Generating Practice Questions as a Preparation Strategy for Introductory Programming Exams

Paul Denny
Department of Computer Science
The University of Auckland
Auckland, New Zealand
paul@cs.auckland.ac.nz

ABSTRACT

Written exams are a common form of assessment in introductory programming courses. Creating exam questions is normally the responsibility of the course instructor, however the process of authoring such questions may be a useful learning activity in itself. We explored this idea with a randomized controlled experiment ($n > 700$) in which a group of first-year programming students generated practice questions prior to an exam. Even though all questions were available to every student in the course for practice, the group that generated the questions performed significantly better on the exam. The effects were most pronounced when students answered exam questions on topics that were targeted by questions they had generated. We suggest that some existing tools for computer science education may benefit from incorporating related activities.

Categories and Subject Descriptors

K.3.2 [**Computers and Education**]: Computer and Information Science Education—*computer science education*

General Terms

Design, Human Factors

Keywords

multiple-choice; exams; programming; PeerWise; student authored questions

1. INTRODUCTION

Assessment is fundamental to education, and exams form a key component of many assessment strategies at the tertiary level. Exams are typically administered after a period of instruction to ensure that intended learning outcomes have been achieved. Indeed, most university courses conclude with a high-stakes final exam which ultimately determines the success or failure of each student. Students are known to adopt a wide variety of techniques, some more effective than others, when preparing for an exam [5, 8]. Identifying successful study strategies is therefore of great interest. In this work, we evaluate the effectiveness of having students generate their own practice questions as preparation for an exam in an introductory programming course.

SIGCSE'15, March 4–7, 2015, Kansas City, MO, USA.

http://dx.doi.org/10.1145/2676723.2677253.

2. BACKGROUND

Paper-based, written exams are not the best way to assess many of the practical skills that students develop when learning to program [12, 17]. Writing code on paper, without the support of an editor or a compiler, is a difficult and unnatural task [2]. Nevertheless, high-stakes written exams are commonplace in programming courses and, for certain concepts, can be an effective way of assessing knowledge. For example, a student's ability to comprehend code can be demonstrated by answering written questions that present code fragments to be traced. Not only is code tracing an important debugging skill, but some minimal competence at code tracing must almost certainly be developed before code can be successfully written [20].

Multiple-choice questions are commonly used to examine code comprehension and numerous studies of novice programming exams involve this format [10]. An investigation by Shuhidan et al. into the use of multiple-choice questions in CS1 exams revealed that a majority of programming instructors believed such questions could adequately test understanding of the low level programming concepts required for successful code tracing [18].

The activity we describe in this paper, in which students author their own multiple-choice practice questions, leverages a highly robust effect from the cognitive psychology literature, the *generation effect*. This refers to the finding that individuals tend to remember information better when they take an active role in generating it, rather than being provided with information generated by others [19]. Consider the various ways that students engage with course concepts when authoring their own questions. They must reflect on the material they are learning to identify the topics that are most relevant. They may review related material from their textbook and execute example fragments of code while constructing answers to their questions. As the questions are to be reviewed by their peers, they must be clear and unambiguous, and possible misunderstandings should be considered to create good plausible alternative answers. In this study, we compare the exam performance of students who generate a small number of practice questions with students who simply use these generated questions to practice.

3. RELATED WORK

The impact on learning of having students author content-related questions has been studied in various settings. Early work in this area focused on text comprehension tasks. For example, Frase et al. conducted an experiment in which students were given a biographical passage and informed they would be examined on its content [7]. They found that students who were asked to generate potential test questions while they read performed significantly better when actually tested than students who simply read the text. This result was reinforced by similar experiments carried out independently by Denner et al. and Lehman et al. [3, 9].

However, these studies also revealed a limitation of the approach. The superior performance of students who had authored questions relating to the studied material was limited to certain post-test items. The question-generating students only performed better on post-test items that targeted the same content as the questions they had authored themselves. This suggests that the benefits observed from authoring questions are a direct result of active engagement with the related concepts.

