The Reverse Exam: A Gamified Exam Structure to Motivate Studying and Reduce Anxiety

Pablo Frank-Bolton
The George Washington University
Washington, DC
pfrank@gwu.edu

Rahul Simha
The George Washington University
Washington, DC
simha@gwu.edu

ABSTRACT

This experience report describes an attempt to improve student attitudes towards exams by encouraging students to craft exam questions that earn game points and by allowing students to defer some questions to a second attempt at the exam a week later, increasing study time while reducing common timed-test anxiety. The approach, inspired by research in gamification and student-generated questions, focuses on: getting students to study broadly across the material; encouraging students to craft good questions; encouraging an honest first attempt; preventing memorization for the second attempt; incorporating teamwork. Data collected from implementations in two different courses indicate that several finer points of the game design are important and that student-generated questions can be just as effective as instructor-generated questions. Survey data shows that students are very positive about having a second chance at learning.

KEYWORDS

Exam Protocol, Exam Design, Reducing Instructor Workload

ACM Reference Format:
Pablo Frank-Bolton and Rahul Simha. 2020. **T**he Reverse Exam: A Gamified Exam Structure to Motivate Studying and Reduce Anxiety. In *Proceedings of The 51st ACM Technical Symposium on Computer Science Education, Portland, OR, USA, March 11–14, 2020 (SIGCSE '20),* 7 pages.
https://doi.org/10.1145/3328778.3366933

1 INTRODUCTION

When we ask students why they get anxious about exams, they point out: (1) just because a professor thinks a question is easy does not mean the students find it easy; (2) the questions can come as a surprise, because students cannot guess which parts of the material the professor thinks are important; (3) it is difficult to know how to prioritize time for study, especially with a large number of topics; (4) it is hard to know how much depth-of-study is required for a topic; (5) you could have a bad day when the exam is administered. Thus, in a student's fantasy, they get to choose what's important, decide how to prioritize, and get a second shot if they have a bad day the first time. On the other side, a professor's fantasy is that students will go all out in solo study of every bit of the material, use all their social time discussing course material in groups, and immediately after an exam points out gaps in understanding, rush back to their dorm rooms to revisit the weakly-learned material. One wonders: can this gap be bridged?

In this paper, we explore setting up an exam structure with multiple incentives to reconcile the students' and professor's goals. The incentives – points earned along the way – constitute a light form of gamification. The two key elements of our approach are: (1) students craft exam questions, from which the actual exam is drawn; (2) students are given two attempts: they can defer answering some questions in the first attempt if they feel unsure. There are several questions one must address to make the approach practical:

- Q1: Do students come up with good questions?
- Q2: Do the questions cover most of the material and is there a way to spread questions and encourage studying all the material?
- Q3: Which kinds of questions get deferred?
- Q4: Does deferral in fact work and induce study for the second attempt?
- Q5: How do the student-generated questions compare with instructor-generated questions?
- Q6: Do students do better on the material for which they created questions?
- Q7: What are student perceptions of the time they spent, how much they learned, and whether they felt a reduction in anxiety given the second chance?

The paper makes the following contributions. First, our approach is a novel combination of ideas from two areas of educational research: gamification and student generation of questions (the next section places this work in the context of the literature). Second, the particular gamification incentives aim to motivate students to produce good questions, a future-work issue pointed out in a recent paper [32]. Third, this pilot study explores the above questions, with lessons learned (including negative ones) from an iterative improvement in piloting the idea in two courses. Because the exam structure is general, the approach can be applied to many types of courses.

We use the term *reverse exam* somewhat tongue-in-cheek because the structure reverses convention in two ways (students write the questions, and choose what to defer to the second attempt). This work also builds on the age-old principle that students learn well when they mimic what instructors go through, in this case, understanding enough to craft a question whose quality will be assessed.

2 RELATED WORK

Gamification. Gamification, often described as applying game rules and conventions to non-game situations [9], has long received significant attention in the education research literature [19, 20, 22,

SIGCSE '20, March 11–14, 2020, Portland, OR, USA

ACM ISBN 978-1-4503-6793-6/20/03…$15.00
https://doi.org/10.1145/3328778.3366933

25, 30] and more recently in computer science [3, 10–12, 18, 23, 26, 27]. Many studies point to increased motivation and its attendant benefits to learning [22, 25, 30], although some studies advocate caution in overstating the impact [20]. Within computer science, a recent review surveys gamification in CS education and reports overall positive impact while nonetheless stating that the field (in CS) is "in its infancy" [27]. Multiple experience reports describe positive outcomes [18, 26], with some [23] proposing to take the concept outside class. Dicheva et al. have developed a gamification platform with authoring tools [11] and recently applied this to a data structures course [12]. As mentioned, our use of gamification is elementary: of the 11 gamification elements listed in the work of Reddy et al. [27], we only use points and teamwork. Yet, it's enough to promote a degree of autonomy and social relatedness through teamwork, and aligns with the some of the principles (blocked-access, submission restriction, choice) discussed in the work by Dicheva et al. [10].

