Every University Should Have a Computer-Based Testing Facility

Craig Zilles, Matthew West, Geoffrey Herman, Timothy Bretl

[zilles,mwest,glherman,tbretl]@illinois.edu

University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

ABSTRACT

For the past five years we have been operating a Computer-Based Testing Facility (CBTF) as the primary means of summative assessment in large-enrollment STEM-oriented classes. In each of the last three semesters, it has proctored over 50,000 exams for over 6,000 unique students in 25−30 classes. Our CBTF has simultaneously improved the quality of assessment, allowed the testing of computational skills, and reduced the recurring burden of performing assessment in a broad collection of STEM-oriented classes, but it does require an up-front investment to develop the digital exam content. We have found our CBTF to be secure, cost-effective, and well liked by our faculty, who choose to use it semester after semester. We believe that there are many institutions that would similarly benefit from having a Computer-Based Testing Facility.

KEYWORDS

assessment, higher education, computer, exams, frequent testing, second-chance.

1 INTRODUCTION

Exams are a commonly-used mechanism for summative assessment in postsecondary education, especially in introductory courses. At many universities, however, introductory courses are large (e.g., 200+ students), presenting logistical challenges to running traditional pencil-and-paper exams, including requesting space, printing exams, proctoring, timely grading, and handling conflict exams [14, 20]. These practical concerns place a significant burden on faculty and their course staff and generally dictate many aspects of how exams are organized.

Unfortunately, because exams are traditionally designed for summative assessment only, faculty seldom use them in ways designed to improve students' learning. However, exams can be formative in function as well, providing a critical mechanism to improve students' metacognition, prime them for future learning, and help them retain knowledge for longer [16, 17]. When exams are given only once, there is no incentive for students to re-learn material. In contrast, when exams are used in a mastery-based learning context, students are required to review and master material before moving on, deepening their learning and helping them take advantage of the feedback that exams provide [3, 10]. In addition to mastery-based paradigms, spaced testing of the same concept over time and frequent testing are two additional techniques that can help students learn content more quickly and retain it for longer. Unfortunately, most exams keep testing new content, rarely returning to previously tested material. Furthermore, the high cost of running exams leads to the use of only a few exams in a course, and they tend to be very high stakes.

In an effort to mitigate the tension between practical and pedagogical concerns in running exams for large classes, we developed our Computer-Based Testing Facility (CBTF, Figure 1). The CBTF's goal is to improve the exam experience for everyone involved—students, faculty, and course staff. Four concepts are central to

Figure 1: The Computer-based Testing Facility (CBTF) is a dedicated, proctored computer lab for summative assessment using complex, authentic items, which permits students to schedule exams around their other commitments.

achieving this goal. First, by running the exams on computers, we can write complex, authentic (e.g., numeric, programming, graphical, design) questions that are auto-gradable, allowing us to test a broad set of learning objectives with minimal grading time and providing students with immediate feedback. Second, rather than write individual questions, we endeavor to write question generators—small pieces of code that use randomness to produce a collection of problems—allowing us to give each student different questions and permitting the problem generators to be used semester after semester. Third, because each student has a unique exam, we allow students to schedule their exams at a time convenient to them within a specified day range, providing flexibility to students, avoiding the need to manage conflict exams, and allowing very large classes to be tested in a relatively small facility. Finally, because exam scheduling and proctoring is handled completely by the CBTF, once faculty have their exam content, it is no more effort to run smaller, more frequent exams, which reduces anxiety for some students [1, 12]. Furthermore, exams can become more formative as instructors can offer second-chance exams to struggling students with relative ease, giving these students a reason to review and demonstrate mastery of concepts that they missed on an exam.

Now operating in its fifth year, our CBTF has become a resource that faculty have come to rely on. In each of the past three semesters, the CBTF has proctored more than 50,000 mid-term and final exams for more than 6,000 unique students enrolled in more than 25 classes. In addition, the CBTF has changed how we teach, leading to more frequent assessment, improved student learning [15] and the re-introduction of more open-ended assignments.