The efficacy and limitations of student-generated questions were further explored in an experiment by Foos et al. where students were provided a 5-page scientific passage to study [6]. Half of the students authored questions and used them to prepare for a post-test, whereas the remaining students did not generate questions but were provided with the questions generated by the first group. There was no performance difference between the student groups on post-test questions that did not target material covered by the generated questions. However, students who generated study questions that were directly related to post-test items performed significantly better on those items than students who simply used the generated questions to study. As a result of this work, Foos et al. conclude that "generating potential test questions while preparing for an examination is a very effective technique".

In computer science, the relationship between exam performance and participation in a question authoring activity was previously explored by Luxton-Reilly et al. [11]. Although in their experiment student participation was self-selected, they reported a positive correlation between the number of practice questions students authored and their final exam scores. Of particular note, the authors concluded their work by encouraging other researchers to push beyond correlational studies when measuring the impact of interventions on student learning.

We present here an investigation of the causal relationship between question authoring and exam performance with a randomized controlled experiment in a large introductory programming course.

4. METHODOLOGY

In this section we provide background on the course in which our experiment was conducted and we outline details of the question authoring task and the exam that followed. We also describe how the student's questions were classified by topic and we list our two primary research questions.

At the University of Auckland, the Bachelor of Engineering degree is a four-year program. In the first year of the degree, all students take a set of compulsory courses before specializing in their second year. One of these courses is a programming course which introduces students to engineering computation in the MATLAB environment. We report here on data collected in 2012 when 729 students were enrolled in the course.

A total of three end-of-module exams were held during the semester at the conclusion of various modules, and a final exam was held at the end of the course. Our experiment concludes with the first end-of-module exam, held on the 9th of August, at the end of the fourth week of instruction. This exam consisted of 10 multiple-choice questions and contributed 5% towards each student's final grade. All students in the course sat the exam at the same time, and had 45 minutes to complete the 10 questions. The topics that had been covered in the course at the time of this exam included logical operators, conditional statements, loops, functions, matrices and matrix operations. We use the results of this exam to answer our first research question:

> **RQ1:** *Do students who generate questions as preparation for an exam perform better on individual questions, and overall, compared with students who simply use the generated questions to study?*

Students were allocated at random to one of two groups. To ensure balanced group sizes, this allocation was performed by generating a random sequence of integers between 1 and 729 inclusive, and using this sequence to partition the class list into two near-equal sized groups. One group, labeled "Authoring", contained 364 students and the other group, labeled "Non-Authoring", contained 365 students.

The task was introduced to students on the 30th of July, 11 days prior to the exam. Students in the "Authoring group" were asked to author *three* multiple-choice questions, along with an explanation of the correct answer, and submit them to an online repository prior to midnight on the 5th of August. This was an unsupervised activity to be completed outside of formal class time. Students in the "Non-Authoring" group were not able to submit questions to this repository, however all students could view and practice answering any of the submitted questions. Answering questions was voluntary, however a small amount of course credit (2%) was awarded to students in the "Authoring" group for completing the task by contributing a minimum of three questions to the repository. For fairness, students in the "Non-Authoring" group were given an equivalent opportunity later in the course, after the completion of our experiment, to earn the same amount of credit.

As a way of motivating students to practice answering the generated questions, the course instructors announced that questions appearing in the repository may "form a basis for questions in the exam". Indeed, once the question authoring deadline had passed, the course instructors reviewed all of the student-authored questions and chose a set of high-quality questions for inclusion on the exam. The researchers were not involved in the selection of these questions. Of the 10 multiple-choice questions that appeared on the exam, 9 were selected from the repository of student-authored questions and just 1 was authored by the course instructors.

