Investigating the Effects of Testing Frequency on Programming Performance and Students' Behavior

David H. Smith IV
University of Illinois
Urbana, IL, USA
dhsmith2@illinois.edu

Chinedu Emeka
University of Illinois
Urbana, IL, USA
cemeka2@illinois.edu

Max Fowler
University of Illinois
Urbana, IL, USA
mfowler5@illinois.edu

Matthew West
University of Illinois
Urbana, IL, USA
mwest@illinois.edu

Craig Zilles
University of Illinois
Urbana, IL, USA
zilles@illinois.edu

ABSTRACT

We conducted an across-semester quasi-experimental study that compared students' outcomes under frequent and infrequent testing regimens in an introductory computer science course. Students in the frequent testing (4 quizzes and 4 exams) semester outperformed the infrequent testing (1 midterm and 1 final exam) semester by 9.1 to 13.5 percentage points on code writing questions.

We complement these performance results with additional data from surveys, interviews, and analysis of textbook behavior. In the surveys, students report a preference for the smaller number of exams, but rated the exams in the frequent testing semester to be both less difficult and less stressful, in spite of the exams containing identical content. In the interviews, students predominantly indicated (1) that the frequent testing regimen encourages better study habits (e.g., more attention to work, less cramming) and leads to better learning, (2) that frequent testing reduces test anxiety, and (3) that the frequent testing regimen was more fair, but these opinions were not universally held. The students' impressions that the frequent testing regimen would lead to better study habits is borne out in our analysis of students' activities in the course's interactive textbook. In the frequent testing semester, students spent more time on textbook readings and appeared to answer textbook questions more earnestly (i.e., less "gaming the system" by using hints and brute force).

CCS CONCEPTS

• **Social and professional topics → Student assessment**.

KEYWORDS

testing frequency, CS1, assessment

SIGCSE 2023, March 15–18, 2023, Toronto, ON, Canada

ACM ISBN 978-1-4503-9431-4/23/03…$15.00
https://doi.org/10.1145/3545945.3569821

ACM Reference Format:
David H. Smith IV, Chinedu Emeka, Max Fowler, Matthew West, and Craig Zilles. 2023. Investigating the Effects of Testing Frequency on Programming Performance and Students' Behavior. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1 (SIGCSE 2023), March 15–18, 2023, Toronto, ON, Canada.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3545945.3569821

1 INTRODUCTION

For many teachers, summative assessment is a necessary evil. We primarily care about our students learning, but we need to measure their knowledge and skills in order to produce grades. Summative assessment and how it is organized, however, can positively contribute to how much our students learn through two mechanisms. First, the act of trying to recall information can improve the long-term retention of that information, a phenomena commonly referred to as the "testing effect" [15, 28]. Second, the existence of the summative assessment can shape student behavior, leading them to practice the material in instances where they otherwise might not [1, 23]. These two mechanisms suggest that increasing the frequency of testing could improve student learning.

The impact of various testing regimens was actively studied in the late 20th century. The findings of these studies (discussed in Section 2) indicated there were benefits from more frequent assessment, but the impact on performance are varied. A meta-analysis by Bangert-Drowns et al. [2] found that students perform better when assessments are divided into more smaller tests, but increased testing frequency doesn't always translate into better final exam performance. There are, however, no studies on the impact of testing frequency in computer science to our knowledge.

In this paper, we study the impact of testing frequency in an introductory programming (CS1) course. There is a reason to believe that frequent testing might be more valuable in CS1 courses, due to the unusually tight coupling of concepts [27] and due to learning challenges and high failure rates [18, 25].

Specifically, we seek to address two research questions:

RQ1: How does performance on the same test items differ between semesters with different testing frequency policies?
RQ2: How are students' behavior, perceptions of learning and fairness, and test anxiety impacted by testing frequency?

We addressed the research questions by performing an across-semester quasi-experiment that compares two offerings of a large

enrollment introductory Python programming class where the only aspect of the course that was changed was the testing regimen. Because the course uses pools of questions to generate unique exams for each student, the same exam questions could be used in both semesters, enabling a comparison of student performance. Course details and the experimental design are presented in Section 3.

This paper makes three main contributions. First, it measures the benefits of frequent testing in a novel subject area relevant to this community (Section 4.1). Second, it couples student learning results with surveys (Section 4.2) and student interviews (Section 4.3) to shed light on students' perceptions of the impact of testing frequency on study behavior, learning, test anxiety, and fairness. Lastly, we provide the first (to our knowledge) large-scale measurements of student study behavior as influenced by testing frequency in the form of an analysis of student interactions with an interactive online text book (Section 4.4).

