Paper or Online? A Comparison of Exam Grading Techniques

Yingjun Cao, Leo Porter, Soohyun Nam Liao, Rick Ord

Department of Computer Science and Engineering

University of California, San Diego

La Jolla, CA, USA

ABSTRACT

As computer science enrollments continue to surge, exam grading requires significant instructional resources. Online grading platforms have been developed in recent years and have been adopted at a number of institutions; however, their effectiveness compared with traditional grading on paper is not fully known. This study is the first in CS to compare online and paper grading. Comparing overall time to grade, including all factors, online grading doesn't show a consistent advantage compared with paper grading. We observed that online grading is much faster during the actual grading phase, but some of this benefit is offset by the additional overhead prior to grading (e.g., scanning) for online grading. Examining student and grader preferences based on feedback, both groups show a strong preference for the online format, predominately due to the convenience of the online platform. Graders report being able to grade more accurately online with the ability to modify rubrics, but that grading on paper tends to result in more social interactions among graders.

CCS CONCEPTS

• **Social and professional topics** → **Computing Education**.

KEYWORDS

Assessment; Exam Grading; Online Grading System

ACM Reference Format:

Yingjun Cao, Leo Porter, Soohyun Nam Liao, Rick Ord. 2019. Paper or Online? A Comparison of Exam Grading Techniques. In *Innovation and Technology in Computer Science Education (ITiCSE '19)*, July 15-17, 2019, Aberdeen, Scotland Uk. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3304221.3319739>

1 INTRODUCTION

The importance of both formative and summative assessments is well established. Formative assessments inform students of where misconceptions lie, enabling future practice and study. In fact, students often focus on various skills or concepts when they believe they will need to use them on an exam. Furthermore, we know that students (and indeed people in general) will learn something better when they are tested on it, a phenomenon known as the "testing effect" [25]. Similarly valuable, summative assessments are

the primary mechanism for instructors to measure how much each student has learned.

In computer science classes, summative (and often formative) assessments frequently take the form of written exams, particularly in lower-division courses [26]. Instructors often seek to make the experience valuable for the students by providing useful feedback for students by (and when) grading. In fact, feedback on exams and quizzes may lead to better knowledge retention [12, 21, 28].

As computer science enrollments surge [9], there is an increased need to scale all the components of the course. Recent work has examined how pedagogies scale [22], but here we focus on scaling assessments, specifically the grading of exams. For instructors of these large classes, they are finding that grading has become an increasingly time-consuming task [19]. If the exams are predominately multiple choice questions, there are automated tools to reduce this time demand (e.g., automatically grading using tools such as Scantron machines). However, multiple choice questions do not always meet instructors' assessment needs. To more authentically assess desired learning outcomes in a course, many computer science exams have short answer or coding questions which often do not have unique answers. These questions require instructional staff to grade manually.

In order to grade the large quantity of exams (and to provide timely feedback), a large number of graders are often needed to grade hundreds of papers in hours. This instructional staff often have limited hours of employment, so time spent grading reduces time spent helping students. As a result, CS educators are seeking means of improving grading efficiency without resorting to exams which are strictly multiple choice questions.

Software tools to improve grading have been proposed with a major goal to improve efficiency. For example, programming projects have been increasingly autograded and feedback automatically generated [17, 20]. Exam grading tools have also been proposed though not as abundant as programming projects [4, 26]. One such tool and the tool used in this study is Gradescope [26]. Gradescope allows instructors to scan student exams, upload them to an online system, and grade online.

Despite claims in support of these tools, there is no empirical evidence to support that exam grading systems offer superior grading over traditional paper grading. This study is designed to investigate this claim in three ways: 1) the comparison of grading time between exam grading systems and paper grading on real computer science exam papers containing coding questions, 2) students' concern on these two forms of grading, and 3) comparing grader experiences using grading tools and grading on paper.

2 BACKGROUND

The role of assessment in student learning is a broad area of inquiry [8, 11, 15]. Assessments strongly influence student learning

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ITiCSE '19, July 15-17, 2019, Aberdeen, Scotland Uk

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6301-3/19/07...\$15.00

<https://doi.org/10.1145/3304221.3319739>

activities [27]. The so-called hidden curriculum, i.e., what students perceive will appear in future assessments, dominates what topics students study and the time they invest in that studying. Students usually try to achieve high study efficiency by aligning their study as much as they can with perceived assessment requirements [2, 5].