Once the exam had been finalized by the course instructors, we examined the 10 questions and identified the main topic they each targeted. We then performed a similar classification of the student-authored questions using the topics identified from the exam questions. For each topic, we could

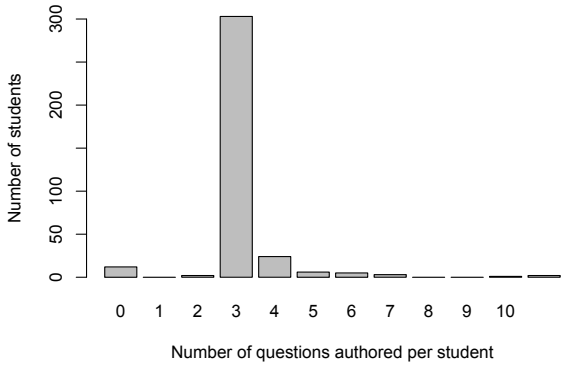

Figure 1: Histogram showing the number of questions authored by students in the "Authoring" group

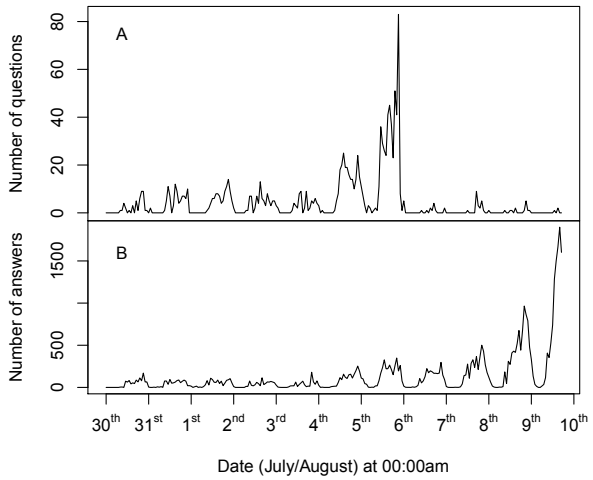

Figure 2: The number of questions authored per hour (A) and the number of answers submitted per hour (B) over the duration of the experiment

therefore partition students into three groups — those in the "Authoring" group who had authored at least one question on the topic, those in the "Authoring" group who had not authored any questions on the topic, and the remaining students in the "Non-Authoring" group (who had not authored any questions). We use this partitioning to explore our second research question:

> **RQ2:** *Do students who author questions on particular topics perform better on exam questions that target those topics compared with students who author questions on different topics, and compared with students who do not author questions?*

To support the question authoring activity in our experiment we used a web-based tool called PeerWise, which enables students to enter and format their questions directly in a web browser [4]. PeerWise also allows students to practice answering questions created by their peers and generates feedback after each attempt which includes the answer and explanation as provided by the question author, along with a summary of all previous responses to the question.

5. RESULTS

We begin this section by presenting an overview of student participation with the question authoring task. We also illustrate when questions were authored and answered with respect to the authoring deadline and the day of the exam.

5.1 Summary of authoring activity

A total of 1,133 questions were published by students in the "Authoring" group prior to the exam (6:15pm on the 9th of August). Of the 364 students in the "Authoring" group, 348 (96%) created at least one question. The overwhelming majority of these students published exactly 3 questions which corresponds to the minimum requirements for earning course credit. Figure 1 illustrates just how infrequently this minimum contribution level was surpassed.

Figure 2A shows the number of questions authored per hour over the duration of our experiment. A large spike corresponds to the deadline for earning course credit (midnight, 5th August), and more than 40% of all questions were authored sometime on the 5th of August. Authoring activity drops sharply after the deadline, with only 50 new questions added to the repository prior to the exam. Figure 2B shows, over the same period, the number of answers submitted per hour to questions in the repository. Although answering questions was voluntary, this chart exhibits a rapid growth in activity as the exam approaches. A total of 34,602 answers were submitted to the student authored questions.

5.2 Effect of authoring on exam performance

Of the 729 students who were enrolled in the course, 712 sat the exam. Overall, these 712 students performed very well, achieving an average score of 7.65 out of 10 with a standard deviation of 1.63. Only 29 students answered fewer than 5 questions correctly. Assuming a minimum exam score of 50% corresponds to a "pass", the class-wide pass rate for students who sat the exam was extremely high at 95.9%. We note however that this pass rate is consistent with that of the corresponding exam both the year before (97.2%) and the year after (96.2%) our experiment, likely due to the highly competitive nature of the course and the inherent reliability of a 10 question, 5-option, multiple-choice exam.

