Middlesex University Research Repository

An open access repository of

Middlesex University research

Copyright:

1 **Student-generated pre-exam questions is an effective tool for participatory**

2 **learning: A case study from Ecology of Waterborne Pathogens course.**

3

4

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

6

7 [1] Soil and Water Science Department, University of Florida, Gainesville, Florida, United States

8 of America

9 [2] Department of Natural Science, Middlesex University, London, United Kingdom.

10 [3] Department of Family, Youth and Community Sciences, University of Florida, Gainesville,

11 Florida, United States of America

12 [4] Department of Marine Science, Eckerd College, St. Petersburg, Florida, United States of

13 America

14

15

16 * Corresponding author

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

18

19

20 Short title: Learning gains from question design

21

23

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.

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

71   2014). Preparation of quiz test banks is also a major time investment by the instructor

72   (O'Flaherty and Phillips 2015). Several studies identified student-led generation of question

73   banks as an opportunity not only to save the instructor's time and institutional resources, but

74   also to engage students more deeply into the learning process (Bottomley and Denny 2011;

75   Hardy and others 2014; Rhind and Pettigrew 2012). With this study, we aimed to test the

76   effectiveness of student-generated multiple-choice question banks in enhancing learning in a

77   blended classroom.

78

79   It is commonly recognized that the act of creating questions enhances learners' understanding

80   of course materials and promotes deep learning (Draper 2009; Rosenshine and others 1996).

81   Students who wrote questions tended to outperform their peers (Foos 1989). Student-generated

82   questions that involved higher cognitive skills (compared to a simple recall) have been linked to

83   self-directed learning and improved conceptual understanding (Chin and others 2002).

84   However, studies also point to the fact that for this activity to be productive, it has to be properly

85   contextualized (Bottomley and Denny 2011; Hardy and others 2014; Nicol 2007; Nicol and

86   Macfarlane-Dick 2006; Rhind and Pettigrew 2012). For example, in the experiments of Rhind et

87   al (2012) conducted in three courses of a veterinary curriculum over a two-year period, students

88   were directed to develop questions that were submitted into an on-line bank without being

89   evaluated (Rhind and Pettigrew 2012). In only 1 of 6 experiments, the correlation between

90   answering questions submitted by peers and the exam grade was significant ($p < 0.05$), despite

91   the fact that over 80% of students agreed that developing original questions and answering

92   questions developed by others helped improve understanding of the material. Students in this

93   study submitted and answered only two multiple-choice questions, and the quality of the

94   student-generated questions was not evaluated. Hardy and others (2014) in a three-school UK

95   study encompassing courses in Physics, Genetics and Chemistry reported only modest, but

96   nevertheless positive, impact of writing and reviewing exam questions on the end-of-term

97 performance (Hardy and others 2014). As was in the study of Rhind (2012), participants in the

98 Hardy (2014) study submitted a limited number of un-moderated questions (between 1 and 2),

99 however, students were required to answer or evaluate between 2 and 20 questions submitted

100 by their peers. Hardy and collaborators (2014) hypothesized that monitoring the quality of the

101 questions that students submit may have improved learning gains. Bottomley and others, 2011

102 reported that when over one hundred biomedical students were asked to develop own multiple

103 choice questions, over 90% of the questions represented the lowest two of Bloom's taxonomy

104 (Bottomley and Denny 2011). However, if students were specifically instructed to write multiple

105 choice questions representing higher levels of Bloom's taxonomy, they did so (Bates and others

106 2014). Thus, it appears, that for this exercise to result in learning gains, students have to

107 engage at multiple higher level tasks, such as creation of the more sophisticated questions,

108 engagement with questions developed by their peers and obtaining timely feedback.

109

110 It is also reasonable to hypothesize that learning gains resulting from the construction of exam

111 questions could be due to the ability of students to recall significantly more information on the

112 final test if they were subjected to intermediate self-administered tests prior to the final

113 examination (Glover 1989). Both the types of the tests and recall activities have an impact on

114 learning. For information that is not complex (i.e. memorization and recall of word pairs in

115 English and in a foreign language) learning gains are most improved with a repeated retrieval

116 practice (Karpicke and Roediger 2008). For complex information, where the ability to recall

117 certain information requires memorization (or thorough understanding) of the context, retrieval

118 has to modify the information in memory, rather than storing it unchanged (Bjork and others

119 2015). Furthermore, it involves either an elaboration of the memorized material (Mcdaniel and

120 Fisher 1991) or requires a certain level of "desirable difficulty" (a condition that initially slows

121 down knowledge acquisition, but improves long-term retention and transfer of knowledge),

122 which can be achieved when the recall activities are reasonably complicated (Bjork and others

123   2015). Therefore, the aim of this study was to test whether student's generation of multiple-

124   choice questions could result in robust learning gains that are similar to those observed in

125   studies of the testing effect and without the undesirable effects of anxiety and biases observed

126   in previous studies (Bjork and others 2015; Kromann and others 2011; Nguyen and McDaniel

127   2015) This activity required not only a simple recall of information, but a thorough understanding

128   of the tested material and additional information.