Formative assessments are also valuable to students for identifying areas of weakness and for helping self-regulate their learning [24]. Summative assessment is a necessary component in the role of certifying a student's knowledge [6, 18]. Due to potential academic integrity issues with take-home exams, summative exams in CS classes are usually conducted on paper with proctors during designated time periods. Given the important role of assessments in computing, we focus here on the practical issue of how to grade given limited instructor resources.

With regard to grading, educators in computer science have designed systems to improve the grading process for a variety of assessments. One recent work on exam grading by Harrington et al. [19] examined the quality of manual TA markings during marking parties, and recommended organized group exam marking sessions. Another area of work has been designing systems to improve grading for coding projects [23]. For example, Dewey et al. [10] used constraint logic from the programming language literature to develop test suites for autograding. We refer readers to this review article on project autograding [20]. However, on the automation of exam grading, most systems rely on Scantron to grade multiple choice questions. Exams in computer science usually involve coding and short answers which are not suitable for Scantron.

Though it is possible to take exams entirely on computers, nearly all online exam grading systems in CS start by scanning handwritten exam papers then grading can be done on computers [4, 26]. One newly developed system, SAFE [7] starts to use student's electronic devices to conduct exams directly. There is no empirical study on how efficient it is to grade scanned exams on computer, and what are potential differences between grading on computer and grading on paper. This study seeks to answer this practical question.

Gradescope [26] is the platform commonly used on our campus for online grading and was hence selected in this study to compare with grading on paper. The system requires that handwritten papers be scanned into PDF formats and uploaded into Gradescope server. It allows instructors to build templates based on where students will write their answers on the exam paper and allows graders to only focus on student answers in the scanned papers. Standard rubrics can be developed for every exam question, and the rubrics are easy to modify. To limit the amount of papers to be scanned, instructors can create answer sheets on which students will write their answers.

3 ONLINE GRADING

In this section, we aim to describe the process of online grading to provide context to those unfamiliar. Using Gradescope to grade an exam, as we did in this study, consists of the following steps:

1. Exam Creation. The exam is created as usual. In large classes, the instructor might create different versions of the exam to ensure no cheating occurs. If this is the case, the instructor should ensure

version numbers are adjacent to each question for ease of grading. Questions on all exams should appear in the same location on the page. Creation of the solutions is similar to standard exams.

2. Giving the Exam. Exams are printed and given to students in the standard manner. Students are reminded that their answers should fall within the assigned question region on the exam.

3. Online Setup. The instructor uploads a copy of the exam as a template and delineates bounding boxes around each question (or sub-part of a question). The instructor can also denote how many points each question is worth.

4. Scanning. One of the drawbacks to online grading is this step, where exams are digitized. At our institution, there are two sets of paper cutters and high-speed scanners. Instructional staff take exams, cut off the staples for ease of scanning, and then scan the exams into large PDF files.

5. Uploading Exams. After scanning the exams, the exams can be uploaded onto the online platform. The platform automatically splits the uploaded PDFs based on the number of pages in the template. If multiple versions are used, uploading one version at a time makes grading easier.

6. Student Identification. Instructional staff have to manually match students to exams. This is a fast process as the tool shows the staff member only the name region and assists the staff with an auto-complete feature.

7. Rubric. This can be developed by the instructor, at least in part, prior to exam grading. The rubric can be changed at any time during grading, and any changes in points for rubric items are applied retroactively. For example, if a syntax error was deducted 2 points in the initial rubric, changing the rubric item to only deduct 1 point applies to all exams which have that rubric item marked. This feature is particularly helpful when graders decide to change rubric criteria.

8. Grading Process. The grading process has generally been streamlined on this platform. A grader can be assigned to grade a single question, and keyboard hot keys allow the grader to apply a rubric and to retrieve the next ungraded exam. This eliminates the need to pass around paper copies and set to the correct page.

9. Monitoring Progress. During paper exam grading, instructors can typically see (and often anticipate) bottlenecks in the grading. The online platform makes this easier by providing the percentage of answers graded per question to aid with load balancing efforts.

10. Totaling Exams. This is a major advantage for online grading as the exams are automatically totaled when grading completes, and because students are already matched to the exam, there is no need to enter grades into any system.