5.2.1 Total exam score

Our hypothesis was that students in the "Authoring" group would perform better on the exam than students in the "Non-authoring" group, due to the potential learning benefits of the question authoring activity. The exam scores achieved by students in the "Authoring" and "Non-authoring" groups are shown on the histogram in Figure 3. Note that the 17 students who did not sit the exam are recorded with a score of 0. While the distribution of marks between these two groups is similar, there are notable differences when looking at scores greater than or less than the average exam mark. For example, students in the "Non-Authoring" group more frequently scored 5, 6 or 7 (below the average mark), whereas students in "Authoring" group more often scored 8, 9 or 10 (above the average mark) .

Our data exhibits a ceiling effect as a result of the high exam scores and so we compare the two groups with a non-parametric test, the Wilcoxon rank sum test, which con-

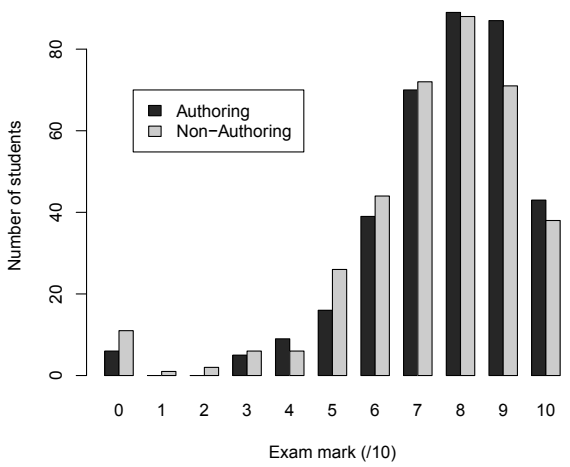

Figure 3: Histogram of exam scores by group ($n = 729$). Scores of 0 correspond to students who did not sit the exam (6 from the "Authoring" group and 11 from the "Non-Authoring" group)

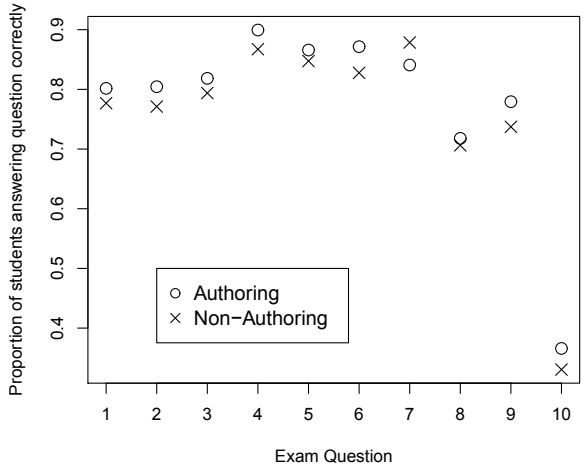

Figure 4: Proportion of students in each group answering each exam question correctly ($n = 712$)

siders only the order in which the observations from the two samples fall. In this case, our alternative hypothesis is that the exam scores achieved by the "Authoring" group are greater than those achieved by the "Non-Authoring" group. Table 1 presents the results of a formal comparison between exam scores of the two groups. Note that two tests have been conducted – one in which all students are included ($n = 729$), and another which excludes the 17 students who did not sit the exam ($n = 712$).

Although the effect sizes are small, for example the average score of students who sat the exam is just 3% higher for "Authoring" students, the distribution of scores differ significantly between the two groups. In other words, students in the "Authoring" group performed better overall than students in the "Non-Authoring" group and our experimental design implies that this difference is a result of the question authoring activity.

Table 1: Comparison of exam performance for students in the "Authoring" and "Non-Authoring" groups. Separate analyses consider all students ($n = 729$) and only those who sat the exam ($n = 712$)

n	Authoring		Non-Authoring		Wilcoxon test
	μ	σ	μ	σ	p-value
729	7.64	1.84	7.31	2.10	0.0197
712	7.77	1.57	7.54	1.68	0.0360

5.2.2 *Accuracy per question*

Analyzing student performance on each of the 10 exam questions individually provides further evidence that students in the "Authoring" group outperformed students in the "Non-Authoring" group. Figure 4 shows the proportion of students in each group that correctly answered each of the 10 exam questions. The "Authoring" group performed better on 9 of the 10 questions. The likelihood of this outcome can be tested formally with a binomial test in which the null hypothesis assumes that the probability of students in the "Authoring" group performing better on a given exam question is 0.5. The results of such a test are statistically significant ($p = 0.0107$), indicating that it is highly likely the question authoring activity had a positive impact on the exam performance of students in the "Authoring" group.