This position paper advocates for other universities to explore and adopt Computer-Based Testing Facilities to improve assessment

at their institutions. We write this paper motivated by the belief that our institution is not unique in its need to offer large enrollment STEM courses nor in its perceived tension between best practice assessment and logistical overhead with pencil-and-paper exams in these classes.[1] We offer our CBTF implementation as a starting point for these investigations, as it is a model that has withstood the test of time and has been operated at scale. To this end, this paper briefly summarizes salient details about the implementation, philosophy, learning benefits, security, and faculty and student experience with the CBTF.

2 CBTF IMPLEMENTATION

In principle, a CBTF implementation is straight forward. It consists of five main components: 1) a physical space with computers, 2) software for delivering exams, 3) the class-specific exam content, 4) staff to proctor exams, and 5) a means for scheduling students into exam times. While details of the implementation, which are discussed elsewhere [21], are important for handling exam accommodations, ensuring security, and providing faculty with the information they need without the burden of excessive communication, there are two concepts that are central to the implementation: question randomization and asynchronous exams.

In our transition to computerized exams, we've gone to great effort to not dumb down our exams. While many learning management systems (LMS) only support auto-grading of a small range of questions types (e.g., multiple choice, matching), the PrairieLearn LMS [18, 19] provides complete flexibility (i.e., the full capabilities of a web browser) to problem authors. This means that we're capable of asking numerical, symbolic, drawing, and programming questions, basically any question type where the answer can be objectively graded by writing a computer program to score the answer. Furthermore, many PrairieLearn questions are written as *question generators* [9] that can produce a wide range of *question instances* by including a short computer program to randomly select parameters or configurations of the question. By writing generators, we can give each student their own instances of the problems and reuse the generators for homework and exams semester after semester, without worrying about students getting an advantage from having access to old solutions.

In addition, we've found that running exams asynchronously, where students take the exam at different times in a given exam window, is key to the efficiency of the CBTF. First, it would be expensive to provision a computer lab large enough for our largest classes (500+ students), and, second, it is practically impossible to get all of the students in a large class to take the exam at the same time due to illnesses and conflicts. Instead, we run our 85-seat CBTF roughly 12 hours a day, seven days a week and allow students the choice of when to take their exam during a 2–4 day exam period. Students make and change their reservations using a fully-automated web-based scheduling tool and they love the flexibility provided by this aspect of the CBTF [22].

Many instructors are initially wary of running exams asynchronously, because of the potential for students to collude to pass information from early test takers to later test takers. The key to mitigating this concern is to generate random exams for each student where not only are the problem parameters changed

(i.e., problem generators are used), but the problems/generators are drawn from pools of problems. Such a strategy makes it harder for students to collect complete information about the exam and harder to memorize (rather than learn) all of that information. An empirical study found that randomizing question parameters and selecting problems from a pool of 2–4 problems was sufficient to make insignificant the informational advantage from colluding with other students [5, 6].

A side-effect of running exams on computers is that it enables us to test a student's ability to use a computer in problem solving. This benefit is most obvious in programming-oriented exams where students can compile, test, and debug their code before submitting it for grading [4], a much more authentic scenario for programming than writing code on paper. Perhaps less obvious, though, is that this capability is also valued in our engineering courses, which are trying to tightly integrate computation into their curricula. Computerized exams permit these courses to pose non-trivial problems to students that require them to write small programs or use computational tools to produce solutions.

Our implementation of the CBTF has proven to be cost effective. We estimate that exams offered in the CBTF have an amortized cost of between 1 and 2 dollars each, including scheduling, proctoring, grading, and supplies [21].[2] By far, our biggest expense is personnel, but the CBTF's economy of scale makes even its staffing cost effective relative to courses proctoring their own exams. Our costs are an order of magnitude lower than commercial proctoring services.