11. Returning Exams. The instructor can release the exam once it is graded and students will receive an e-mail notifying them it is available. They can then view the exam online and view the rubric items applied per question.

12. Regrade Requests. Regrade requests can also be submitted online and answered online. A positive benefit for Academic Integrity is that students cannot modify original exams.

4 METHOD

In this section, we first list the questions we seek to address with this work, then provide details on the setting for the work.

4.1 Research Questions

This study intends to answer three related research questions on exam grading techniques:

- **RQ1:** Are there significant time efficiency differences between online exam grading and grading on paper?
- **RQ2:** What are the differences, if any, on experiences from students between online and paper exam grading?
- **RQ3:** What experiences and preferences do exam graders have on the two forms of exam grading?

RQ1 is the primary question, as grading expediency is often cited as the reason for adopting online grading systems. As instructors, we recognize that time spent grading is only one factor we need to consider. As such, RQ2 and RQ3 offer more insights on the most relevant parties of exams, the students and graders. These questions focus on their experiences and preferences with these grading techniques.

4.2 Setting and Experimental Design

The study was carried out in a 10-week advanced data structure class at a large research institution in North America during the 2016-2017 academic year. The class had two sections (A and B), both meeting on MWF and the instructor of both sections is an author of this paper. The course is structured to have two midterms, each one month apart, and the two sections hold exams on the same day. Section A started with 205 students and section B has 200 students at the beginning of the term.

For all practical purposes relevant to this study, the two sections were essentially identical. The only key difference is how the exams were graded: section A's first midterm was graded and returned online while section B's first midterm was graded and returned on paper. For the second midterm, the technique was switched (section A on paper and section B online) to limit potential confounding due to possible differences in students in each section. This experimental design allows each student from both sections to be exposed to both exam grading formats.

Exams on both midterms and both sections were printed on paper, and the exams were graded in a large on-campus eatery either one or two days after the exam. The class had 24 graders for both sections, and most graders are familiar with Gradescope. About 10–15 staff participated in each exam grading, but they arrived and departed during grading at various times to accommodate their class schedules.

To compare the efficiency of the two grading techniques (RQ1), a detailed log of the grading process was recorded with respect to the number of people participating in grading over various time periods. Additionally, time spent before exams (e.g., exam design, answer sheet creation, preparation with online tools, etc.) were also recorded. We also kept track of events after the exam, including the number of students who collected (or viewed) their graded exam and the number of regrade requests received by course staff. To further investigate students' experience and preference with the grading techniques (RQ2), a customized survey with five questions was distributed via a Google form to students at the end of the term to collect their experience on both exam grading formats. The survey contains four Likert-scale questions and one open-ended question. A minimal extra credit was offered to students who completed

the survey. To examine graders' experiences (RQ3), we provided graders with a survey with four open-ended questions.

Each midterm exam had four versions with similar questions but different numbers in those questions. The exam contents and question types are as follows:

Midterm 1. Write the outputs of 3 segments of code (fill in the blanks); write a method for a binary search tree (BST) (coding); draw a multiway trie (MWT) or a ternary search tree (TST) (graphical question); complete a hashtable and calculate probability of collision (short answer), and one runtime analysis question comparing BST with TST and MWT (short answer).

Midterm 2. Complete and analyze an example using the count-min sketch algorithm (short answer); fill in the blanks on a red-black tree insertion example; fill in the blanks for a Huffman encoding question; a code writing problem to perform bitwise operations, and fill in the blanks related to Dijkstra's algorithm. This exam also had two multiple choice questions for code efficiency analysis.

5 RESULTS

In this section, we will provide results for the research questions.

5.1 Grading Time Efficiency (RQ1)

5.1.1 Pre-grading Time. A log was recorded for each midterm including exam design phase and the grading phase. Table 1 includes the time for events that occurred before the grading phase.