Student-generated questions. The theory behind asking students to produce questions dates back to 1975 [15]: as one might expect, it promotes deeper engagement, facilitates sense-making, and higher-order thinking [14, 32]. Since then, papers have proposed variations [2] or reported empirical studies that explore gains in learning, including some in computer science [4, 7, 8]. Denny et al. devised an online tool called Peerwise to facilitate question generation among other forms of collaboration [6] that has been used by others as well to validate the approach [17, 28, 29]. As mentioned in the work of Yu et al. [34], most of these have focused on using student-generated questions for study and preparation while few are focused on generating an entire exam. Ahn et al. [2] propose a student-generated midterm but do not evaluate. Note that some research posits that students can memorize questions [24], an issue we address in our approach.

Two-stage exams. In a two-stage exam, students first take an exam as individuals and then do the same exam questions again as a team [5]. The final score is typically some weighted combination. Empirical studies have shown the effectiveness of this form of collaborative learning [35], including in computer science [4, 33]. Yu et al. [34] describe a system to enable such exams. However, our approach is somewhat orthogonal to whether the studying is individual or collaborative. In particular, in our first stage, an individual student chooses which questions to *defer* to the second attempt because they feel additional study is warranted. Also, they don't take the questions with them after the first round and instead rely on vaguely remembering the topics of questions they deferred, which promotes broader studying than solely focusing on just the questions in the exam.

Other. A few other areas of educational research are relevant to our work. Although the general approach is not restricted to multiple-choice questions (MCQs), we preferred MCQs because they are practical and make it easier to vet student-generated questions. Scully et al. [31] show that MCQs are able to reach all but the highest two levels of Bloom's taxonomy. We explicitly instruct students via examples to lead them to craft questions at a higher Bloom-level, and include two game-point incentives for this purpose (see next section). The other relevant body of research is text anxiety, with timed-tests noted as commonly causing anxiety in technical subjects [16]. Ergene et al. [13] quantify the impact of common interventions to reduce anxiety. In this sense, the strongly positive reaction of students (survey data) affirms the form of anxiety reduction achievable through question deferral to the second attempt. Also, it has been noted that intrinsic motivation reduces test anxiety [21], which in our case comes from gamification-induced autonomy.

Contributions in the context of the literature. To summarize, the main contributions of this paper are: a novel combination of light gamification and student question-generation that aims to address the motivation factor mentioned in the question-generation literature; a new form of two-stage exams aimed at anxiety reduction (and additional study) that can be applied individually or collaboratively; and preliminary data that explores the details of the approach. Because gamification points directly affect student behavior, careful choices in the incentive system are important - we describe the structure and these incentives next.

3 METHODS

3.1 Steps in creating and conducting the exam

Phase 1: Set up. Instructor lists topics, along with rules and a sequence of deadlines. For labeling convenience, we assume the material is broadly divided into numbered modules and each module has numbered sections.

- What is critical is that the procedure is explained and the point system clarified so that students very clearly understand the incentive system.
- In our example, a good question is: a *four-choice multiple-choice question*, with choices labeled **a** through **d**, and that meets the following constraints: it has the module/section clearly identified; includes an explanation of the correct answer (with reference to the chapter material); it can be answered solely based on the chapter material (and within a few minutes); requires some thinking and is not solely depend on fact-recall; must have one unambiguously correct answer, and an explanation of how the plausibly correct but demonstrably erroneous distractors were created.
- The description should include both positive (higher Bloom-level) and negative (fact-recall) sample questions. Students should understand that they gain points for submitting good questions.
- Students are informed that a *random sample* from their submission will be vetted (graded) for quality, which encourages them to make sure all their questions are good questions.

Phase 2: Individual question crafting.

- Suppose K is the target number of questions for the exam. Each student must submit $K_I \geq K$ questions (as individuals), to make sure each student studies across the material.
- The K_I questions must include at least m questions per module to ensure spread across the material.
- All questions must not be shared at this phase.

Phase 3: Team question crafting.