3 PHILOSOPHY

By delegating much of the work of proctoring to the CBTF and the work of grading to computer programs, we free up faculty and course staff resources for higher-value activities in courses. Our goal is not automation for automation's sake, but rather to automate tasks that are improved through automation (e.g., web-based homework systems can provide immediate feedback, provide an endless supply of problems, and be adaptive) to allow the humans to focus on the tasks that cannot be effectively automated (e.g., one-on-one question answering, grading open-ended projects). We believe that the CBTF improves testing by enabling faculty to offer shorter, more-frequent exams [2], by providing students immediate feedback [11], and by enabling faculty to offer second-chance tests[3].

Furthermore, we don't advocate that auto-graded questions need to make up the entirety of a course's summative assessment. To date we've had the most success writing auto-graded questions for "building block" skills and structured design tasks in STEM courses. For courses that want to include higher-level, open-ended, or integrative tasks (e.g., creative design, requirements gathering, critique), we recommend a blended assessment strategy where the CBTF is used for the objectively gradable tasks and subjective grading tasks are performed manually. In fact, we've seen a number of large-enrollment courses reintroduce team activities, lab reports, and projects because the teaching assistants are no longer burdened with traditional exam proctoring and grading.

---

[1]This belief is validated by the existence of the Evaluation and Proficiency Center [7] at the University of Central Florida, which was developed concurrently with our CBTF and shares much with it in the way of philosophy and implementation.

[2]This cost estimate does not include the cost of the space or utilities, which were too hard to isolate.

[3]Second-chance testing is the practice of providing students feedback about what they got wrong on an exam, permitting them to remediate the material, and then offering them a second (equivalent but different) exam for some form of partial grade replacement.

In general, it is our view that proctoring exams and checking the correctness of completely correct answers is not a good use of highly-skilled faculty and teaching assistant time. For example, the wide spread practice in introductory programming classes of having teaching assistants grade pencil-and-paper programming exams by "compiling" student code in their heads seems particularly inefficient. Instead, faculty and course staff time can be directed to improving student learning through more face time with students and developing better materials for the course. Our experience has been that not only do students highly value these activities, but faculty and staff prefer them to doing routine exam grading.

4 LEARNING GAINS

In addition to the reduction in recurring grading effort and exam logistics, the biggest motivation for using the CBTF is improved learning outcomes. Two quasi-experimental studies have been performed in the CBTF that had the same basic structure. In both cases, a course taught by the same faculty member was compared from one semester to the same semester in the following year. An effort was made to ensure that the only thing that changed in the course was to convert some of the summative assessment to use the CBTF. Student learning was compared across semesters through the use of a retained pencil-and-paper final exam.

In the first experiment [13], a sophomore-level mechanics of materials course was modified to replace two two-hour pencil-and-paper mid-terms with five 50-minute exams in the CBTF, each with a second-chance exam offered in the following week. As shown in Figure 2, the more frequent testing enabled by the CBTF led to a more than halving of the number of D and F grades and a doubling of the number of A grades on an identical retained final exam.
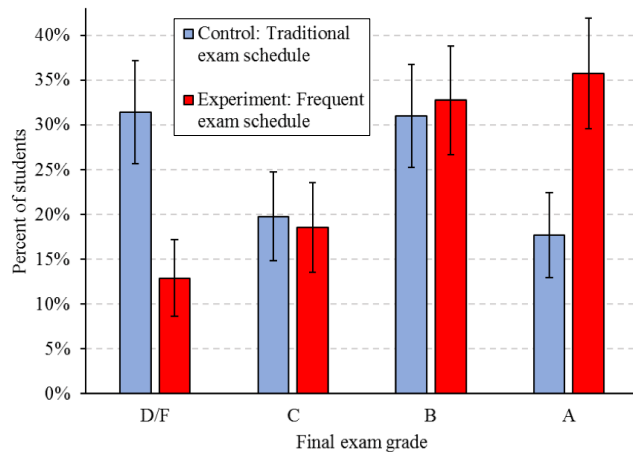

Figure 2: Replacing long pencil-and-paper mid-terms with shorter, more frequent computer-based exams led to a reduction of failing grades on the final exam and a commensurate increase in the number of A grades.