129

130   In addition to asking students to develop and submit questions on-line, we wanted to further

131   capitalize on this learning opportunity and instructed students to use questions that were

132   developed by their peers and posted within an e-learning platform for self-paced quizzes. The

133   most difficult questions were discussed during two one-hour class periods. Self-paced and un-

134   graded pop quizzes were shown to be effective learning tools, which confer the benefits of the

135   activity without the negative impact of stress and anxiety (Khanna 2015).

136

137   **Methods**

138   **Ethics Statement**

139   All survey questions and the study design were reviewed by the University of Florida

140   Institutional Review Board. All study participants gave written consent to take part in the study.

141   Consent letter was drafted by T.I. (who was not the course instructor), and were administered

142   by a third party. Whether or not students gave their consent to participate in the study was not

143   revealed to the course instructors until after the final grade for the course was finalized and

144   submitted. University of Florida Institutional Review Board reviewed the consent procedure.

145

146   **Participants**

147   This experiment was conducted in the course "Ecology of Waterborne Pathogens", which is

148 taught once a year. This course is an approved elective for Biology, Microbiology, Soil and

149 Water Science majors. Over the duration of the experiment, 162 students chose to participate

150 in this study from 2009 to 2015 (of them, ~55% where female and ~45% male). Development of

151 the pre-exam questions for the bank and performance on exams were graded activities in the

152 course. Two control cohorts, enrolled in the course in 2009 the year prior to the intervention

153 consisted of 32 students, and the 2015 cohort included 30 students.

154

155 **Course Design**

156 The course included traditional lectures, in-class discussions, pre-recorded on-line instructional

157 videos, weekly on-line discussions, a nine-week long investigation of a simulated outbreak and

158 an activity in which students constructed multiple-choice questions. Approximately two weeks

159 prior to the first class, the instructor emailed students a list of suggested course topics, and

160 students ranked them according to their interest in the subject. Based on the results of the

161 students' votes, more or less time was allocated to certain topics. Some topics (food safety,

162 urban microbiology and diseases of aquatic animals) were not discussed in all years, however,

163 these topics were typically covered in the third module of the course (after the second exam).

164 The first module typically included surveys of outbreaks of water- and foodborne illnesses,

165 environmentally-transmitted pathogens (including emerging and re-emerging pathogens),

166 sample collection and preparation, culture-based methods for pathogen isolation and detection,

167 physiological, immunological and nucleic acid-based detection and characterization, source-

168 tracking of human pathogens, a survey of environments where pathogens survive (soil, air,

169 water, plant-associated environments). Upon completion of the first module, students completed

170 the first exam. During the second module, the following topics were typically covered: biofilms

171 and microbial cell-to-cell signaling, evolution of antibiotic resistance and virulence, pathogens as

172 prey, microbiological components of water quality, indicator organisms, drinking and waste

173 water treatment and disinfection. At the end of the second module students completed the

174 second exam (Figure 1).

175

176 **Study Design**

177 As mentioned above the course was divided into three modules. The assessment structure is

178 shown in Figure 1. From 2010 to 2012 the intervention was implemented before Exam 2. In

179 2015, an additional control was introduced: the intervention was implemented before a third

180 exam instead of the Exam 2. For the intervention, students were instructed to design 20

181 multiple-choice questions. Prior to the assignment, a 15-minute presentation about Bloom's

182 taxonomy of learning was offered in class by one of the instructors. Students also received a

183 handout summary of the presentation, and examples of questions that would not be accepted

184 for credit. Students were instructed that less than 20% of their questions could represent the

185 lower two domains of Bloom's taxonomy and that True/False questions (even when disguised

186 as multiple-choice questions such as: "Which of the following four statements is correct...")

187 would not be accepted for credit. Students were also instructed that questions based on a

188 single power-point slide from lectures would not be accepted for credit, thus forcing them to tap

189 into higher levels of learning. Questions had to be scientifically accurate and grammatically

190 correct, and stems of questions had to be based directly on the topics that were discussed

191 during the second module, although integration of the topics from the first module was also

192 encouraged (see, for example, Questions 1, 6 and 8 in the Table 1). These criteria were used

193 for grading the questions that students submitted. All questions, without the answers, were

194 posted by students in the e-learning system to be accessible to all students. Through the

195 duration of this study, the institution used two different e-learning platforms (Sakai and Canvas).