- By the second deadline, students must meet in teams to pick $K_T \geq K$ questions to submit as the team submission.
- Teams can select from among the individual questions or create new ones based on their discussion.

- Teams are not allowed to share questions with other teams.

Phase 4: Question vetting and exam creation.

- The instructor randomly selects K'_I questions from each individual to grade for quality.
- With T teams and a target of K exam questions, a reasonable choice is to select $K' = K/T$ questions from each team's submission to grade for quality.
- The instructor then applies the following algorithm to select questions for the exam:

Algorithm 1: Question selection

```
let k_{m,s} be the desired number of questions for module m,
  and section s
let n_{m,s} be the number of questions generated thus far for
  m and s
let n_t be the number of questions generated so far for team t
let L_{t,m,s} be the questions submitted by team t for m and s
set n_{m,s} = 0 for all m and s
set n_t = 0 for all m and s
while any n_t < K' do
    Remove all teams t where n_t ≥ K'
    Build a list of all questions from remaining teams for all
      sections m, s where n_{m,s} < k_{m,s}
    Randomly select a question from this list and add to the
      exam
    Remove the question from the appropriate list L_{t,m,s}
    Update all variables
end
```

  As an example, in our first trial we created a 32-question exam from $T = 16$ teams, selecting $K/T = 2$ questions from each team.
- The instructor then edits the exam as needed for clarity and correctness, perhaps going back to the pool if some questions are of insufficient quality or too hard to fix.

Phase 5: First round.

- Students take the exam as individuals, and are required to answer a minimum number of questions, K_M, (they cannot defer all of them). Typically, $K_M = K/2$ (they must answer at least half).
- Both the questions and their answer sheets (with their names) are kept by the instructor, who will hand each student the same questions and exam in the second round.
- Students leave the exam without taking anything with them except for what they remember as challenging. They are of course encouraged to study for the second round.

Phase 6: Second round.

- A week or so later, students have a second round in which they complete only the deferred questions and any additional new questions (from the pool, or from the instructor's own collection). Because some questions have been answered in first round, the second round could be shorter.
- The final score is computed based on the point system described below.

3.2 The point system and rationale

The exam structure works when students clearly understand the point system and buy into the rationale. Below we describe the different types of points the student can obtain.

- Each student gets a P_i score for the quality of their individual questions, as assessed from the random sample graded.
- Each student in a team gets a P_t score for the quality of their team questions, as assessed from the randomly selected questions.
- Each student gets $p_{1,c}$ points per correctly answered question in the first round, and $-(p_{1,w}/4)$ (negative) points per incorrectly answered first-round question; they get $p_{2,c}$ points per correctly answered question in the second round, for totals of P_1 and P_2 per student. Here, $p_{2,c} < p_{1,c}$ to encourage studying for the first round, and the negative points (a lesson learned from the first course) are to ensure students do not carelessly answer questions because they don't want to study further.
- Each student gets $P_{e,c}$ points for correctly answering the extra instructor questions.
- Each student gets p_d points for every time one of their team's questions gets deferred by someone else for a total of P_d. This is a crucial incentive in encouraging students to strike a good balance between a challenging question (higher on the Bloom scale) and satisfying the requirements. If they go overboard in making the question too hard, they could lose points when the question is assessed by the instructor. Students were allowed to defer up to half of the first round questions.
- Lastly, we include p_u "uniqueness" points for each question in the topics with less coverage (possibly the harder topics), for a total of P_u points per team.

Students get excited about the notion that if their questions are strong, others will defer them as an indication of their quality.

3.3 The instructor's work and scalability

One major goal was to make the design work for instructors, ideally, with no more work than a standard exam. Assuming vetting a question takes a little less time than crafting a new one, the instructor can select as few as one random question to assess from each individual, or could restrict quality assessment solely to team questions. This, along with the K questions randomly drawn for the exam becomes the "work" for the instructor. For example, with 50 students in 16 teams and a 32-exam format, this results in 82 questions to assess (50 individual and 32 team questions), and 32 questions to carefully edit and use in the exam. Because the exam is multiple-choice, scoring is trivial even if one chooses to use partial points for some choices.

An alternative approach to reduce instructor workload while increasing the depth of the vetting process is proposed in the Discussion.

3.4 The Courses and Point construction

This protocol was tested on two upper-level CS undergraduate courses: Algorithms (ALG) and Systems Programming (SYS). Both had a similar number of students: ALG had 47 and SYS had 53. In terms of modules and sections, ALG had 7 modules and a total of 50 sections; SYS had 8 modules and a total of 51 sections.