In the second experiment [15], a junior-level programming languages course was modified to convert its two two-hour pencil-and-paper mid-terms into two, two-hour computer-based exams in the CBTF. In addition, four of the 11 programming assignments were no longer collected, but instead students were asked to go to the CBTF to re-write a random fifth of the assignment in the CBTF. As shown in Figure 3, the combination of the computerized exams, the higher level of accountability for the programming assignments,

and the more frequent testing enabled by the CBTF, all led to a substantial reduction in the number of failing grades on the final exam.
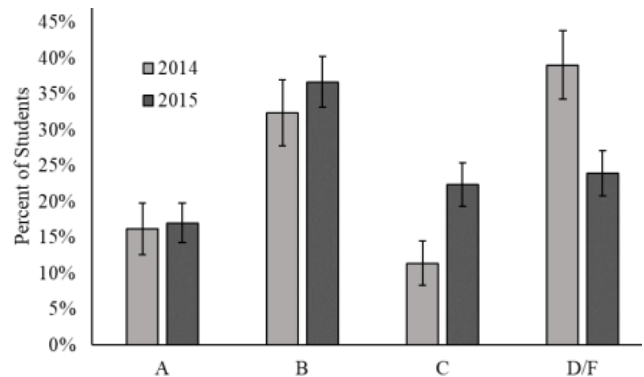

Figure 3: Replacing pencil-and-paper mid-terms (2014) with computer-based exams and requiring portions of four programming assignments to be re-written in the CBTF (2015) led to a significant reduction in the number of failing grades on the final exam.

Faculty and students are both overwhelmingly positive about shorter, more frequent exams [22]. Students prefer them because each exam is less stressful, because it is a smaller fraction of their overall grade. Faculty like them because they prevent student procrastination. As one faculty member said:

> "The CBTF has allowed us to move from a standard 3-midterm model to a weekly quiz model. As a result, students are staying on top of the material, which has made a substantial impact to their learning, but also feeds back into the lecture and lab components of our course. Students are more participatory in these sections because they have not fallen behind." [22]

5 FACULTY AND STUDENT EXPERIENCE

Faculty on the whole are very positive about their experience with the CBTF; we provide here an overview of findings from a collection of surveys of faculty users of the CBTF [22]. The majority find that the CBTF reduces their effort to run exams, reduces their effort to deal with student exceptions, improves student learning in their course, and improves their ability to test computational skills. Furthermore, these faculty see the CBTF as a necessity to support enrollment growth, and half of those we surveyed would be willing to accept a reduced number of teaching assistants to be able to continue using the CBTF. Faculty demonstrate that they value the CBTF through their actions as well; in the past 4 semesters, 90% of courses have returned to the CBTF in the next semester that the course was offered, and many faculty that have used the CBTF introduce it into new courses as their teaching assignments change.