196 However, because functionalities of the two platforms were similar, this was not considered as a

197 variable that impacted the outcome of the study. Students were offered an opportunity to send

198 some or all of their questions for the instructor's formative feedback prior to posting them (and

199 approximately a third of students took advantage of that opportunity). The instructors provided

200 feedback and graded the posted questions prior to the exam. As the questions were submitted

201 in the e-learning system without the answers, all students were encouraged to use them for self-

202 paced tests (not graded by the instructors), and approximately 80 of the most difficult questions

203 were discussed during two class periods prior to the exam, however review was offered in all

204 sections. None of the questions that were discussed during the review session were included in

205 the actual exam.

206

207 **Exam and Assessment**

208 All questions from Exam 1 were designed by the instructor. Exam 1 consisted of multiple-

209 choice questions and a list of five questions for short essays from which students selected three

210 to answer. Multiple-choice questions accounted for ~70% of grade, while short essays made up

211 the rest. Approximately 80% of the multiple-choice questions on the first exam were the same

212 for the three semesters in which the intervention took place; there was only ~50% overlap

213 between questions on the first exam in the semesters when the intervention took place and in

214 the two semesters prior to the intervention.

215

216 The second exam consisted of 20 questions, ~75% of them were drawn from over 600

217 questions submitted by the students to the question bank during that semester, and ~25% of the

218 questions were designed by the instructor. Questions developed by students were edited by the

219 instructor for clarity and grammar. Multiple-choice questions offered by the instructors in the

220 semesters when the intervention took place were essentially the same. In the year prior to the

221 intervention, all questions were designed by the course instructors. To accommodate a variety

222 of learning styles, students had a choice to complete multiple-choice or short essay questions

223 on the second exam; students self-selected which version of the exam to take. Essay exams

224 included a "menu" of 15-20 questions, from which students selected 10-12 to answer. Each

225 question required a clear understanding of at least 2 different concepts discussed throughout

226   the course. For example, "We have discussed several examples of water quality indicators.

227   Can indicator organisms be used in programs aimed at controlling accidental or opportunistic

228   waterborne pathogens (such as *Aeromonas, Vibrio cholera, Vibrio vulnificus*)? Why or why

229   not?"; "How would you use multiplex PCR to identify subspecies of *Salmonella*? Design primers

230   for this experiment, include picture of the gel that you expect will result from this PCR

231   experiment. Make sure to include all appropriate positive and negative controls. In addition to

232   multiplex PCR targeting genes you identified as important, what 3 other genetic techniques

233   could be used to distinguish between subspecies of *Salmonella*? Why? What would the data

234   look like (provide examples of gels)?", "Resistance to antibiotics is often associated with

235   acquisition of the new genes and functions. Please provide 3 hypothetical scenarios in which

236   loss of a gene or a mutation of a gene within the core genome would increase resistance of a

237   pathogen to antimicrobials." Mastery of the same concepts was required to complete multiple

238   choice (MC) exams. Over the duration of this study, approximately 40% of students chose to

239   take the essay version of the exam. All exams were graded and moderated by the two

240   instructors in the course to ensure consistency (with the exception of one semester prior to the

241   intervention). Essays were graded based on whether students demonstrated understanding of

242   individual concepts, ability to integrate them and to provide clear, concise and accurate

243   answers. When achievements of students who took the essay or multiple-choice exam were

244   compared, no statistically significant differences were observed. Students' achievement on the

245   second exam was compared with their achievement on the first exam. As a control, there was a

246   formative pre-exam review session prior to the first exam, in which the instructor led an in-class

247   discussion of the key concepts covered during the first module.

248

249   **Coding of student questions according to Blooms's taxonomy.** Pre-exam questions

250   developed by students were considered to be recall-type questions if they required a

251   restatement of a specific fact or a definition and were based on a single power point slide;

252 comprehension questions required basic understanding of at least two different concepts, and

253 were often exemplified by simple analogies; application questions involved basic case study

254 analyses, data interpretation or calculations; analysis and synthesis questions required original

255 analysis of complex data.