The construction of final exam points was different for the courses since we applied the lessons learned from ALG to the SYS course. For SYS, points were given to individual submissions to motivate independent study. In addition, the negative points were included to discourage guessing and encourage strategic deferrals.

The relative weights of the quality scores in the construction of the exam points is the following:

Table 1: Relative Score Weights for SYS

Score	Maximum Weight	Components
P_i	1.5%	24 questions per individual
P_t	2.5%	20 questions per team
P_1	70%	3.5% for each **correct** answer
P_2^*	25*%	2.5% for each **correct** answer (only deferred)
P_e	15%	6 extra instructor questions
P_d	10%	up to 50 deferrals (by other students)
P_u	1%	with respect to 50 sections

Note that $P_i + P_t + P_1 + P_e + P_d + P_u$=100%. The negative points in round 1 is set to reduce the number of guesses. P_t would need to be higher if the weights are made known to students ahead of time. In the second round, correct answers give slightly less points than in the first round. This is designed to promote strategic deferrals rather than the automatic deferral of harder questions. Finally, and given that the deferral points P_d and the uniqueness points P_u are relative to the actions of their peers, we scale the exam points to a 100%.

4 RESULTS

4.1 Scaled Points in ALG and SYS

In order to answer the questions posed in Q1-Q7, we first show that the point system works similarly for both courses. We obtained a measure of the difference in results between the courses using the scaled and raw grades in the final exam. The scaled exam points for ALG (M=70.4, SD=9.2) for $N_A = 47$, and SYS (M=73.6, SD=15.7) for $N_B = 53$ were not significantly different; t(85)= -1.255, p = 0.2. In addition, we compared the distributions of raw grades and scaled points. Raw grades were calculated as the percentage of correct answers including both attempts (including student and instructor questions). For ALG, there was a significant difference in the scores for raw grades (M=77.1 SD=9.1) and the scaled points (M=70.4, SD=9.2); t(92)=3.54, p < 0.01. For SYS, we employed the Wilcoxon rank sum test, which showed there was no significant difference in the scores for raw grades (M=68.7) and the scaled points (M=73.6);W=1166, p = 0.132. Note that raw grades are biased in favor of the students since they do not distinguish between correct answer in one or two attempts. We call this the idealized traditional raw grade. The distribution of raw and scaled points are shown, for both courses, in Figure 1.

Observation 1: The scale grades are comparable between courses but in SYS, scaling the game points allowed final grades similar to an idealized traditional raw grade.

4.2 Quality of Questions

With respect to the quality of the submitted questions (Q1), we found that the quality of the questions submitted by teams was high. Grades were assigned by the Instructors and TAs as a quality rating for each question. These ratings were averaged and scaled to the range 0 to 100. For ALG, the median rating was 80/100 and for SYS the median rating was 90/100. Team quality points differed significantly according to Welch's t-test, t(17) = -3.5419, p < .01, with higher question qualities in SYS than in the ALG course.

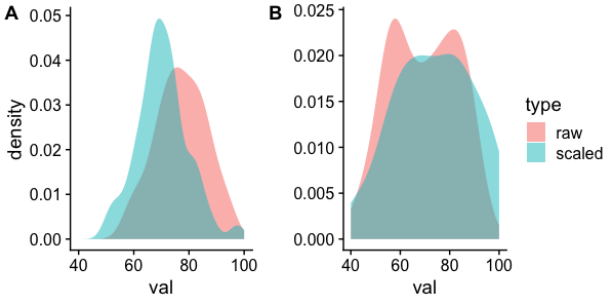

Figure 1: Raw vs scaled grades for ALG (A) and SYS (B).

Observation 2: Small changes in the protocol (inclusion of the individual question points P_i) could have resulted in a higher question quality for SYS.

4.3 Coverage

With respect to coverage of the material (Q2), we found that the requested uniform coverage by module was achieved with few issues. In a few cases, submitted problems were misclassified as belonging to a module with similar content, causing a couple of modules to have slightly more or less coverage than the average.

With respect to sections, coverage in both courses follow a similar distribution. In the ALG course all sections were covered; in the SYS course all but two sections were covered. The distribution shows the importance of implementing good coverage policies.

Observation 3: Most sections were covered, and the use of Algorithm 1 (random picking) with bad-question replacement maintains the question selection frequency of the submitted questions.

4.4 Analysis of Deferral

These sections allow us to answer which questions get deferred the most (Q3), and if students gain any advantage from the second round of study (Q4).