2 LITERATURE REVIEW

The majority of the early empirical work on frequent testing was largely positive with multiple studies indicating more frequent testing leads to higher performance [21, 30]. One of the earliest studies [16] took a bank of questions and divided them into weekly assessments for one group and monthly assessments for the other. They found that students given weekly assessments outperformed those given monthly assessments in terms of overall test performance and performance on a retention assessment that was given to both groups. Similar experiments that divide set banks of questions into smaller, more frequent assessments found similar results, with the students who were more frequnetly tested answering more questions correctly than their less frequently tested counterparts [8, 11]. As noted by Foss and Pirozzolo [9], this may be partially due to those under the frequent testing condition being tested on material closer to when they learned it and the smaller size of the more frequent assessments.

The impact of frequent testing on final exam performance is more mixed. Courses with some midterm exams/quizzes outperform those with none on final exams by 0.57 standard deviations [2]. When comparing some number of midterms to another number of midterms, most studies have found no significant difference in final exam performance [4, 6, 7, 10, 13, 16, 32] with a few exceptions where more frequent testing lead to better performance [17, 22]. Fulkerson and Martin [10] suggests that the week prior to the final assessment was sufficient for the students in their infrequent condition to catch up to their peers in the frequent condition and thus reduce the gap in knowledge between the two groups.

2.1 Perceptions and Behavior

Gaynor and Millham [11] and McDaris [24] found that students who were tested weekly expressed more enjoyment and perceived learning from their testing frequency compared to students receiving just a midterm and a final. Holmes [14] found that employing the use of weekly, low-stakes (1%) e-assessments was associated with perceived increases in engagement and learning. Results from anxiety questionnaires given to students undergoing different assessment frequencies suggest that increasing the number of assessments reduces anxiety [8], and this reduction may disproportionately benefit highly test anxious students [10]. A survey by Vaessen et al. [31] found that students valued the more frequent assessments both as a motivator to study and as a diagnostic tool they could use to learn and improve. While frequently-tested students rate their classes higher on average [2], a study also found that more frequent testing increased the stress and reduced the self-confidence of students that were performing poorly on the assessments [31].

Empirical work has evaluated how the frequency of testing impacts student behavior by storing study materials exclusively in a room that could be monitored through a one-way mirror [23]. This small-scale ($N = 8$) within-subjects study found that daily testing produced consistent study patterns compared to tests that occurred weekly and once every three-weeks. When students were placed in the latter two conditions they studied in bursts that occurred prior to their tests but spent roughly the same amount of time studying overall. Anderson [1] had students ($N = 10$) log their amount of studying in a cross-over study with and without weekly quizzes. They found that students studied more in weeks that they had weekly quizzes, but both conditions had noticeable increases in studying before major exams.

3 STUDY DESIGN

We present a quasi-experimental, between-subjects study comparing student performance on assessments between two semesters of an introductory Python programming course for non-CS majors. The semesters were taught identically and by the same instructor, except for the frequency of summative assessment. We will refer to the semesters as *baseline* (Spring 2022) and *frequent* (Fall 2021). Both semesters had a large population with 540 students in the baseline semester sitting the first exam and 488 of those continuing through the course all the way to the final. The frequent semester saw 727 students take the first exam and 671 of those continuing on to sit the final exam. In both semesters, the students are predominantly first year students (Baseline=65.6%, Frequent=70.6%) and business is the most represented major (Baseline=38.1%, Frequent=65.6%).

The course involves five main components: 1) weekly readings to be completed before lecture using an interactive textbook from zyBooks, where credit was earned by completing multiple-choice and short-answer participation activities, 2) peer instruction-based lectures, 3) weekly two-hour active learning lab sections, 4) weekly web-based homework, and 5) computer-based exams. The homework and exams use similar questions and the same platform (PrairieLearn [33, 34]); both include a broad range of question types, including true-false, multiple-choice, tracing, syntax questions (where students write a single line of code), Parsons problems, code fixing questions, and programming questions. Weekly homework consists of a mix of 20–40 questions of these types.

Computerized exams are taken in a computer-based testing facility [35], a locked-down computer lab with proctors present. Because students take their exams at different times due to the size of the class, the computerized exams are generated randomly from a collection of question pools [36]. Each pool consists of questions of the same type that test the same learning objective at a similar difficulty. The exam is graded interactively and students are allowed to attempt questions multiple times for reduced credit, as configured by the instructor.

