepts and principles we want them to learn. Adv Physiol Edu 38(4):339-342

McDaniel MA, Anderson JL, Derbish MH, Morrisette N. 2007. Testing the testing effect in the classroom. E J Cogn Psychol 19(4-5):494-513

Mcdaniel MA, Fisher RP. 1991. Tests and test feedback as learning sources. Cont Edu Psychol 16(2):192-201

McLaughlin JE, Roth MT, Glatt DM, Gharkholonarehe N, Davidson CA, Griffin LM, Esserman DA, Mumper RJ. 2014. The flipped classroom: a course redesign to foster learning and engagement in a health professions school. Acad Med 89(2):236-43

Morin KH. 2014. Fostering student accountability for learning. J Nurs Edu 53(10):547-548

Morris CD, Bransford JD, Franks JJ. 1977. Levels of processing versus transfer appropriate processing. J Verb Lear Verb Behav 16(5):519-533

Nalliah RP, Allareddy V. 2014. Weakest students benefit most from a customized educational experience for Generation Y students. PeerJ. 2:e682.

Nguyen K, McDaniel MA. 2015. Using quizzing to assist student learning in the classroom: the Good, the Bad, and the Ugly. Teach Psychol 42(1):87-92

Nicol DJ. 2007. E-assessment by design: using multiple-choice tests to good effect. J Furth High Edu 31(1):53-64

Nicol DJ, Macfarlane-Dick D. 2006. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. Studies High Edu 31(2):199-218

O'Flaherty J, Phillips C. 2015. The use of flipped classrooms in higher education: A scoping review. Internet High Edu 25:85-95.

Rhind SM, Pettigrew GW. 2012. Peer generation of multiple-choice questions: student engagement and experiences. J Vet Med Edu 39(4):375-379

Richland LE, Kornell N, Kao LS. 2009. The pretesting effect: do unsuccessful retrieval attempts enhance learning? J Exp Psychol Appl 15(3):243-257

Rosenshine B, Meister C, Chapman S. 1996. Teaching students to generate questions: A review of the intervention studies. Rev Edu Res 66(2):181-221

Spangler J. 2014. Costs related to a flipped classroom. Acad Med 89(11):1429

van Gog T, Sweller J. 2015. Not new, but nearly forgotten: the testing effect decreases or even disappears as the complexity of learning materials increases. Edu Psychol Rev 27(2):247-264

Wood WB, Knight J. 2004. Teaching large biology classes: active-engagement alternatives to lecturing and evidence that they work. Mol Biol Cell 15:233a-233a