256

257 **Activity satisfaction questionnaire.** Upon completion of the semester, we administered a

258 questionnaire, which aimed to determine which factors (such as clarity of the assignments,

259 recognition of students' contributions, appropriateness of guidance by the course instructors,

260 relevance of the assignment) correlated with the overall satisfaction with the activity. Students

261 answered questions by assigning either 1 (strongly disagree) or 5 (strongly agree) to each

262 question. Participation in the questionnaire at the end of the course was voluntary, and students

263 received extra credit points for filling it out. Eighty seven students completed this post-activity

264 questionnaire consisted of 11 questions. In order to measure factors associated with the

265 student satisfaction with the intervention, linear regressions of each question correlated with the

266 overall satisfaction was performed. Statistical analysis was carried out with JMP (SAS) software

267 package.

268

269 **Statistical analysis to assess the impact of the pre-exam activity on student achievement.**

270 Individual student grades were converted into percentages and three tests were performed: 1) A

271 one-way ANOVA across cohorts was inferred to determine differences in mean grades on Exam

272 1, the pre-exam and Exam 2, and in changes between Exams 1 and 2. 2) A quintile regression

273 analysis to identify differential effects of the pre-test and Exam 1 performance among different

274 groups of grades in Exam 2. In other worlds, the quintile analysis allowed us to identify to what

275 extent students grouping into different grades (0-20%, 20-40%, 40-60%, 60-80% and 80-100%)

276 were affected by the implementation of the activity. To test for differences in the performance of

277 cohorts in Experiment 1 on each exam (Exam 1, pre-exam, and Exam 2) a one-way ANOVA

278 was performed on the mean of the grades and on the differences between grades on Exam 2

279 and grades on Exam 1. To check for the effects of the pre-test among different groups of

280 grades, a quintile regression has been performed including both pre-test and grades on the

281 Exam 2. Statistical analyses were carried out in R and JMP (SAS) software package. While

282 preliminary analyses of the data were carried out by the course instructors, final analyses

283 presented in this paper were conducted by the individuals who were not associated with the

284 study design or course instruction.

285

286

287 **Results**

288 ***Quality of student-generated pre-exam questions.*** Prior to completing the assignment,

289 students were introduced to Bloom's Taxonomy of Learning and were given specific instructions

290 on what pre-exam questions would be accepted for full credit. Only approximately 15% of the

291 questions submitted by students were simple recall-type questions. The majority of the

292 questions engaged higher levels of learning. Approximately 35% of the questions were scored

293 as comprehension-type questions, 28% were scored as application, 20% were analysis and the

294 rest were synthesis-type questions, according to Bloom's Taxonomy. For example, Q1 (Table 1)

295 required that students were able to recall definition of "fecal coliform", analyze composition of

296 the xylose lysine deoxycholate (XLD) medium and deduce how fecal coliforms would behave

297 when plated on XLD agar. Construction of Q2 and Q3 required comprehension of the topics

298 discussed under Fundamentals of Microbial Evolution, using examples from the instructor's

299 presentations on common food- and waterborne pathogens as well as indicators of water

300 quality. To design Q4, students had to recall that *S. bongori* carries *Salmonella* Pathogenicity

301 Island 1 (discussed in Evolution of Enteric Pathogens) and synthesize it with the information

302 discussed under Virulence Mechanisms. Virulence of *S. bongori* was not an explicit topic that

303 was discussed in class.

304

305 Because majority of students in this course were on the pre-med track, they used this as an

306 opportunity to apply information they learned throughout the course to what was discussed in

307 this course and in other classes they have taken (see, for example Q5 and Q6). Construction of

308 these questions required comprehension of the mechanisms of action of antibiotics (discussed

309 in this course under Antibiotics: Mechanisms of Action and Resistance) as well as the

310 instructor's presentation Pathogens As Prey and a number of concepts discussed under

311 Quorum Sensing, and a presentation on Detection of Waterborne Pathogens. Students were

312 specifically instructed that stems of questions had to be based on the material covered in the

313 second module of the course, and questions that could have been answered without taking the

314 course were not accepted for credit.

315

316 We note that analogies represented a significant number of questions designed by students

317 (over 22% overall). Analogy questions designed by students clearly represented higher levels

318 of learning (for example, Q7 and Q8). Of these two questions, Q7 was not accepted for full

319 credit because it was based on the same power point slide in the instructor's presentation on

320 Quorum Sensing. Q8 demonstrated synthesis of topics discussed under Pathogens As Prey

321 and introductory presentations on the life cycle of waterborne human pathogens.

322

323 ***Impact of the pre-exam activity on student achievement.***

324 The first hypothesis tested with this study was that student achievement on an exam would be

325 improved by constructing pre-exam questions before Exam 2. Student achievement on this pre-

326 exam assignment (construction of exam questions) averaged 89.56±14.16%, 98.24±4.12%, and

327 96.60±7.37%, for 2010, 2011 and 2012 respectively (Table 2). No significant differences in

328 student achievement on the pre-exam were observed over the three years (F = 2.12, p=0.125).