(a) Baseline - Testing Schedule

(b) Frequent - Testing Schedule

Figure 1: The testing schedules for each of the two semesters. Red cells indicate a week with no tests, light green is a week with a quiz, dark green is a week with an exam.

To familiarize them with this computerized exam format, students in both semesters were given an "Exam 0" in the third week of instruction that is worth only 2% of their final grade. Because both semesters were identical up to this point and the contents of Exam 0 was identical, this exam allows us to compare the student populations across semesters. Students in the baseline semester scored slightly better on Exam 0 (baseline: μ=88.1, σ=10.4; frequent: μ=87.2, σ=9.7), but this difference was not statistically significant. Since the results show the frequent semester performing better on the rest of the exams, these Exam 0 scores suggest that the primary effect that we're seeing is not due to population differences.

After Exam 0, the two semesters differed in their frequency of summative assessment, as shown in Figure 1. Both semesters included five hours of exams, but distributed differently over the semester. The baseline semester included a single 2-hour midterm (20% of final grade) and a 3-hour final exam (30%). The frequent semester had three 1-hour midterm exams (each worth 10%) and a 2-hour final exam (20%). The baseline semester's midterm was comprised of the pools of questions used in the frequent semester's first and second exams (E1, E2). Similarly, the baseline's final exam question pools corresponded to those on the frequent semester's third midterm and final exam (E3, E4). During both semesters, practice exams with similar structure to the exams were released in the week prior to the exam.

Outside of exams, the frequent semester had four self-proctored (formative) quizzes (worth 2% each), one in each week preceding an exam. In the baseline semester, students were offered "self-assessment" quizzes every two weeks, but these quizzes weren't for credit.[1]

With IRB permission, we received access to analyze the anonymized student data associated with the course and publish it in aggregated forms. Students were informed that their data would be used in this way and given an opportunity to have their data excluded from the analysis; few did. In addition to comparing exam scores (Section 4.1), we surveyed both course offerings, interviewed a modest number of students from each semester, and analyzed the textbook behavior of students. We present the details of these methods as part of the results in Sections 4.2, 4.3, and 4.4, respectively.

4 RESULTS

4.1 Item Performance

The students in the frequent testing condition performed better than the students in the baseline section. We first show this through comparisons of raw exams scores. In Figure 2, we show average exam scores for both semesters. Note that the frequent semester's exam 1 and exam 2 scores had higher averages than the baseline semester's midterm, which covers the same material using largely the same pools of questions as exams 1 and 2. Likewise, the frequent semester's exams 3 and 4 had an average about 10% higher—a full letter grade—than the baseline semester's final that is composed of same pools of questions.

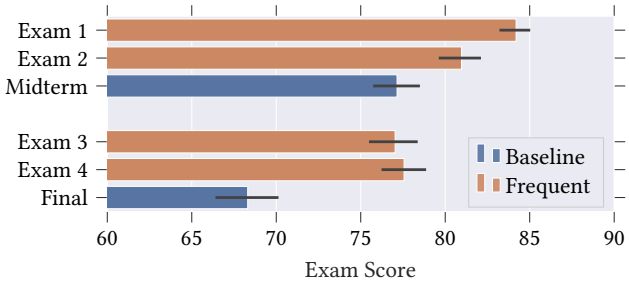


Figure 2: Exam averages with 95% confidence intervals

These raw exam scores, however, are a rather coarse instrument for comparing performance, because the baseline semester's exams are **not** merely the concatenation of the frequent semester's exams. Instead, a more apples-to-apples comparison is to compare performance at the granularity of individual questions. Because the question types follow different performance trends and use different grading procedures, we present our analysis at the granularity of question type, focusing on three types—programming, syntax[2], and tracing—that test the bulk of the course's core learning objectives.