Table 1: Logs for pre-grading events (Time unit: minutes)

Pre-grading Events	Exam 1		Exam 2	
	Online	Paper	Online	Paper
Exam question design	480	480	270	270
Answer key preparations and rubric design	48	48	34	34
Create answer sheet and enter rubrics on online system	48	0	42	0
Copy answer key for paper grading	0	11	0	14
Exam digitization (including sorting, cutting, scanning, up-loading, and name matching)	230	0	420	0
Total Time	806	539	766	318

Exam and answer key preparations are the same for both grading formats. The difference is that online grading requires a rubric to be created on the platform with a certain format. For paper grading, answer keys need to be copied to distribute to graders during grading. The one big difference between the grading formats is the scanning time. To use the online platform, the answer sheet from each student will be sorted in page order and will be scanned using scanner(s). For midterm 1, one grader scanned all the answer sheets for section A and the total time was reported to be 200 minutes in total. For the second midterm, 4 graders worked together for 100 minutes to complete the scan process for section B. The total time is calculated as the number of graders multiplied by the time they spent. The reason for a longer exam 2 scanning time is the scanner being used had a very slow interface with old USB drives, and the scanning staff were unaware of this limitation. Note that only the section that is being graded online has its exam papers scanned. The digitization time also includes name matching by a

single grader: 30 minutes total for midterm 1 and 20 minutes total for midterm 2.

5.1.2 Grading Time. The grading phase log is shown in Figure 1 and total times appear in Table 2.

Figure 1: Time logs during grading for two midterms. Time unit is minute. The heights of boxes in the figure are proportional to the number of graders grading in a time period. For both midterms, we started with online grading then moved to grading on paper.

Table 2: Grading time in minutes.

Format	Exam 1			Exam 2		
	t_{total}	# exams	average	t_{total}	# exams	average
Online	861	195	4.42	1392	185	7.52
Paper	1526	197	7.75	1562	194	8.05

The metric we choose to compare grading techniques is the average time it takes to grade an exam paper by a single grader. To calculate this metric, the total amount of time for grading is obtained by weighing the time by the number of people grading. For example, for exam 1 online grading, the total time is calculated as $13 \text{ graders} \times 16 + 13 \text{ graders} \times (68 - 27) + 8 \text{ graders} \times (77 - 68) = 813$ minutes. There were two extra delays in midterm 1 for online grading due to a misplaced answer sheet and more time spent grading that paper which are included in the table results. Table 2 also includes the time the instructor spent grading questions that existing rubrics do not fully apply for each grading technique in both exams (about 30 minutes). For grading on paper, the time to total exam scores and to enter exam grades to a spreadsheet is also included. We acknowledge that the metric is not the only choice though we believe it is one of the fairest for comparison purposes.

RQ1 is answered by Tables 1 and 2: while the online grading format has more overhead before grading, the total actual grading time as well as the average time per exam paper seems to be shorter with online grading. This pattern is true for both exams. However, if we combine both pre-grading and the actual grading time, and calculate the average time per grader per exam paper, there is no consistent pattern. For midterm 1, the average per paper is 12.51 minutes for paper grading, and for online, it is 8.55 per paper. For midterm 2, the trend reversed. It took 9.69 minutes to grade an exam on paper, and 11.66 minutes to grade an exam online.

5.2 Student Experience (RQ2)

An anonymous survey containing five questions was distributed to students near the end of the term. The survey results of the four Likert-scale questions are shown in Table 3. Results indicate that students overwhelmingly prefer to receive exams back from the online grading system and they report being more likely to “pick up” and review exams if they are returned via the online system. Regrade submissions follow a similar trend. The last question on a sense of being more “personally connected” is relatively more balanced: 50.2% of students prefer online compared with 31.9% preferring paper.

Table 3: Survey results on student experience with two grading techniques. Number of entries in the survey is 256. The annotation for ++ is strongly prefer and + is for prefer.

Question	Online ++	Online +	Neutral	Paper +	Paper ++
How do you prefer to receive your exams back?	49%	23.3%	8.9%	9.7%	8.9%
In which grading scenario are you likely to pick up/ review your graded exam?	47.5%	20.6%	12.8%	9.7%	9.3%
In which grading scenario do you feel comfortable submitting re-grades?	43.6%	26.5%	17.5%	6.6%	5.8%
Which grading scenario do you feel personally connected to your exam results?	32.7%	17.5%	17.9%	14.4%	17.5%

5.3 Grader Experience (RQ3)

After the final exam, graders were asked to complete the following four-question online survey concerning the two grading formats: **Q1** - Convenience to grade exams; **Q2** - Speed to grade exams; **Q3** - Communication and team-bonding during grading; and **Q4** - Comment on the overall grading experience

Out of 24 graders, 13 responded and two of these responses were dropped as their responses indicated they misunderstood the questions. The remaining 11 responses show the following themes.