4.4.1 Deferral vs Correctness by Question. For SYS, we looked at which questions were deferred the most (Q3). A logistic regression was performed to ascertain the relationship between number of deferrals a question had and the proportion of correct answers it had over the set of students in the first round. The logistic regression model was statistically significant, $\chi^2(1) = 30.83$, p < .001. The model explained 13.8% (McFadden's R^2) of the variance in percentage of correct answers. Increased deferrals for each question were associated with a slight decrease in the proportion of correct answers for that question (Figure 2).

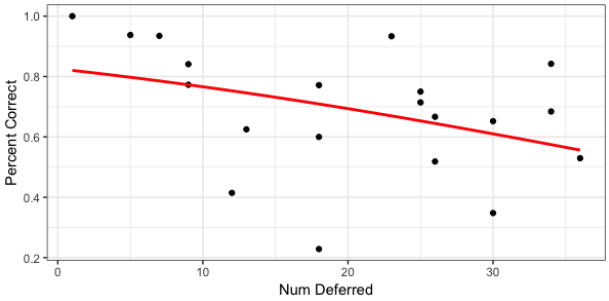

Figure 2: Correctness by Deferrals for SYS

Observation 4: Questions with higher rates of deferral had lower proportion of correct responses. We interpret these as harder questions.

4.4.2 Deferral vs Correctness. This section shows the importance of encouraging strategic deferrals. In ALG, 227 questions were deferred by 47 students (average of 4.9 per student) while in SYS, 399 questions were deferred by 53 students (average of 7.5 per student). For both courses, we performed a logistic regression to establish the relation between the number of deferred questions by student and their final proportions of correct answers. For ALG, the logistic regression model was statistically significant, $\chi^2(1) = 10.76$, $p < .01$. The model explained 4% (McFadden's R^2) of the variance in percentage of correct answers. Increased deferrals by a student was associated with a very slight decrease in the total proportion of correct answers. In the case of SYS, the model was not found to be statistically significant, $\chi^2(1) = 3.32$, $p = 0.068$.

Observation 5 : Before implementing measures to encourage the deferral of questions, students did not defer much, and performed poorly on the those questions that they deferred (presumably because there was little to gain in getting some of these few questions right).

4.4.3 Deferral by Others vs Correctness. We analyzed SYS to explore if making questions that were deferred by others had an effect on the author's results. We found no correlation between the number of deferrals a submitted question obtained, and the correctness rate of the students that generated the questions $F(1, 51)=0.0005$, $p = 0.9823$.

Observation 6: Making harder questions does not necessarily promote higher grades.

4.4.4 Number of Attempts and Correctness. Combining the above results, we can address the question of how students do when employing deferral (Q4). To determine the effect of having more than one attempt at solving a question, we compared the proportion of questions answered correctly when attempted using one opportunity with those that were deferred and then attempted in the second round. For ALG, a significant difference was found between the proportion of correct answers given a single attempt (74.1%) and given two attempts (52.4%), $\chi^2(1) = 44.989$, $p < 0.001$. For SYS, there was no significant difference in correctness with one attempt (70.1%) and two attempts (65.7%), $\chi^2(1) = 2.36$, $p = 0.1242$. Both can be seen in Figure 3

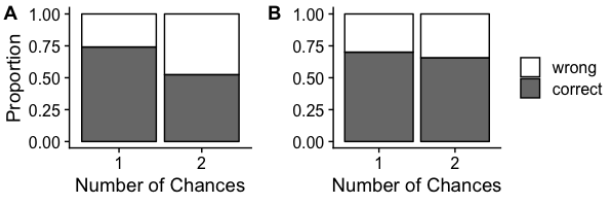

Figure 3: Correctness by attempts for ALG (A) and SYS (B).

Observation 7: After promoting strategic deferrals, in a second chance, students do as well in the most deferred (hard) questions as in the easier ones, which is suggestive that learning occurred between the two attempts.

4.5 Student vs Instructor-generated questions

We extracted the level of correctness for student vs instructor-generated questions (Q5). For ALG, a significant difference was found between the proportion of correct answers in questions created by students (77.1%) and by instructors (51.1%), $\chi^2(1) = 99.885$, $p < 0.001$. For SYS, there was no significant difference the proportion of correct answers in questions created by students (69.2%) and by instructors(67.6%), $\chi^2(1) = 0.203$, $p = 0.652$.

Observation 8: After promoting strategic deferrals, in a second chance, students do as well in the instructor questions as in the ones generated by students.