For each question type and group of exams[3], we fit the following ordinary least squares (OLS) regression:

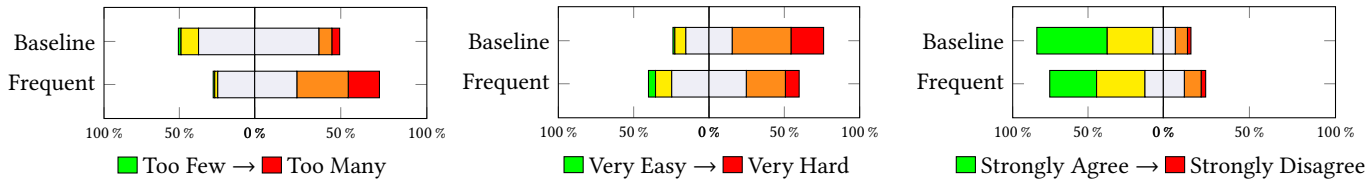
$$QuestionScore = \sum_{i=0}^{n} \beta_i Exam_i + \sum_{j=0}^{m} \alpha_j QuestionName_j \quad (1)$$

where $QuestionScore$ is the percentage score a student received on a question. $Exam_i$ and $QuestionName_j$ are categorical dummy variables that indicate the exam the question was on—which implicitly tells us which semester it is from—and the specific question. The regression estimates two sets of regressors. The β_is captures the relative performance of the question type by exam. The α_js estimate the average difficulty of a given question which, while not useful independently, improves the regression's estimate of β_i.

[1] In the baseline semester, the other formative activities (e.g., textbook, lecture participation, homework) are worth more. Since we are comparing exam scores and not overall grades, this difference should not be important.

[2] These code writing questions involve students writing a single line of code.

[3] The regression is run twice; once for exams E1, E2, and Midterm and once for exams E3, E4, and Final.



(a) The number of exams in the course is? **(b)** The computer based exams are **(c)** I find the quizzes and exams very stressful.

Figure 3: Survey responses as collected from informal early feedback forms given in each semester.

Table 1: Percentage points that E1 and E2 scores were higher than Midterm scores

	E1	E2	R-Squared
Programming	0.33	11.00***	0.135
Syntax	3.54***	9.10***	0.058
Tracing	2.31**	8.06***	0.102

Table 2: Percentage points that E3 and E4 scores were higher than Final scores

	E3	E4	R-Squared
Programming	-0.79	9.77***	0.224
Syntax	9.00***	13.48***	0.094
Tracing	1.33	4.31***	0.132

In Tables 1 and 2, we present how much better the frequent semester's students did than the baseline semester. The tables present how many percentage points higher the frequent semester students performed on a given question type relative to the baseline semester students did on the corresponding exam. For example, Table 1 indicates that frequent semester students scored 11 percentage points higher on programming questions found on E2 than the baseline semester students did on the corresponding questions when they took them on their midterm.

When we compare exams that occur at roughly the same time in the semester (i.e., E2 and Midterm; E4 and Final) students are consistently scoring higher in the frequent testing semester to a statistically significant degree for all question types. Effect sizes for E2 vs. midterm are 0.29, 0.29, and 0.27 and for E4 vs. final are 0.27, 0.43, and 0.14 for programming, syntax, and tracing questions, respectively.

In addition, we see the frequently-tested students score statistically significantly higher on some questions when they are taken three weeks earlier in the semester. For example, frequently-tested students score an average of 9 percentage points higher on syntax questions found on E3 than baseline semester students do *on those same questions* when they take them three weeks later on the final exam. In other question types, the frequent testing semester students performance is statitically equivalently to that of the baseline semester students three weeks later, which still seems noteworthy.

4.2 Survey Results

Students in both semesters were given informal feedback surveys as a part of routine course procedure. Three questions from those surveys are of interest to this work, those relating to perceptions of the quantity, difficulty, and stressfulness of the exams. 5-point Likert items were used. Due to an oversight by the instructor, the surveys were delivered differently in the two semesters. In the frequent semester (Fa21), the survey was given using SCANTRON forms during lab section between exams 2 and 3, for no credit. In the baseline semester (Sp22), the survey was given using a Google form in the last week of classes, with a .2% overall course grade incentive for completion. In both semesters, the majority of students completed the surveys (baseline = 81.1%, frequent = 70.6%).

Students reported somewhat inconsistent perceptions. Students report that the two exams in the baseline semester is the appropriate number of exams and that the frequent testing semester tended toward too many exams (Figure 3a). They reported, however, that the exams in the frequent semester were easier (Figure 3b), in spite of the fact that they are made up of the same material. This perception that the exams were easier does correlate with their higher performance on the questions. They also report the exams to be slightly less stressful in the frequent semester (Figure 3c). All of these differences are statistically significant (all $p<0.001$) according to the results of a series of Mann-Whitney U tests between each pair of responses.