Q1. *Most graders feel that online grading is more convenient.* Arguments for using the online grading system include 1) ease of rubric adjustment, 2) ease of final score totaling, 3) ease of returning of exams, 4) the ability to easily skip hard-to-grade questions, 5) ease of splitting grading tasks into subparts for parallel grading, 6) support for grading remotely, and 7) ease of the regrading process. As for paper exams, graders’ responses indicate that grading on paper seemed to be harder and less convenient.

Q2. *Most graders feel that online grading is faster than grading on paper* with one grader thinking that the speed is not that different between the two grading techniques. One grader commented that grading on paper is faster on coding problems while grading online seemed to work better for other types of questions.

Q3. Six graders commented that grading on paper seemed to encourage better communication among graders. Two graders believed there was no difference, one grader thought grading online in front of a machine facilitates better communications, and two graders

did not respond. For this question, grading on paper is a slight preference for graders.

Q4. We did not observe any consistent theme on Q4. Five graders stated that they like grading on paper for the fun factor and four graders preferred the online system because of speed and accuracy. Two expressed no preference.

6 DISCUSSION

6.1 Time Efficiency Comparison

Tables 1 and 2 show two distinctive patterns. Online grading has more initial overhead for scanning and answer sheet design. This is especially true for midterm 2 where the overhead is more than twice of that from paper grading. However, the availability and proper use of high-speed scanners may reduce this overhead. For the actual time grading the exams, the online platform was faster for both midterms. One benefit with paper grading is that it allows for grading immediately after the exam (no need to wait for scanning). We couldn't measure the grading accuracy of these two formats except counting the number of regrade requests. Though grading on paper seems to have fewer regrade requests, the difference may simply be due to the easiness of submitting regrades online.

6.2 Student Experience

As can be seen from Table 3, more than 70% of students preferred receiving graded exams electronically. In comparison, slightly less than 20% prefer to receive the paper exams back. More than 67% of students reported getting exams back online also made them more likely to review their exams. Again, about 20% of students reported they prefer retrieving physical exams for review.

These differences are large and to understand the reasons behind these preferences, we analyzed the open-ended question on the survey on students' opinions on receiving exams online and on paper. We coded the themes from the open survey question, and top themes are shown in Table 4.

Table 4: The list of top themes in the open-ended question on students' experience on the online exam grading system and paper. Number of valid responses is 256.

	Theme	%
O N L I N E	Convenient and easily accessible	40.8
	Regrade is easy	21.2
	Detailed rubrics are helpful	12.6
	Online storage of exam is safe	7.7
	Online grading system works great for the final exam	7.1
	No cheating on regrades on the online system	4.0
	Some rubrics are vague	2.4
	Better privacy when "picking up" papers	1.8
	Too much technology already	1.2
	Answer sheets without exams are not convenient to review	1.2
P A P E R	Getting exam paper back is time consuming and inconvenient	29.4
	Having paper exam back is better for review and study	24.8
	Paper feels better and embodies a personal touch	20.9
	Hand graded exams have more personal comments	5.2
	Grading rules are easy to understand on paper	4.6
	Writings on paper exams may not be clear	4.6
	Getting the paper by the end of class is nice	3.3
	Can write notes on paper exam	2.6
	Want to keep a hard copy	2.6
	Midterms on paper is good	2.0

The primary reason that students like the online grading format is its convenience. The availability of exams online allows students to review their exams remotely without worrying about losing their exam papers. Students stressed that being able to access the graded exam remotely was especially useful for the final exam as students often leave campus prior to exams being graded. Also convenient for students is the ability to submit a regrade request through the platform by only providing a brief justification. The grader who graded the problem is also conveniently notified that a request was submitted. Students also reported valuing the detailed rubrics provided online which are typically not provided for paper exams (but could be). There are also some side-benefits to support the use of online grading systems. For example, online systems have a scan of the original papers so students cannot modify their exam when they submit a regrade request.

The most important reason that students do not prefer paper exams back is surprisingly the inconvenience of getting it back. Most instructors try to hand back exams near the end of a class section but with hundreds of students, multiple assistants have to be on site and students will go to a specific assistant based on their last name or some other identifiers. If a student misses that class, she has to go to the instructor's office to collect the exam. Moreover, students reported being concerned about the privacy of their grade as when exams are returned in-class, grades are sometimes seen by other students.