329 Following the intervention from 2010 to 2012, average grade on the post-intervention exam

330 increased by 7.44%.

331

332 On Exam 1, which was administered following an in-class discussion of topics covered prior to

333 the exam, but without the intervention, the four-year average was 78.70% (Table 2, Figure 2A).

334 No differences in students' performance were observed for Exam 1 (F=0.67, p=0.575) in the 4

335 years of observations (Figure 2 and Table 2).

336

337 The three-year average on the second exam, administered following the intervention, which

338 involved development, submission and discussion of the student-generated questions, was

339 86.58±12.86% (Figure 2 B). There were no statistically significant differences in student

340 achievement on the second exams over the three years when the intervention was implemented

341 (86.46±10.38%, 86.42±11.66%, and 86.86±16.56% in 2010, 2011 and 2012, respectively).

342 Student achievement on the same exam in 2009 (when students did not submit pre-exam

343 questions prior to the second exam) was significantly lower than in the years when the

344 intervention was implemented (Figure 2 B).

345 When comparing students' performance on Exam 2, the average grade in 2010-2012 (when the

346 pre-exam was carried out) was statistically higher than the average grade in 2009 (when no pre-

347 exam was performed) (F=4.84, p=0.003) (Figure 2, B). The change in performance between

348 Exam 1 and Exam 2 (see Exam2-Exam1 column, Table 2) also shows a significantly greater

349 improvement for the 2010-2012 cohorts than for the 2009 cohort (F=4.90, p=0.003).

350 An additional control experiment was carried out in 2015. Students were instructed to design

351 multiple choice questions as in the Experiment 1 prior to the third (final, and not the second)

352 exam. This was done to address the possibility that the differences in learning gains observed

353 on the second exam during previous years were due not to an intervention but due to the

354 inherent relative "easiness" of the material covered during the second module or due to some

355   other factors that may be associated with the particular timeframe during the spring semester.

356   The grades for the second exam were used as a control for all the years, in which the students

357   were asked to write the pre-exam questions before the second exam.   When grades from

358   Exams 2 where compared along the years, the two years (2009 and 2015) in which the pre-

359   exam was not implemented were significantly lower (p=0.001) when compared with the years in

360   which the intervention was implemented (Figure 2 B), thus indicating that learning gains on the

361   second exam were most likely due to the intervention, and not to other factors.

362

363   Despite the fact that the intervention (construction and discussion of the pre-exam questions)

364   appears to have been associated with a significant increase in student learning gains, there was

365   no correlation between points earned for the design of the pre-exam questions and their grade

366   on the second exam (p=0.198).

367

368   Quintile regression was also performed in order to measure to what extent the implementation

369   of the pre-exam affected each quintile of the grades (0 to 100%). In the quintile regression for

370   each unit, an increase in student performance on the second exam varied by quintile.

371   Interestingly when the pre-test was used, a significant improvement in student achievement on

372   Exam 2 was observed.

373

374   It is important to point out that not all students benefited equally from this activity (Table 3).

375   Based on the student achievement on the first exam, ninety-nine students participating in the

376   study were divided into five quintiles to track their improvement. Following the intervention, six

377   out of nine students in the second quintile (20-40%) improved their scores on the second exam

378   on an average of 6.88% percentage points. Students who were in the fourth quintile based on

379   the results of the first exam demonstrated the highest achievement (measured as a score on the

380   second exam). Students in the 4th quintile (60% to 80% of correct answers), twenty-nine out of

381 thirty-eight students improved their performance on an average by 12.37% percentage points.

382 Finally, of the students who were within the top (fifth) quintile (from 80% to 100% of correct

383 answers on the first exam), thirty-one students out of fifty-two improved their performance by an

384 average of 10.88% percentage points (Table 3).

385

386 ***Factors associated with the student satisfaction with the intervention***. In order to assess

387 the impacts of the intervention (construction and discussion of the multiple-choice pre-exam

388 questions), students were surveyed on how they felt their contributions were received.

389 Students were given a questionnaire consisting of 11 questions (Table 4) upon completion of

390 the course. We wanted to determine whether students felt that the assignment was clear,

391 challenging and relevant, whether they have received the appropriate level of guidance, whether

392 their contribution was properly recognized. Overall, study participants felt that construction of

393 the pre-exam questions was a rewarding experience (3.9/5.0, Table 4), despite the fact that they

394 were generally unsure that this activity would be relevant in their future careers (Table 4).