4.3 Interview Results

With IRB permission, interviews were performed with students from both the baseline ($N = 12$) and frequent ($N = 7$) semesters, in the summer after both semesters had completed. Students were recruited by email, provided informed consent, and compensated with an Amazon gift card at a rate of $15/hour. These semi-structured interviews were conducted and recorded via Zoom. Students were presented with the baseline and frequent testing schedules side by side and asked to compare the two testing frequencies using each of the following criteria:

- in how they would affect your approach to studying?
- in how they would contribute to your ability to learn the course content?
- in terms of stress and test-related anxiety?
- in terms of fairness?

The interviews were transcribed and inductively coded by two researchers. The codes were then reconciled and consolidated to 14 codes for validation purposes. The two researchers and an external coder re-coded 4 randomly selected transcripts. Inter-rater reliability was computed with Krippendorff's alpha [19]. We examined whether raters had assigned the same codes for a student's entire response to a question, i.e. the student's turn in the conversation. The Krippendorff alpha value ($\alpha = 0.86$) was found to be satisfactory [20].

Themes were extracted by grouping inductive codes together [12], the most prevelant of which are discussed below.

Studying and Learning: We identified three theme groups related to the impact on studying and learning. First, a significant number of students indicated that infrequent testing leads to sub-optimal studying behaviors, like procrastination and massed practice (i.e., cramming). The students reported that these behaviors led to diminished retention of knowledge. One student noted, "*If I only had like a midterm and a final, I'd probably just start reviewing the week before, and almost just start cramming for it. And I don't know if I would retain the material as well if I . . . had some sort of pressure point to make sure I knew the material.*" Another student indicated, "*I feel like students with the [infrequent testing] will be more likely to not recall what they've been learning as much while they're studying, but instead, they'll try to cram a lot before the exam, which I'm guilty of, so I understand.*".

In contrast, many students reported that frequent testing promoted good study habits and learning. Students reported more frequent testing leads to more reaching out for help, increasing their studying, increasing their usage of the textbook, and generally paying more attention to the course and it's content. When considering the frequent testing schedule one student remarked, "*I would have made sure to... when I'm [working] through zyBooks and stuff, I would make sure that I'm actually understanding what I'm doing instead of just trying to get the participation points.*". Furthermore, students remarked that frequent testing required them to constantly review concepts taught in the course, and it improved their ability to recall information. One student noted, "*People still need to learn the same amount of material, but I feel like with the frequent testing, they're more likely to retain what they learned*".

Finally, students indicated that the frequent exams lead to a more manageable workload, often by distributing the work more evenly across the semester. For example, one student said, "*I probably would have been forced into more review and, thus, a little bit more solid understanding in a week by week basis*".

Stress and Anxiety: Most students suggested that having more exams made the exams less stressful. Students frequently noted that with only two exams in the baseline semester, both exams were high stakes which led to them being stressful. As one student stated, "*I think the weight of the midterm and the weight of the final, because there's much less exams, because they're worth so much more, that can add a lot of pressure to the student.*" Students also frequently indicated that frequent assessments reduced the anxiety of any one exam, because each exam has a smaller impact on their grade. As one student noted, "*I feel like the frequent testing would be less stressful just because it's weighted less*," and another stated, "*So, I think that frequent testing kind of spreads out the stress a little bit more than just a midterm and final.*".

This sentiment, however, was not universal. A few students stated that frequent testing could cause more stress overall because there were more exams or the exams occurred more frequently. One student noted, "*I think the [frequent testing] might be a little more stressful, just because seeing the syllabus or the curriculum, right in the first week of school, might be a little overwhelming, seeing the amount of exams and tests there are.*".

Fairness: Perspectives were somewhat mixed when students were comparing the fairness of the baseline and frequent assessment. The dominant perspective, by a small margin, was that frequent testing was more fair because it reduced the likelihood that low performance on a single assessment would irreversibly depress a student's grade. Students also stated that frequent testing provided more opportunities to demonstrate an understanding of course concepts. As one student stated, "*The frequent testing is a lot more fair, because you have a bunch of opportunities, to really test your knowledge. . . [with infrequent testing], you really only have two opportunities, which is the midterm and the final. And if one of those doesn't go according to plan, then no matter what, you're not going to get a good grade in the class, whereas [with frequent testing], maybe if one of the exams doesn't go that well, you still have a pretty high chance of getting the grade that you want in the class.*".