The biggest hurdle to adoption of the CBTF in a course is the up-front investment required to develop the digitized exam content. In addition, one's exam construction mind set has to change, as some commonly-used practices (e.g., questions with multiple dependent parts) aren't as appropriate for computerized exams. As one faculty member stated, "CBTF exams are *not* a drop-in replacement for traditional pencil-and-paper exams. They are different. Your exams (and policies) have to change." We recommend that courses develop
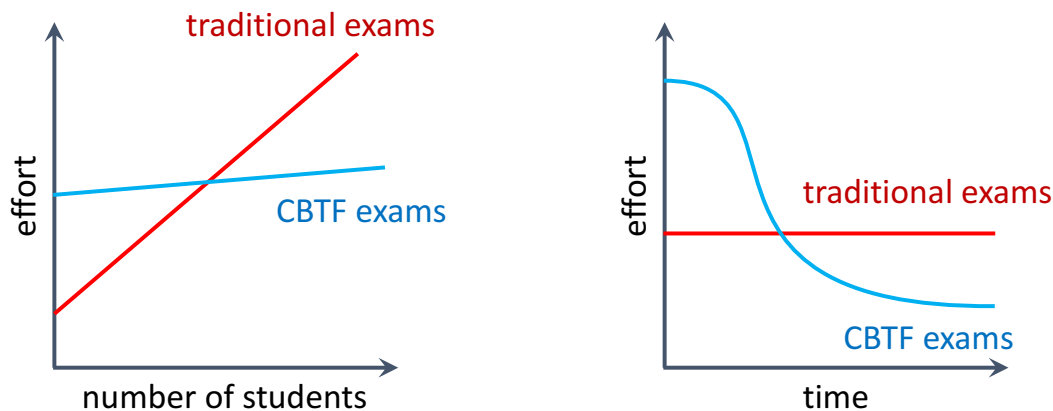
Figure 4: Scaling properties of CBTF exams in terms of instructor effort. Left: traditional exams are less effort for a small number of students, but CBTF exams become more efficient for hundreds of students in a course. Right: CBTF exams require more up-front effort to create pools of question generators, but are much less effort to repeat in the future.

auto-graded questions and deploy them as homework for a semester before offering computerized exams. This ensures that enough content will be available and that questions are tested in a low stakes environment before being used on exams. Another faculty member noted, "It is easy to underestimate how much effort it is to develop good question generators."

Our experience has been, however, that this investment pays off quickly with reduced recurring exam construction, proctoring, and grading time, especially in large classes (see Figure 4). With question generators and question pools resulting in unique exams for each student, faculty can heavily reuse their exam content from semester to semester with less concern for exam security. Most faculty then focus on incrementally refining and enhancing their exam content rather than unproductively churning out new exams each semester. In addition, we've found that the necessity for the grading scheme to be designed before the question is given to students (in order to implement an auto-grader) has led many faculty to think more deeply about what learning objectives their questions are testing and the design of their exams. One faculty remarked:

> "This has revolutionized assessment in my course. It is much more systematic, the question quality is much improved, and my TAs and myself can focus on preparing questions (improving questions), rather than grading."

Over the CBTF's lifetime, a number of student surveys have been performed; we summarize here salient findings of those surveys [22]. Student satisfaction with the CBTF is broadly high, as shown in Figure 5. Many students appreciate the asynchronous nature of CBTF exams, which allows them to be scheduled at convenient times of the day and around deadlines in other courses. In addition, students find the policies to be reasonable, find opportunities to take second-chance exams to be valuable for their learning, and like getting immediate feedback on their exam performance. Finally, students generally prefer more frequent testing because each exam is less anxiety provoking, as they are each worth a smaller portion of the final grade. Some students, however, report fatigue from frequent testing, especially if they are taking multiple CBTF-using courses. In light of the learning gains presented above, which we largely attribute to more frequent testing, the optimal frequency

of testing considering both cognitive and affective impacts is a question that deserves further study.
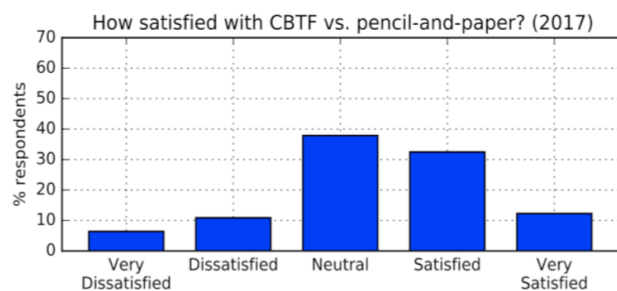

Figure 5: Many more students are satisfied than dissatisfied with CBTF exams (in relation to pencil-and-paper exams).