395

396 Whether or not rules of the assignment were clear, appropriateness of instructor's guidance,

397 proportionality of earned points to the investment of energy, relevance of the experience to the

398 future career, recognition of the student's investment by the instructor all had a statistically

399 significant correlation with the overall satisfaction with the assignment (Table 4). Nevertheless,

400 the r^2 values for all of these correlations were low, with the relevance to the future career,

401 instructor recognition and the proportionality of the investment to the overall grade having the

402 strongest impact on the overall satisfaction with the assignment (Table 4).

403

Discussion

405 As instructors retreat from relying on lectures as a main instructional tool to adopt more

406   participatory strategies, student-centered educational models are becoming more wide-spread.

407   In a classroom, whether traditional or flipped, testing remains an important component of the

408   educational process and institutional assessment. Beneficial effects of testing on long-term

409   retention of knowledge have been well documented (Brame and Biel 2015; Jacoby and others

410   2010; Kang and others 2007; Kromann and others 2011; McDaniel and others 2007; van Gog

411   and Sweller 2015). Learning gains are observed even when students attempt to answer

412   questions, but either fail to answer them correctly or do not receive timely feedback (Kornell

413   2014; Richland and others 2009). It is hypothesized that attempting to answer questions

414   activates cognitive networks, potentially allowing retrieval of related content and identifying the

415   need for related information (Richland and others 2009).

416

417   Exams and quizzes are stress- and anxiety-inducing activities (Nguyen and McDaniel 2015) and

418   references therein), and benefits of test-enhanced learning maybe gender-biased (Kromann and

419   others 2011). Furthermore, poorly constructed multiple-choice pre-exams with implausible

420   alternative answers do not result in meaningful learning gains (Bjork and others 2015). Frequent

421   quizzes may also be seen by students as an attempt to assert (or even usurp) authority by the

422   instructor, and thus undermine development of the student-centered participatory classroom. To

423   reduce the levels of stress and anxiety associated with exam-taking, instructors have

424   experimented with ungraded pop quizzes (Khanna 2015), peer-graded quizzes (Coppola and

425   Pontrello 2014) and also "flipped exams", in which students work collaboratively to solve exam

426   questions (Lujan and DiCarlo 2014). Clearly, engaging students into an active learning process

427   can take different forms.

428

429   When the benefits of testing were further dissected, several hypotheses (e.g., elaborative

430   retrieval and transfer-appropriate processing theories, retrieval induced facilitation, unspecific-

431   goal perspective) attempted to explain its cognitive benefits (Carpenter 2009; Carpenter and

432 DeLosh 2006). According to the elaborative retrieval hypothesis, activation of the elaborative

433 semantic networks (especially when presented with weak cues) during the retrieval process

434 improves long-term retention. Thus, the elaborative retrieval hypothesis postulates that the

435 intensity of mental effort invested during the intervention phase accounts for the gains of testing

436 (Carpenter 2009; Carpenter and DeLosh 2006). Furthermore, follow-up studies showed that it

437 was the mental effort *per se* involved in the retrieval of the information during the intervention

438 (intermediate test) that accounted for much of the gains (Endres and Renkl 2015). Bjork and

439 others (2015) agree and highlight the need for a certain level of "desirable difficulty" for the

440 intermediate testing to result in increased recall (Bjork and others 2015). Therefore, it is

441 reasonable to hypothesize that benefits of the testing may be obtained from any activity

442 requiring high mental effort. Therefore, with this study, we tested to what extent student

443 learning outcomes are improved by engaging students in developing pre-exam questions.

444 In this study, the best learning gains were obtained by students in the 4th quintile (Table 3, who

445 earned between 60% and 80% on the first exam, prior to the intervention). This cohort consisted

446 of students who improved their grade by 12.37% compared with the year when the intervention

447 was not implemented. While encouraging, we note that other studies report that a variety of

448 customized educational experiences tend to benefit weakest students (Nalliah and Allareddy

449 2014).

450 Engaging students with developing multiple-choice questions is not entirely novel. The benefits

451 of the activity obtained in this study are significantly higher than those reported previously. In

452 previous studies carried out in several institutions and in courses covering diverse STEM

453 disciplines, student learning gains from the question development activities were modest and/or

454 inconsistent (Hardy and others 2014; Rhind and Pettigrew 2012). In analyzing designs of these

455 studies (Hardy and others 2014; Rhind and Pettigrew 2012), several commonalities became

456 apparent: students were asked to develop a limited number of questions (up to 5), quality of the

457 submitted questions was not monitored and, when, evaluated, student-generated questions

458 represented lower levels of Bloom's Taxonomy of Learning. Therefore, in this study, we

459 provided detailed instructions on how to design questions that focus on testing higher levels of

460 learning. We agree with the authors who suggest that the act of developing multiple-choice

461 questions represents a higher order learning activity regardless of the complexity of the question

462 (Chin and others 2002). It is also likely that multiple-choice questions that are more complex

463 may lead to higher learning gains (Hardy and others 2014), since development of these

464 questions requires a more significant mental effort ("desirable difficulty" (Bjork and others 2015).