A slightly smaller number of students indicated that fairness was independent of testing frequency. Students expressing this sentiment explained that learners were responsible for their performance on assessments however they are structured. For example, one student said, "*I think both [frequent and infrequent testing] are fair. . . . If we're tested on material that we were actually taught by the professors, then that is fair.*"

4.4 Textbook Engagement

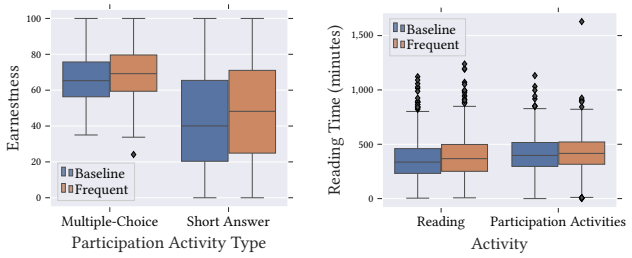
Motivated by this remark from Section 4.3 and others like it,

> "(With frequent exams) I would have made sure to... when I'm [working] through zyBooks and stuff, I would make sure that I'm actually understanding what I'm doing instead of just trying to get the participation points."

we wanted to see if we could observe this behavior more generally in the textbook usage of the two semesters. Specifically, we wanted to measure the degree to which students appeared to use the book exercises earnestly to learn as opposed to trying to earn the points with the least expended effort. The online textbook used two kinds of participation exercises: the multiple-choice questions can be brute-forced by selecting each option in turn, because the questions are static and there is no penalty for wrong answers, and the short answer questions have a help feature that permits students to show themselves the answer.

In Figure 4a, we compare the *earnestness* of students in each semester for each of the two kinds of questions. Because brute-forcing students typically rapidly select options until the answer is found, we operationalize earnestness for multiple-choice questions as the number of questions completed where the second response occured more than one second after an initially incorrect response. For short answer questions, we operationalized earnestness as the number of short answer questions that students answered without clicking the "show answer" button after one incorrect attempt.

Students in the frequent testing semester had statisitically significantly higher earnestness scores than those in the baseline semester. For multiple choice questions, the effect size was 0.24 (Frequent: μ=69.4, σ=13.4; Baseline: μ=66.2, σ=13.3). For short answer questions, the effect size was 0.20 (Frequent: μ=48.8, σ=27.4; Baseline: μ=43.3, σ=26.8). Significance was evaluated using an independent



(a) Earnestness **(b)** Time-on-Task

Figure 4: Students in the frequent semester were statistically significantly more "earnest" in their work in the text book (a) and in the amount of time they spent reading the textbook (b).

samples t-test, which indicated these results to be highly significant (t(1344)=3.68, p<0.001 and t(1344)=4.32, p<0.001, respectively).

In addition, Figure 4b compares the time students spent reading and doing participation activities in the textbook. Students in the frequent testing semester spent more time than the baseline semester on reading the textbook. This effect size was 0.12 (Frequent: μ=386.0, σ=195.0; Baseline: μ=362.6, σ=192.4) and statistically significant (t(1344)=2.18, p<0.05). There was no statistically significant difference in the time spent on participation activities (Frequent: μ=410.8, σ=182.2; Baseline: μ=403.5, σ=185.6, t(1344)=0.72, p>0.05).

5 DISCUSSION

The improvements on the final exam observed in Section 4.1 are somewhat inconsistent with the findings of most studies described in Section 2. While most prior studies found no statistically significant impact on final exam scores by varying the number of mid-term exams [2], the impact in our study was highly significant ($p < .001$) with effect sizes around 0.3 in most cases. We suspect these difference arise from the highly cumulative nature of the material in CS1 courses. Where many college courses cover a range of loosely coupled topics that may be forgotten and have to be relearned for a final, the key learning objective in a CS1 course—how to program—is presented as series of modules that each build on the previous one. If increased testing frequency results in better learning of each unit, the increased mastery of each unit potentially facilitates learning the next unit's worth of material resulting in higher scores at the end of the semester.

The results of the interviews and analysis of textbook interactions suggest that the differences in performance may be partially explained by students engaging in more consistent and deliberate practice. In particular, students routinely stated that, under the baseline testing frequency, they believed they would engage in more massed practice, otherwise referred to as "cramming". Such study patterns can lead to poor long term retention [3, 29]. Increasing the number of tests appears to encourage students to prepare on a more regular basis rather than engaging in a small number of massed practice events. Though the analysis presented in Section 4.1 focuses on the high-stakes exams, the low stakes quizzes that preceded each exam may also have played an important role in incentivizing students to engage in more consistent studying.