4.6 Coverage vs Correctness

We analyzed the relation between correctness rate and the coverage in the submitted question for the SYS course. This was done to verify if students do better in the sections for which they created questions (Q6). We found a significant difference between the proportion of correct answers in a section where the student submitted a question (72.4%) and the in the ones they did not cover (65.6%), $\chi^2(1) = 7.2323$, $p < 0.01$.

Observation 9: Students had higher correctness rates in sections for which they submitted questions of their own.

4.7 Student Responses

For the SYS course, students filled a survey, a 5-level Likert scale, where they gave their impressions on the protocol (Q7). The summary of results are the following: Students believe they learn more when working by themselves (mean=3.6, median= 4) than when creating the team submissions (mean=2.9, median=3), $W = 1952$, $p < 0.01$. Unsurprisingly, students assigned a very high value (mean=4.6, median= 5) of getting a second chance at answering questions. Anonymous student comments indicated that having two opportunities was the best feature of the exam. When asked about their choice in difficulty when crafting questions, students indicated their strategy was to design them with intermediate difficulty (mean=3.6, median= 3.5). Students reported spending an average of 13.7 hours studying for the first round and 5.3 hours for the second.

Lastly, student comments about the approach focused heavily on the benefit of having two attempts, with over 60% mentioning this aspect explicitly. The following are three typical student comments:

- "Two attempts did take a lot of the stress off. It was much more relaxed than one large final exam."
- "The two part exam is nice, lets you study material you didn't know."
- "Having two attempts is definitely good and allows students to study after the exam, which means you don't just study for one test and might actually learn something."

Student responses were mostly positive when discussing the question-crafting aspect of the protocol, with negative comments predictably focusing on the added work. Some examples are:

- "The building question part does help in learning."
- "Making questions pursuant to particular topics forced us to realize what the topics of the exam actually are, and to prepare accordingly."
- "Make less questions, as I believe even making 1-2 questions requires going through all the material for a certain chapter, and would likely make it easier on the teaching staff."

Observation 10: Students appreciated having a second chance at attempting deferred questions and had to devote a considerable amount of study time before the second round. In addition, they learned from creating questions of fair difficulty.

5 DISCUSSION

The overall design of the protocol was modified from the ALG course to the SYS one. We used the ALG version to note trends and student attitudes towards the question creation and deferral mechanisms. We noted that there was little incentive to employ the deferrals, and that question quality was lacking.

The changes between the ALG to SYS implementations of the protocol were put in place to increase the quality of questions and the value of a second chance at learning. To do so, we added a small incentive to create good individual questions which we found to be effective. In order to motivate strategic deferrals, we added mechanisms to penalize guessing in the first round (negative points for erroneous answers) and prevent frivolous deferrals (by having a limited number of deferrals and awarding fewer points in the second round). All of these mechanisms are set via the relative point weights. This was done empirically with the following rationale:

- Game points should reflect exam correctness. The weight awarded to correct answers (P_1, P_2, and P_e) covered approximately 85% of the grade (depending on the number of erroneous and deferred questions). Observation 1 indicates that scaled scores are not significantly lower than the scaled and weighted points.
- Question creation should receive a significant weight (up to 15% from P_i, P_t, P_d, and P_u) while balancing individual and team work. Results show that covering a section increased the odds of performing better in another question from the same section.
- Coverage should be uniform for modules and reach most sections. We added a small incentive to promote this (P_u). Observation 3 points out that this seems to be the case.
- Students should try to build questions that were fair. Since students design a large part of the exam, competitive features are put in place to balance the difficulty, like P_d and the final grade scaling.
- Sharing is not allowed unless students are in the same team. Not sharing questions increases the chances of obtaining high P_d points, which must be significant to motivate keeping the questions private. A small additional incentive is also that P_u is maximized when few students chose the same sections.

The addition of instructor-generated questions in the second round forces the student to study across concepts and not simply across question instances. This is a direct countermeasure against memorization. In that respect, the results show that students perform as well in the student questions as in the ones made by the instructor.

An important assumption we are using is that students defer questions that they feel are harder. While question difficulty might be subjective, the results support this interpretation. Using this inference, the results show that students performed as well in the second round as in the first. While initially disappointing (shouldn't they do better?), this result actually means that by using the deferral mechanism, hard questions become easier after the second round of study.

Students seem to greatly appreciate and take advantage of the second round of studying and execution, even stating explicitly that the mechanism helped reduce anxiety. This self-reported data, as well as the aforementioned leveling of difficulty with the second attempt give indirect evidence that this assessment approach may represent an alternative that helps reduce anxiety.