465 Therefore, it is reasonable to conclude that our requirement for students to develop multiple-

466 choice questions targeting higher levels of learning is one of the main differences responsible

467 for greater learning gains in this study compared to those of others (Hardy and others 2014;

468 Rhind and Pettigrew 2012). The requirement that students design 20 questions spanning all

469 topics covered during the period leading up to the test may have also ensured that a significant

470 mental effort was involved in preparation for the exam.

471

472 At least three other possibilities may account for the students' learning gains. First, Linton and

473 others (2014) reported that writing exercises that focused on student comprehension of the

474 material had a positive effect on student achievement on multiple-choice exams (Linton and

475 others 2014). Second, approximately 75% of the questions on the actual exam were derived

476 from the questions submitted by the students. Considering that all questions submitted by

477 students were posted on-line prior to the exam, one may suspect that students simply

478 memorized all the questions and correct answers. We do not feel that this is likely given that the

479 student pre-exam questions were posted without answers, and there were at least 600

480 questions in the test bank. It is doubtful that students memorized that many questions and

481 answers. Another possible explanation for the increased learning gains is the notion of transfer-

482 appropriate processing (Morris and others 1977). According to it, similarities between

483 intervention (e.g., pre-exam questions) and the tests drive the testing effect. While this was not

484   specifically tested in this study, others found no experimental support for the transfer-

485   appropriate processing (Carpenter and DeLosh 2006; Endres and Renkl 2015).

486

487   We note that similar learning gains were not observed in the two semesters in the years without

488   the intervention. In fact, in the semesters without the intervention, students grades on the first

489   exam was higher than on the second exam. Even though the course and module content were

490   very similar over these years, the questions used on the first exam overlapped by only 50%, and

491   there was negligible overlap in the questions used on the second exam over these years. Given

492   differences in the exam content and presentation, we conclude that "vertical" comparisons

493   (between the cohorts pre- and post-intervention) are less informative than "horizontal"

494   comparisons (within the cohorts that either experienced or did not experience the intervention).

495   Intervention in 2015 prior to the final, and not the second exam was pivotal in determining that

496   the learning gains on the second exam following the intervention were due to the intervention

497   itself, and not a host of other possibly confounding factors. Once the pre-exam activity was

498   removed in 2015, learning gains on the second exam were lost, and the grades were similar to

499   those achieved in 2009 when no intervention took place. The effect size between the control

500   and treatments can be described as medium to large on the bases of calculated Cohen's d

501   statistics.

502   Concluding, to being a valuable learning tool, design of the pre-exam questions was an activity

503   that students enjoyed. We can easily envision this activity being incorporated in a variety of

504   STEM courses.

505

Acknowledgements

508

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

534   authored and evaluated multiple-choice questions. Biochem Mol Biol Edu 39(5):352-361

535 Brame CJ, Biel R. 2015. Test-enhanced learning: the potential for testing to promote greater

536   learning in undergraduate science courses. Cbe-Life Sci Edu 14(2)

537 Carpenter SK. 2009. Cue strength as a moderator of the testing effect: the benefits of

538   elaborative retrieval. J Exp Psychol Learn Mem Cogn 35(6):1563-9

539 Carpenter SK, DeLosh EL. 2006. Impoverished cue support enhances subsequent retention:

540   support for the elaborative retrieval explanation of the testing effect. Mem Cognit

541   34(2):268-76

542 Chin C, Brown DE, Bruce BC. 2002. Student-generated questions: a meaningful aspect of

543   learning in science. Int J Sci Edu 24(5):521-549

544 Coppola BP, Pontrello JK (2014) Using errors to teach through a two-staged, structured review:

545   peer-reviewed quizzes and "What's Wrong With Me?". J Chem Edu 91(12):2148-2154