One of the most surprising results for us were the exam frequency survey results presented in Section 4.2. By all other indications students seem to prefer more frequent testing: both surveys and interviews suggest that more frequent testing is less stressful for most students, survey results suggest students find the material easier when split among more exams, and students report in the interviews that the more frequent testing regimen is more conducive to learning. In spite of all this, however, students report that the frequent testing regimen has too many tests. Two explanations seem plausible. First, that fewer exams would mean less work, or a little like a failure on a Piagetian conservation task [26], students might expect that fewer exams would require less preparation in spite of the same amount of material being covered in both courses. Second, that students aren't reliable at choosing the pedagogy that best serves their learning, seen, for example, in preference for passive lecture over active pedagogies [5]. If this second explanation is true, students are advocating for fewer exams in spite of their belief expressed in the interviews that frequent testing improves their learning.

6 LIMITATIONS

This work has two primarily limitations. First, differences in the composition of the two student populations could have impacted the results. While the similar demographics between the two offerings and their statistically-equivalent performance on Exam 0 provide some assurances, the quasi-experiment is not a randomized controlled trial and should not be interpreted as such.

Second, as is common in qualitative research, the number of students interviewed is modest (N = 19), and the students are self-selected. As such, their responses may not be entirely representative of the sentiment of the student population as a whole. That said, the survey results (Section 4.2) and textbook analysis (Section 4.4) corroborate a number of the interview findings.

7 CONCLUSION

In this work, we present the results of a quasi-experiment comparing a frequent testing schedule and baseline (midterm + final) testing schedule in an introductory Python programming course. Using multiple lenses—linear regression of performance data, surveys, interviews, and textbook analysis—we find that students perform better in the frequent testing scenario, which appears to be the result of the frequent testing regimen leading to better studying behavior that translates into better learning. In addition, we find that students predominantly report that the frequent testing regimen produces less test anxiety and is more fair.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1915257 and 2013334. We also would like to thank Chelsea Gordon and zyBooks for development of the earnestness metrics used in this paper.

REFERENCES

[1] Jean E. Anderson. 1984. Frequency of Quizzes in a Behavioural Science Course: An Attempt to Increase Medical Student Study Behavior. *Teaching of Psychology* 11, 1 (1984), 34–34. https://doi.org/10.1207/s15328023top1101_7

[2] Robert L Bangert-Drowns, James A Kulik, and Chen-Lin C Kulik. 1991. Effects of frequent classroom testing. *The journal of educational research* 85, 2 (1991), 89–99.

[3] Kristine C Bloom and Thomas J Shuell. 1981. Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *The Journal of Educational Research* 74, 4 (1981), 245–248.

[4] D William Deck Jr. 1998. *The effects of frequency of testing on college students in a principles of marketing course.* Ph.D. Dissertation. Virginia Polytechnic Institute and State University.

[5] Louis Deslauriers, Logan S McCarty, Kelly Miller, Kristina Callaghan, and Greg Kestin. 2019. Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences* 116, 39 (2019), 19251–19257.

[6] Parsla Dineen et al. 1989. The Effect of Testing Frequency upon the Achievement of Students in High School Mathematics Courses. *School Science and mathematics* 89, 3 (1989), 197–200.

[7] Samuel D Downs. 2015. Testing in the college classroom: Do testing and feedback influence grades throughout an entire semester? *Scholarship of Teaching and Learning in Psychology* 1, 2 (2015), 172.

[8] David S Dustin. 1971. Some effects of exam frequency. *The Psychological Record* 21, 3 (1971), 409–414.

[9] Donald J Foss and Joseph W Pirozzolo. 2017. Four semesters investigating frequency of testing, the testing effect, and transfer of training. *Journal of Educational Psychology* 109, 8 (2017), 1067.

[10] Frank E Fulkerson and Glen Martin. 1981. Effects of exam frequency on student performance, evaluations of instructor, and test anxiety. *Teaching of Psychology* 8, 2 (1981), 90–93.

[11] Jessica Gaynor and Jim Millham. 1976. Student performance and evaluation under variant teaching and testing methods in a large college course. *Journal of Educational Psychology* 68, 3 (1976), 312.

[12] Barney G Glaser and Anselm L Strauss. 2017. *The discovery of grounded theory: Strategies for qualitative research.* Routledge.

[13] Cathy A Grover, Angela H Becker, and Stephen F Davis. 1989. Chapters and units: Frequent versus infrequent testing revisited. *Teaching of Psychology* 16, 4 (1989), 192–194.