Although nearly all students acknowledge the convenience of online grading systems, grading on paper exams has benefits as well. We suspect a large benefit of paper exams is in having more verbose and personalized feedback. Particularly on code-writing questions, it is easier to identify and mark errors on paper than online. (Note that the online platform allows for custom rubric items and for graders to write on exams using a tablet; in our experience, this feature has not been used heavily.) The next important reason is that having a hard copy is more tangible and may connect students more personally to the exam.

The ease of regrades may impact student preference. Our data shows that for midterm 1, there are 9 regrade requests from the online grading system, and 4 from exams graded on paper. For midterm 2, there were 48 regrade requests online while there were 14 from exams returned on paper. All but one online regrade request is frivolous. Therefore, the ease of regrades may make the grading process more fair by reducing the number of grading errors. We do note, however, that ease in submitting regrades may cause students to abuse the system. In our classes, we typically caution students that regrades are to be used to correct obvious grading errors, not to argue for more points.

As described for online grading systems, some students stated that getting midterm exams back on paper is better while receiving exams back online is better for the final exam. Students also like personal comments graders may put on the exam paper though the online system also provides similar functionalities. Last but not least, some students find grading rules on paper are easier to understand compared to the online system. Though they didn't elaborate, we hypothesize that these students probably refer to coding or complicated short answer questions. Graders can draw and write on the location where certain mistakes have occurred while it is not easy to do it on the online system, at least not efficiently.

6.3 Grader Experience

In general, graders reported preferring online grading as it reduced many of the administrative chores compared with grading on paper. However, paper grading might promote better communication among graders and, possibly, community building. With the current discussion on utilizing undergraduate students as teaching assistants to help scaling CS classes [13, 16], this feedback from graders might be valuable. The instructor of the class also observed that graders seem to enjoy the grading process and interact with each other more during grading sessions on paper. They more frequently share exams with each other to decide if a certain solution is worth some partial credit. Having the physical copy may have made this kind of sharing more convenient.

Graders also echoed some of the points raised by students regarding different types of questions may impact which system is preferred. For example, graders tended to remark that code-writing problems are easier to grade on paper as they can make marks on the paper. Multiple choice or short answer questions might be more easily graded on the online system without having to flip through pages on paper exams. Graders also discussed the convenience of changing grading rubrics on the online system. It is much more convenient to apply a change on the allocation of points on the online system than grading on papers. In particular, if rubrics have to be changed after an exam has been distributed, paper exams will involve a huge administrative cost to collect all distributed papers and regrade them. This can be done easily with the online system.

6.4 Quality of Feedback

Existing research has shown that quality feedback may have one of the most influential effects in student learning [3, 14]. Students indicated in the survey that having detailed rubrics from the online platform is helpful to them. However, they may sometimes focus on trying to determine how well their answer matches the rubric to argue for more points instead of trying to understand the concepts. Bennett et al. [1] shows that if feedback is provided together with the marks of assessment, students tend to over-focus on the exam scores instead of examining and reflecting on the feedback. Paper exams are more convenient for students to put notes on their mistakes and it might be easier for graders to provide customized comments, but paper exam return typically does not include providing detailed rubrics. Most of the time, students will only know the correct answer if it is written on their exam papers. Future research on the effect of rubrics on student learning might reveal more nuanced findings.

7 CONCLUSION

This work is the first to empirically compare online grading systems and paper grading in computing from three important perspectives: grading time efficiency, student preferences, and grader preferences. Results show online grading was faster during actual grading, but required scanning before grading. A more streamlined scanning process might be the key to achieve better overall time efficiency using online grading. Students, in general, prefer the online grading system because of its convenient accessibility, regrading system, and available detailed rubrics. Exams graded on paper tend to result in difficulties receiving the exam due to absence on days the exams are returned and time to return the exam in large classes. Exams

graded on paper may offer other advantages including students possibly feeling more connected to the physical exam. Graders tend to prefer online grading for convenience, yet paper grading may result in a better atmosphere among graders. For future studies, we plan to randomly sample exams from students and compare the accuracy of two grading formats on different types of problems.