One issue they report is that creating the questions takes a lot of time. Both trials of the protocol featured question generation in a brief period at the end of the semester. A possible improvement is to create and vet the questions throughout the course, thereby easing both the student's and the instructor's workload. In addition, further work needs to be put into separating the effects of crafting the questions from having two attempts. An additional negative result was that additional effort in designing hard questions (those deferred the most) did not influence the author's grades.

6 CONCLUSIONS AND FUTURE WORK

The standard exam's one-shot format not only causes stress, it rarely promotes post-exam study to correct gaps pointed out by the exam. This work presents a protocol that can be used to promote re-study of gaps revealed by the exam; employ the concept of learning-by-teaching (or in this case, designing); and reduce exam anxiety by offering a second chance. The incentives need to be in place to motivate a thorough coverage of materials, an honest first attempt, and an additional study session for the second round. This protocol is designed so that students have a stake in studying effectively (for the first round), realize what they don't know (the deferred questions), and adapt to re-study to complete maximum coverage of concepts (the second round).

Future work will focus on streamlining question creation throughout the course and increasing the learning value and payoff of making good questions. In addition,the effects of the question-creation must be disentangled from the two-chances in the exam.

APPENDIX: QUESTION EXAMPLES

Here are two examples of questions submitted by students for the SYS course.

The following is an example of a good question:

Question: When pressing a button, a 2-bit increment and decrement counter is used for debouncing. At which *Time* (A, B, C or D) is the button press detected (the maximum value reached)?

Bit	0	1	0	1	1	0	1	0	1	1	1
Time			A		B					C	D

Here, the answer is **C** since that is the moment where the counter's maximum value is reached (3) including increments and decrements. All distractors are plausible (A is at the 3rd read bit; B is at the 3rd read **1**; D is at the 3rd straight **1**)

The following is an example of a poor question:

Question: What does the term "signal under-loading" refer to?

(a) The output range of a device is far smaller than the full input range of an ADC
(b) The output range is too large to be converted by a ADC
(c) The input signal is too weak to be converted by an ADC
(d) The input signal is to too strong to be converted by an ADC

Here, the answer is **A** but the question relies solely on the memorization of the term "signal under-loading", a low level in the Bloom taxonomy, and explicitly in contradiction to instructions given to the students.

REFERENCES

[1] O.O.Adesope, D.A.Trevisan and N.Sundararajan. Rethinking the Use of Tests: A Meta-Analysis of Practice Testing *Review of Educational Research*, Vol. 87(3), pp. 659 –701, 2017.

[2] R.Ahn and M.Class. Student-Centered Pedagogy: Co-Construction of Knowledge through Student-Generated Midterm Exams, *Int.J. Teaching and Learning in Higher Education*, Vol. 23(2), pp.269-281, 2011.

[3] K.A.Behnke. Gamification in Introductory Computer Science, PhD Thesis, University of Colorado Boulder.

[4] Y.Cao and L.Porter. Evaluating Student Learning from Collaborative Group Tests in Introductory Computing, *SIGCSE '17*, pp.99-104, 2017.

[5] D.Cohen and J.Henle. The Pyramid Exam, *UME Trends*, Vol. 10(2), 1995.

[6] P.Denny, A.Luxton-Reilly and J.Hamer. The PeerWise System of Student Contributed Assessment Questions. *Proc. 10th Conf. Australasian Computing Education (ACE 2008)*, Vol.78, pp.69-74, Wollongong, Australia, 2008.

[7] P.Denny. Generating Practice Questions as a Preparation Strategy for Introductory Programming Exams. *Proc. SIGCSE '15*, pp.278-283, Kansas City, MO, 2015.

[8] P.Denny, E.D.Tempero, D.Garbett and A.Petersen. Examining a Student-Generated Question Activity Using Random Topic Assignment, *ITiCSE '17*, pp.146-151, 2017.

[9] S.Deterding, D.Dixon, R.Khaled and L.Nacke. From Game Design Elements to Gamefulness: Defining "Gamification". *15th International Academic MindTrek Conference*, Tampere, 2011.

[10] D.Dicheva, C.Dichev, G.Agre and G.Angelova. Gamification in Education: A Systematic Mapping Study. *Educational Technology & Society*, pp. 75–88, 2015.

[11] D.Dicheva, K.Irwin and C.Dichev. OneUp Learning: A Course Gamification Platform, in J.Dias at al. (Eds): *Games and Learning Alliance* Springer LNCS 10653, 148-158, 2017.