5.3 Exam performance by topic

Although question authoring has been shown to be a useful learning strategy in various disciplines, previous research has highlighted that the benefits are often limited to the specific topics targeted by the authored questions. This may help to explain the small effect sizes seen in our previous analysis. For example, some students in the "Authoring" group may have created perfectly valid questions, but none of those questions directly targeted concepts needed to answer the exam questions. Similarly, certain exam questions may have required the understanding of concepts that were targeted by only a small number of student authored questions. Either scenario may mask some of the benefits of the activity if we consider all "Authoring" students as a group.

In this section we examine groups of exam questions that target a specific topic. We investigate how well these groups of questions were answered by those students who authored questions on the same topic compared with other students who also authored questions, but on different topics.

We begin by presenting our topic-classification of the 10 exam questions. Two of the questions required students to trace through provided while loops, two of the questions involved evaluating boolean expressions consisting of both logical and relational operators, and two of the questions involved tracing through code that operated on matrices using built-in MATLAB operators such as element-wise multiplication and matrix composition. The remaining four exam questions were more complex, compound style questions that targeted multiple topics. For the purposes of this analysis we focus only on the exam questions that targeted a single topic. Table 2 shows the three primary topics we identified, lists the exam questions to which they relate, and gives a brief description of each question.

Table 2: Exam questions targeting a single topic

Topic	Q	Description of exam questions
While loops	2, 7	Tracing through while loops. In one case identifying the correct output; in the other identifying which one of the provided loops would not terminate
Boolean expressions	3, 4	Evaluating boolean expressions involving relational and logical operators. In one case tracing through provided code; in the other determining which of the provided input values would produce the given output
Matrix operations	8, 9	Tracing code that manipulates matrices. In one case transposition and element-wise multiplication of two matrices; in the other the composition of arrays through concatenation

Students in the "Authoring" group could freely select the topics on which to base their three questions. Of the 1,133 questions generated in total, 99 targeted "While loops", 99 targeted "Boolean expressions" and 124 targeted "Matrix operations" and these were authored, respectively, by 86, 87 and 106 students. Figure 5 shows how well students answered each pair of questions on the exam, based on whether or not they authored questions on the corresponding topic. Students who did not author questions at all performed almost identically to students who authored questions that did not target the exam question topics. Consistent with previous research, it appears that the topic on which questions are authored plays an important role in the effectiveness of the question generation activity.

We use a two-sample test for equality of proportions to test our hypothesis that students in the "Authoring" group who authored at least one question on a related topic performed better on the corresponding pair of exam questions than students who authored questions on different topics. Although the differences are larger than those from our earlier analyses (effect sizes of between 10% and 20% across the question pairs), the differences are statistically significant only for the "Matrix operations" topic, shown as Q8+Q9 on Figure 5 ($p = 0.0197$).

6. DISCUSSION

In our experiment, asking students in an introductory programming course to author three practice questions was sufficient to cause a measurable, but small, improvement in scores on a subsequent exam. Students randomly allocated to the "Authoring" group performed significantly better overall, and on individual questions, than students in the "Non-authoring" group who were able to review the generated questions but did not author any of their own.

We also replicated results from other domains which show that the topics on which students author their questions are important. Students who generated questions on a particular topic were more successful answering related exam questions compared with students who authored questions on different topics. We note that in our experiment topics were self-selected, and so students may have tended to author questions on topics they already understood well. Nonetheless, the act of generating questions does appear to have had a positive effect as the "Authoring" students exhibited superior performance with respect to the control group.