[14] Naomi Holmes. 2015. Student perceptions of their learning and engagement in response to the use of a continuous e-assessment in an undergraduate module. *Assessment & Evaluation in Higher Education* 40, 1 (2015), 1–14.

[15] Jeffrey D. Karpicke and Henry L. Roediger. 2008. The Critical Importance of Retrieval for Learning. *Science* 319, 5865 (2008), 966–968. https://doi.org/10.1126/science.1152408 arXiv:https://www.science.org/doi/pdf/10.1126/science.1152408

[16] Noel Keys. 1934. The influence on learning and retention of weekly as opposed to monthly tests. *Journal of Educational Psychology* 25, 6 (1934), 427.

[17] Abdulkhalig SS Khalaf and Gerald S Hanna. 1992. The impact of classroom testing frequency on high school students' achievement. *Contemporary Educational Psychology* 17, 1 (1992), 71–77.

[18] Jinwoo Kim and F Javier Lerch. 1997. Why is programming (sometimes) so difficult? Programming as scientific discovery in multiple problem spaces. *Information Systems Research* 8, 1 (1997), 25–50.

[19] Klaus Krippendorff. 2009. Testing the reliability of content analysis data. *The content analysis reader* (2009), 350–357.

[20] Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology.* Sage publications.

[21] Daniel H Kulp. 1933. Weekly tests for graduate students. *School and Society* 38, 970 (1933), 157–159.

[22] Frank C Leeming. 2002. The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology* 29, 3 (2002), 210–212.

[23] VT Mawhinney, DE Bostow, DR Laws, GJ Blumenfeld, and BL Hopkins. 1971. A COMPARISON OF STUDENTS STUDYING-BEHAVIOR PRODUCED BY DAILY, WEEKLY, AND THREE-WEEK TESTING SCHEDULES 1. *Journal of Applied Behavior Analysis* 4, 4 (1971), 257–264.

[24] Marsha A McDaris. 1984. Test Frequency Revisited: A Pilot Study. (1984).

[25] Rodrigo Pessoa Medeiros, Geber Lisboa Ramalho, and Taciana Pontual Falcão. 2018. A systematic literature review on teaching and learning introductory programming in higher education. *IEEE Transactions on Education* 62, 2 (2018), 77–90.

[26] J. Piaget. 1997. *The Child's Conception of Number.* Routledge. https://books.google.com/books?id=1ET995R5VHEC

[27] Anthony Robins. 2010. Learning edge momentum: A new account of outcomes in CS1. *Computer Science Education* 20, 1 (2010), 37–71.

[28] Christopher A Rowland. 2014. The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological bulletin* 140, 6 (2014), 1432.

[29] Richard A Schmidt and Robert A Bjork. 1992. New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological science* 3, 4 (1992), 207–218.

[30] Austin H Turney. 1931. The effect of frequent short objective tests upon the achievement of college students in educational psychology. *School and Society* 33, 858 (1931), 760–762.

[31] Bram E Vaessen, Antoine van den Beemt, Gerard van de Watering, Ludo W van Meeuwen, Lex Lemmens, and Perry den Brok. 2017. Students' perception of frequent assessments and its relation to motivation and grades in a statistics course: a pilot study. *Assessment & Evaluation in Higher Education* 42, 6 (2017), 872–886.

[32] Eric F Ward. 1984. Statistics mastery: A novel approach. *Teaching of Psychology* 11, 4 (1984), 223–225.

[33] Matthew West, Geoffrey L. Herman, and Craig Zilles. 2015. PrairieLearn: Mastery-based Online Problem Solving with Adaptive Scoring and Recommendations Driven by Machine Learning. In *2015 ASEE Annual Conference & Exposition.* ASEE Conferences, Seattle, Washington.

[34] Matthew West, Nathan Walters, Mariana Silva, Timothy Bretl, and Craig Zilles. 2021. Integrating Diverse Learning Tools using the PrairieLearn Platform. In *Seventh SPLICE Workshop at SIGCSE.*

[35] Craig Zilles, Matthew West, David Mussulman, and Timothy Bretl. 2018. Making testing less trying: Lessons learned from operating a Computer-Based Testing Facility. In *2018 IEEE Frontiers in Education (FIE) Conference.* San Jose, California.

[36] Craig B Zilles, Matthew West, Geoffrey L Herman, and Timothy Bretl. 2019. Every University Should Have a Computer-Based Testing Facility.. In *CSEDU (1).* 414–420.