[12] D.Dicheva, K.Irwin and C.Dichev. OneUp: Engaging Students in a Gamified Data Structures Course. *Proc. SIGCSE 19*, pp. 386-392, Minneapolis, MN, March 2019.

[13] T.Ergene. Effective Interventions on Test Anxiety Reduction: A Meta-Analysis *School Psychology International*, Vol. 24(3), pp.313-328, 2003.

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

[15] L.T.Frase and B.J.Schwartz. Effect of question production and answering on prose recall. *J. Edu. Psychology*, 67(5): 628-635, 1975.

[16] E.Geist. The Anti-Anxiety Curriculum: Combating Math Anxiety in the Classroom *Journal of Instructional Psychology*, Vol. 37, No. 1.

[17] J.Hardy, S.P.Bates, M.M.Casey, K.W.Galloway, R.K.Galloway, A.E.Kay, P.Kirsop and H.A.McQueen. Student-Generated Content: Enhancing learning through sharing multiple-choice questions, *Int.J. Science Education* Vol. 36(13), 2014.

[18] A.Iosup and D.Epema. An experience report on using gamification in technical higher education. *Proc. SIGCSE 14*, pp.27-32, Atlanta, Georgia, March, 2014 .

[19] J.Hamari, D.J.Shernoff, E.Rowe, B.Coller, J.Asbell-Clarke and T.Edwards. Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning, *Computers in Human Behavior* Vol. 54, pp. 170-179, 2016.

[20] M.D.Hanus and J.Fox. Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance *Computers & Education* Vol. 80, pp.152-161, 2015.

[21] R.Khalaila. The relationship between academic self-concept, intrinsic motivation, test anxiety, and academic achievement among nursing students: Mediating and moderating effects, *Nurse Education Today* Vol. 35(3), pp.432-438, 2015.

[22] M.Lister. Gamification: The effect on student motivation and performance at the post-secondary level. *Issues and Trends in Educational Technology*, 3(2), 2015.

[23] M.Mejias, K.Jean-Pierre, Q.A.Knox, E.Ricks, L.Burge and A.N.Washington Meaningful Gamification of a Computer Science Department: Considerations and Challenges, *Int'l Conf. Frontiers in Education: CS and CE*, FECS, 2015.

[24] T.Papinczak, R.Peterson, A.S.Babri, K.Ward V.Kippers and D.Wilkinson. Using student-generated questions for student-centred assessment, *J. Assessment & Evaluation in Higher Education*, Vol.37(4), 2012.

[25] T.A.Papp. Gamification Effects on Motivation and Learning: Application to Primary and College Students. *Int. J. Cross-Disciplinary Subjects in Education (IJCDSE)*, Vol. 8, No. 3, September 2017.

[26] M.Piteira, C.J.Costa, M.Aparicio. Computer Programming Learning: How to Apply Gamification on Online Course. *J.Information Systems Engineering & Management*, Vol. 3(2), pp. 2468-4376, 2018.

[27] M.Reddy, G.S.Walia and A.D.Radermacher. Gamification in Computer Science Education: a Systematic Literature Re- view. *ASEE Annual Conf*, 2018.

[28] S.M.Rhind and G.W.Pettigrew. Peer Generation of Multiple-Choice Questions: Student Engagement and Experiences, *J. Veterinary Medical Education*, Vol. 39(4), 2012.

[29] G.W.Rieger and C.E.Heiner Examinations That Support Collaborative Learning: The Students Perspective, *J.College Science Teaching*, Vol. 43(4), 2014.

[30] M.Sailer, J.U.Hense, S.K.Mayr, and H.Mandl. How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction *Computers in Human Behavior* Vol.69, pp.371-380, 2017.

[31] D.Scully. Constructing Multiple-Choice Items to Measure Higher-Order Thinking *Practical Assessment*, Vol. 22(4), May 2017.

[32] D.Song. Student-generated Questioning and Quality Questions: A Literature Review em Research Journal of Educational Studies and Review, Vol. 2(5), pp.58-70, 2016.

[33] B.Yu, G.Tsiknis and M.Allen. Turning Exams Into A Learning Experience, *SIGCSE '10*, Milwaukee, WI, 2010.

[34] F-Y.Yu and C-L.Su. A student-constructed test learning system: The design, development and evaluation of its pedagogical potential. *Australasian Journal of Educational Technology*, 31(6). 685, 2015.

[35] J.F.Zipp. Learning by Exams: The Impact of Two-Stage Cooperative Tests, *Teaching Sociology*, Vol. 35(1), pp. 62-76, 2007.