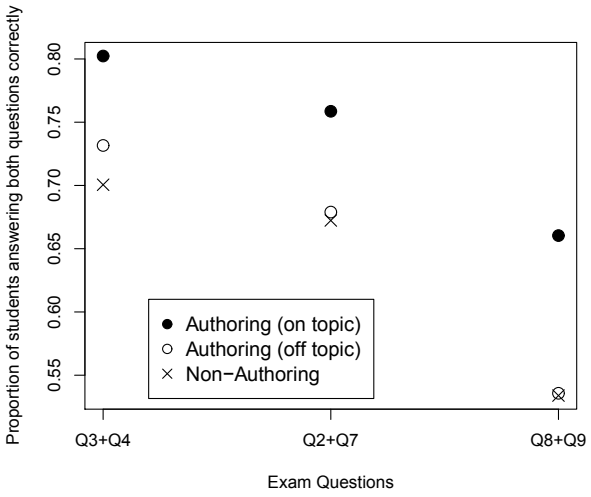

Figure 5: The proportion of students who correctly answered *both* questions on a related topic, based on whether or not they had authored a question on that topic

Despite the apparent benefits of question generation, we must acknowledge a potential disparity in the time that students spent preparing for the exam in our experiment. It is difficult to measure this time accurately, as each student adopts their own strategies and these are conducted outside of the classroom. However, at least for activities directly related to our experiment, we can approximately measure the effort and time expended by students in each group.

First, we can examine differences between the two groups with respect to the number of *answers* submitted to questions in the repository. We find that students in the "Authoring" group were more likely to answer questions than students in the "Non-Authoring" group. Of the 664 students who answered at least one question prior to the test, 345 (52%) were from the "Authoring" group and 319 (48%) were from the "Non-Authoring" group. Moreover, the 345 active "Authoring" students submitted 58.2 answers on average compared with just 45.5 answers from the 319 active "Non-Authoring" students, and these answer distributions differ significantly ($p = 0.0023$). A likely explanation for this is that students in the "Authoring" group were essentially required to log in to PeerWise to receive course credit for authoring their questions. Even though they were not required to answer any questions, doing so is relatively simple once the barrier of logging in is overcome.

We can also estimate the actual time spent by each student in the "Authoring" group creating their questions by examining log files maintained by the PeerWise tool. The median time taken to author a single question was 12.8 minutes, indicating that the majority of students in this group would have spent approximately half an hour generating their three questions. Combined with the fact that these students also tended to answer more questions, it could be argued that even once all other study activities are taken into account, the "Authoring" students spent more time preparing for the exam thus explaining their superior performance. We cannot rule this out, and a more careful measure of the efficacy of question authoring would involve having the "Non-

Authoring" students participate with a competing activity for a similar period of time. However, students are known to adopt various strategies for exam preparation, and not all are effective [8]. Our data shows that question authoring can be an effective activity, even when just a few questions are generated taking most students little more than half an hour.

The main focus of this work was to investigate the benefits to students of authoring practice questions prior to an exam. An associated benefit was the rapid creation of a large repository of questions, without any instructor effort, which became a well utilized practice resource. Practice testing is known to be a useful learning strategy, particularly when immediate corrective feedback is produced [16], for which the multiple-choice format is well suited. Recent educational research in various disciplines has shown that student generated questions are of acceptably high quality for such purposes [1, 15]. Specifically in computer science education, various tools have been developed which provide students with banks of instructor-created exercises with which to practice [14, 13]. In light of the benefits associated with generating questions, we encourage designers of such tools to consider approaches that would allow students to develop their own exercises and contribute them to these platforms.

7. CONCLUSIONS

In this work, we investigated the impact of a student-generated question activity on subsequent exam performance in an introductory programming course. Previous work involving similar activities in computer science courses have reported positive correlations between question authoring and exam performance, suggesting a relationship exists, but these studies have not been controlled.

We conducted a large-scale ($n > 700$) randomized controlled experiment, where each student in the experimental group was asked to author three practice questions prior to a summative exam. Students in the control group used these questions to prepare for the exam but did not generate questions of their own. We found that students who authored questions performed better on the exam overall, and on individual questions, and these effects were most pronounced when students answered exam questions on topics that their own questions targeted.