In our surveys, we found that computer science and electrical/computer engineering students are disproportionately fond of the CBTF. In part, these students are more comfortable with computers generally and benefit from taking programming exams on computers where they can compile, test, and debug their programs to avoid losing points to easy to find bugs. In addition, these computing-oriented students have a lot of prior experience with finicky all-or-nothing systems like compilers. In contrast, the physical engineering disciplines report below average affinity for the CBTF, and their primary concern is the manner that partial credit is granted in the CBTF. While many exams in the CBTF grant partial credit for students that can arrive at the correct answer on their 2nd or 3rd attempt (for example), the auto-graders give no credit if the student answer doesn't meet any of the desired criteria. This is in stark contrast (for the students) to common practice in paper exams in STEM subjects, where partial credit is often granted to students that correctly do some of the set-up steps (e.g., writing relevant equations) in problem solving questions even if the calculation isn't performed correctly.

DeMara et al. have developed a technique that they call *score clarification* that simultaneously improves student satisfaction with auto-graded exams and induces deeper metacognition in students about the questions they got wrong [8]. With score clarification, the students' scratch paper is scanned when they complete the

exam. Then, after the exam period is over, students can review their exam, the correct answers to their questions, and their scratch paper under the supervision of a TA. Students can then (verbally) make a case to the TA to get some partial credit by demonstrating how their scratch represents part of the solution process and an understanding of how they failed to reach the correct answer. Key to this process relative to traditional partial credit grading is that it is the student that has to reconcile their work with the correct answer and articulate why they deserve credit. We have begun prototyping score clarification at our CBTF with similar positive results.

Lastly, the majority of students report that the CBTF is more secure than traditional pencil-and-paper exams [22]. Student comments explain how the CBTF's physical and electronic security prevents common cheating strategies and indicate that "CBTF staff check for cheating more intensely than instructors in regular tests". Students are surprisingly positive about the inclusion of security cameras in the CBTF; student written comments on the survey suggest that most students want an exam environment that doesn't encourage cheating. A number of students did remark that it is commonplace for students after leaving the CBTF to discuss their exams with friends waiting to take that same exam. These anecdotes only reinforce our belief that it is necessary to randomize exams as discussed above.

6 CONCLUSIONS

In this position paper, we have argued for the benefits of Computer-Based Testing Facilities in higher education, like the one we have implemented at the University of Illinois. We have provided evidence that our own facility improves student learning outcomes (e.g., by reducing the number of failing grades on final exams), allows practical adoption of exactly those course policies that are thought to lead to these outcomes (e.g., the use of frequent and second-chance testing as a proxy for mastery-based learning), can be operated efficiently at very large scales (e.g., using one room with 85 seats to serve 50K exams each semester for a total cost of less than $2 per exam) and—despite requiring changes both in how exams are designed and how they are taken—leads to broad faculty and student satisfaction (e.g., positive survey results and continued use by courses from one semester to the next). We have described the architecture of our facility and, in particular, the two key concepts—question randomization and asynchronous exams—that are central to its implementation. We have noted that the University of Central Florida concurrently developed a similar facility that shares many of the same principles and methods as our CBTF, and we believe that this demonstrates the potential for creation of Computer-Based Testing Facilities at other institutions.

Although we have emphasized the utility of our CBTF to very large courses (more than 200 students) in this position paper, it is important to note that our facility is also used by—and provides the same significant benefits to—many smaller courses (less than 100 students). The existence of large courses, often prompted by steady growth in student enrollment and a decline in state funding for public universities, are a key driver for adopting facilities like ours. However, we have seen that once a Computer-Based Testing Facility is available, it is attractive to a broad range of faculty teaching both large and small courses.

REFERENCES

[1] Joni K Adkins and Diana Linville. 2017. Testing frequency in an introductory computer programming course. *Information Systems Education Journal* 15, 3 (2017), 22.

[2] Robert L. Bangert-Drowns, James A. Kulik, and Chen-Lin C. Kulik. 1991. Effects of Frequent Classroom Testing. *Journal of Educational Research* 85 (1991).

[3] B.S. Bloom. 1968. Learning for mastery. *Evaluation Comment* 1, 2 (1968), 1–12.