REFERENCES

- [1] R. E. Bennett. Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1):5–25, 2011.
- [2] J. Biggs. Enhancing teaching through constructive alignment. *Higher Education*, 32(3):347–364, 1996.
- [3] P. Black. Formative assessment: raising standards inside the classroom. *School Science Review*, 80(291):39–46, 1998.
- [4] A. Bloomfield and J. F. Groves. A tablet-based paper exam grading system. *SIGCSE Bull.*, 40(3):83–87, June 2008.
- [5] A. Cain and M. A. Babar. Reflections on applying constructive alignment with formative feedback for teaching introductory programming and software architecture. In *ICSE 2016*.
- [6] Y. Cao and L. Porter. Evaluating student learning from collaborative group tests in introductory computing. In *SIGCSE 2017*.
- [7] K. Chebrolu, B. Raman, V. C. Dommeti, A. V. Boddu, K. Zacharia, A. Babu, and P. Chandan. Safe: Smart authenticated fast exams for student evaluation in classrooms. In *SIGCSE 2017*.
- [8] N. R. Council et al. *How people learn: Brain, mind, experience, and school: Expanded edition*. National Academies Press, 2000.
- [9] CRA. Computing Research Association, The Taulbee Survey. <https://cra.org/wp-content/uploads/2018/05/2017-Taulbee-Survey-Report.pdf>. Accessed Jan. 15, 2019.
- [10] K. Dewey, P. Conrad, M. Craig, and E. Morozova. Evaluating test suite effectiveness and assessing student code via constraint logic programming. In *ITiCSE 2017*.
- [11] F. Dochy, M. Segers, D. Gijbels, and K. Struyven. Assessment engineering. In *Rethinking assessment in higher education: learning for the longer term.*, pages 87–100. Oxford: Routledge, 2007.
- [12] M. L. Epstein, A. D. Lazarus, T. B. Calvano, K. Matthews, R. A. Hendel, B. B. Epstein, and G. M. Brosvic. Immediate feedback assessment technique promotes learning and corrects inaccurate first responses. *The Psychological Record*, 52:187–201, 2002.
- [13] J. Forbes, D. J. Malan, H. Pon-Barry, S. Reges, and M. Sahami. Scaling introductory courses using undergraduate teaching assistants. In *SIGCSE 2017*.
- [14] B. J. Fraser, H. J. Walberg, W. W. Welch, and J. A. Hattie. Syntheses of educational productivity research. *International Journal of Educational Research*, 11(2):187–212, 1987.
- [15] G. Gibbs, C. Simpson, P. Gravestock, and M. Hills. Conditions Under Which Assessment Supports Students' Learning. *Learning in Teaching in Higher Education*, 1(1):3–31, 2004.
- [16] C. Gregg and C. M. Lewis. Working with undergraduate teaching assistants: Best practices and lessons learned. In *SIGCSE 2015*.
- [17] S. Gulwani, I. Radiček, and F. Zuleger. Feedback generation for performance problems in introductory programming assignments. In *FSE 2014*.
- [18] W. Harlen and M. James. Assessment and learning: differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, 4(3):365–379, 1997.
- [19] B. Harrington, M. Ahmadzadeh, N. Cheng, E. H. Wang, and V. Efimov. TA Marking Parties: Worth the Price of Pizza? *ICER 2018*.
- [20] P. Ihanntola, T. Ahoniemi, V. Karavirta, and O. Seppälä. Review of recent systems for automatic assessment of programming assignments. In *Koli Calling 2010*.
- [21] H. Keuning, J. Jeuring, and B. Heeren. Towards a systematic review of automated feedback generation for programming exercises. In *ITiCSE 2016*.
- [22] S. N. Liao, W. G. Griswold, and L. Porter. Impact of class size on student evaluations for traditional and peer instruction classrooms. In *SIGCSE 2017*.
- [23] T. MacWilliam and D. J. Malan. Streamlining grading toward better feedback. In *ITiCSE 2013*.
- [24] D. J. Nicol and D. Macfarlane-Dick. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, 31(2):199–218, 2006.
- [25] H. L. Roediger and J. D. Karpicke. The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1(3):181–210, 2006.
- [26] A. Singh, S. Karayev, K. Gutowski, and P. Abbeel. Gradescope: A fast, flexible, and fair system for scalable assessment of handwritten work. In *L@S 2017*.
- [27] B. R. Snyder. *The hidden curriculum*. New York: Knopf, 1970.
- [28] J. A. Stone. Using reflective blogs for pedagogical feedback in CS1. In *SIGCSE 2012*.