While our work has been limited to multiple-choice questions, we note that numerous tools have been created by the computer science education community which may lend themselves well to similar activities. We encourage designers of such tools to consider approaches that would allow students to contribute their own exercises to these platforms.

8. REFERENCES

[1] S. P. Bates, R. K. Galloway, J. Riise, and D. Homer. Assessing the quality of a student-generated question repository. *Phys. Rev. ST Phys. Educ. Res*, Jul 2014.

[2] J. Bennedsen and M. E. Caspersen. Assessing process and product: a practical lab exam for an introductory programming course. In *Proc. of the 36th Annual ASEE/IEEE Frontiers in Education Conf.*, 2006.

[3] P. R. Denner and J. P. Rickards. A developmental comparison of the effects of provided and generated questions on text recall. *Contemporary Educational Psychology*, 12(2):135 – 146, 1987.

[4] P. Denny, A. Luxton-Reilly, and J. Hamer. The PeerWise system of student contributed assessment questions. In *Proc. 10th Australasian Comp Ed Conf (ACE 2008)*, pp 69–74, Wollongong, Australia, ACS.

[5] J. Dunlosky. Strengthening the student toolbox: Study strategies to boost learning. *American Educator*, 37(3):12–21, September 2013.

[6] P. W. Foos, J. J. Mora, and S. Tkacz. Student study techniques and the generation effect. *Journal of Educational Psychology*, 86(4):567–576, 1994.

[7] L. Frase and B. Schwartz. Effect of question production and answering on prose recall. *Journal of Educational Psychology*, 67(5):628–635, 1975.

[8] R. Gurung. How do students really study (and does it matter)? *Teaching of Psychology*, 32:238–240, 2005.

[9] J. R. Lehman and K. M. Lehman. The relative effects of experimenter and subject generated questions on learning from museum case exhibits. *Journal of Research in Science Teaching*, 21(9):931–935, 1984.

[10] R. Lister, E. S. Adams, S. Fitzgerald, W. Fone, J. Hamer, M. Lindholm, R. McCartney, J. E. Moström, K. Sanders, O. Seppälä, B. Simon, and L. Thomas. A multi-national study of reading and tracing skills in novice programmers. *SIGCSE Bull.*, 36(4):119–150, June 2004.

[11] A. Luxton-Reilly, D. Bertinshaw, P. Denny, B. Plimmer, and R. Sheehan. The impact of question generation activities on performance. In *Proceedings of SIGCSE '12*, pp 391–396, New York, USA, 2012. ACM.

[12] M. D. Medley. On-line Finals for CS1 and CS2. *SIGCSE Bull.*, 30(3):178–180, Aug. 1998.

[13] A. Papancea, J. Spacco, and D. Hovemeyer. An open platform for managing short programming exercises. In *Proceedings of ICER '13*, pp 47–52, New York, USA, 2013. ACM.

[14] N. Parlante. Nifty reflections. *SIGCSE Bull.*, 39(2):25–26, June 2007.

[15] H. Purchase, J. Hamer, P. Denny, and A. Luxton-Reilly. The quality of a PeerWise MCQ repository. In *Proc. 12th Australasian Comp Ed Conf (ACE 2010)*, pp 137–146, Brisbane, Australia, ACS.

[16] H. L. Roediger and A. C. Butler. Retrieval practice (testing) effect. *In H. L. Pashler (Ed.), Encyclopedia of the Mind, Sage Publishing Co.*, pages 660–661, 2013.

[17] J. Sheard, Simon, A. Carbone, D. D'Souza, and M. Hamilton. Assessment of programming: Pedagogical foundations of exams. In *Proceedings of ITiCSE '13*, pp 141–146, New York, USA, 2013. ACM.

[18] S. Shuhidan, M. Hamilton, and D. D'Souza. Instructor perspectives of multiple-choice questions in summative assessment for novice programmers. *Computer Science Education*, 20:229–259, 2010.

[19] N. Slamecka and P. Graf. The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology*, 4:592–604, 1978.

[20] A. Venables, G. Tan, and R. Lister. A closer look at tracing, explaining and code writing skills in the novice programmer. In *Proceedings of ICER '09*, pp 117–128, New York, USA, 2009. ACM.