546 Draper SW. 2009. Catalytic assessment: understanding how MCQs and EVS can foster deep

547   learning. B J Edu Technol 40(2):285-293

548 Endres T, Renkl A. 2015. Mechanisms behind the testing effect: an empirical investigation of

549   retrieval practice in meaningful learning. Front Psychol 6:1054

550 Foos PW. 1989. Effects of student-written questions on student test performance. Teach

551   Psychol 16(2):77-78

552 Glover JA. 1989. The testing phenomenon - not gone but nearly forgotten. Journal of Edu

553   Psychol 81(3):392-399

554 Hardy J, Bates SP, Casey MM, Galloway KW, Galloway RK, Kay AE, Kirsop P, McQueen HA.

555   2014. Student-generated content: enhancing learning through sharing multiple-choice

556   questions. Internat J Sci Edu 36(13):2180-2194

557 Jacoby LL, Wahlheim CN, Coane JH. 2010. Test-Enhanced Learning of Natural Concepts:

558   Effects on Recognition Memory, Classification, and Metacognition. J Exp Psychol Learn

559   Mem Cogn 36(6):1441-1451

560   Kang SHK, McDermott KB, Roediger HL. 2007. Test format and corrective feedback modify the

561   effect of testing on long-term retention. E J Cogn Psychol 19(4-5):528-558

562   Karpicke JD, Roediger HL. 2008. The critical importance of retrieval for learning. Science

563   319:966-968

564   Khanna MM. 2015. Ungraded pop quizzes: test-enhanced learning without all the anxiety.

565   Teach Psychol 42(2):174-178

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

567   Kornell N. 2014. Attempting to answer a meaningful question enhances subsequent learning

568   even when feedback is delayed. J Exp Psychol Learn Mem Cogn 40(1):106-114

569   Kromann CB, Jensen ML, Ringsted C. 2011. Test-enhanced learning may be a gender-related

570   phenomenon explained by changes in cortisol level. Med Edu 45(2):192-199

571   Linton DL, Pangle WM, Wyatt KH, Powell KN, Sherwood RE. 2014. Identifying key features of

572   effective active learning: the effects of writing and peer discussion. Cbe-Life Sci Edu

573   13(3):469-477

574   Lujan HL, DiCarlo SE. 2014. The flipped exam: creating an environment in which students

575   discover for themselves the concepts and principles we want them to learn. Adv Physiol

576   Edu 38(4):339-342

577   McDaniel MA, Anderson JL, Derbish MH, Morrisette N. 2007. Testing the testing effect in the

578   classroom. E J Cogn Psychol 19(4-5):494-513

579   Mcdaniel MA, Fisher RP. 1991. Tests and test feedback as learning sources. Cont Edu Psychol

580   16(2):192-201

581   McLaughlin JE, Roth MT, Glatt DM, Gharkholonarehe N, Davidson CA, Griffin LM, Esserman

582   DA, Mumper RJ. 2014. The flipped classroom: a course redesign to foster learning and

583   engagement in a health professions school. Acad Med 89(2):236-43

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

585   Morris CD, Bransford JD, Franks JJ. 1977. Levels of processing versus transfer appropriate

586   processing. J Verb Lear Verb Behav 16(5):519-533

587 Nalliah RP, Allareddy V. 2014. Weakest students benefit most from a customized educational

588   experience for Generation Y students. PeerJ. 2:e682.

589 Nguyen K, McDaniel MA. 2015. Using quizzing to assist student learning in the classroom: the

590   Good, the Bad, and the Ugly. Teach Psychol 42(1):87-92

591 Nicol DJ. 2007. E-assessment by design: using multiple-choice tests to good effect. J Furth

592   High Edu 31(1):53-64

593 Nicol DJ, Macfarlane-Dick D. 2006. Formative assessment and self-regulated learning: a model

594   and seven principles of good feedback practice. Studies High Edu 31(2):199-218

595 O'Flaherty J, Phillips C. 2015. The use of flipped classrooms in higher education: A scoping

596   review. Internet High Edu 25:85-95.

597 Rhind SM, Pettigrew GW. 2012. Peer generation of multiple-choice questions: student

598   engagement and experiences. J Vet Med Edu 39(4):375-379

599 Richland LE, Kornell N, Kao LS. 2009. The pretesting effect: do unsuccessful retrieval attempts

600   enhance learning? J Exp Psychol Appl 15(3):243-257

601 Rosenshine B, Meister C, Chapman S. 1996. Teaching students to generate questions: A

602   review of the intervention studies. Rev Edu Res 66(2):181-221

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

604 van Gog T, Sweller J. 2015. Not new, but nearly forgotten: the testing effect decreases or even

605   disappears as the complexity of learning materials increases. Edu Psychol Rev

606   27(2):247-264

607 Wood WB, Knight J. 2004. Teaching large biology classes: active-engagement alternatives to

608   lecturing and evidence that they work. Mol Biol Cell 15:233a-233a