[4] Jacabo Carrasquel, Dennis R. Goldenson, and Philip L. Miller. 1985. Competency testing in introductory computer science: the mastery examination at Carnegie-Mellon University. In *SIGCSE '85*.

[5] Binglin Chen, Matthew West, and Craig Zilles. 2017. Do Performance Trends Suggest Wide-spread Collaborative Cheating on Asynchronous Exams?. In *Learning at Scale*.

[6] Binglin Chen, Matthew West, and Craig Zilles. 2018. How much randomization is needed to deter collaborative cheating on asynchronous exams?. In *Learning at Scale*.

[7] Ronald F. DeMara, Navid Khoshavi, Steven D. Pyle, John Edison, Richard Hartshorne, Baiyun Chen, and Michael Georgiopoulos. 2016. Redesigning Computer Engineering Gateway Courses Using a Novel Remediation Hierarchy. In *2016 ASEE Annual Conference & Exposition*. ASEE Conferences, New Orleans, Louisiana. https://peer.asee.org/26063.

[8] Ronald F. DeMara, Tian Tian, and Wendy Howard. 2018. Engineering assessment strata: A layered approach to evaluation spanning Bloom's taxonomy of learning. *Education and Information Technologies* (04 Oct 2018). https://doi.org/10.1007/s10639-018-9812-5

[9] Mark J Gierl and Thomas M Haladyna. 2012. *Automatic item generation: Theory and practice*. Routledge.

[10] Chen-Lin C. Kulik and James A. Kulik. 1987. Mastery Testing and Student Learning: A Meta-Analysis. *Journal of Educational Technology Systems* 15, 3 (1987), 325–345. https://doi.org/10.2190/FG7X-7Q9V-JX8M-RDJP arXiv:https://doi.org/10.2190/FG7X-7Q9V-JX8M-RDJP

[11] James A. Kulik and Chen-Lin C. Kulik. 1988. Timing of Feedback and Verbal Learning. *Review of Educational Research* 58, 1 (1988), 79–97. https://doi.org/10.3102/00346543058001079 arXiv:http://rer.sagepub.com/content/58/1/79.full.pdf+html

[12] T. Kuo and A. Simon. 2009. How many tests do we really need? *College Teaching* 57 (2009), 156–160. Issue 3.

[13] J. Morphew, M. Silva, G. L. Herman, and M. West. 2019. Improved learning in a university engineering course from an increased testing schedule. (2019). (preprint).

[14] Robyn Muldoon. 2012. Is It Time to Ditch the Traditional University Exam? *Higher Education Research and Development* 31, 2 (2012), 263–265.

[15] Terence Nip, Elsa L. Gunter, Geoffrey L. Herman, Jason W. Morphew, and Matthew West. 2018. Using a Computer-based Testing Facility to Improve Student Learning in a Programming Languages and Compilers Course. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education (SIGCSE '18)*. ACM, New York, NY, USA, 568–573. https://doi.org/10.1145/3159450.3159500

[16] M. A. Pyc and K. A. Rawson. 2010. Why testing improves memory: Mediator effectiveness hypothesis. *Science* 330 (2010), 335.

[17] K. A. Rawson and J. Dunlosky. 2012. When is practice testing most effective for improving the durability and efficiency of student learning? *Educational Psychology Review* 24 (2012), 419–435. https://doi.org/10.1007/s10648-012-9203-1

[18] M. West. url. https://github.com/PrairieLearn/PrairieLearn.

[19] 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.

[20] C. Zilles, R. T. Deloatch, J. Bailey, B. B. Khattar, W. Fagen, C. Heeren, D Mussulman, and M. West. 2015. Computerized Testing: A Vision and Initial Experiences. In *American Society for Engineering Education (ASEE) Annual Conference*.

[21] 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.

[22] C. Zilles, M. West, D. Mussulman, and C. Sacris. 2018. Student and Instructor Experiences with a Computer-Based Testing Facility. In *10th annual International Conference on Education and New Learning Technologies (EDULEARN)*.