Analyzing the decline of student scores over time within self-scheduled asynchronous exams

Binglin Chen, Matthew West, Craig Zilles
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{chen386, mwest, zilles}@illinois.edu

Abstract

- **Background:** When students are given a choice of when to take an exam in engineering and computing courses, it has been previously observed that average exam scores generally decline over the exam period. This may have implications both for the design of interventions to improve student learning and for data analysis to detect collaborative cheating.
- **Purpose/Hypothesis:** We hypothesize that average exam scores decline over the exam period primarily due to self-selection effects, where weaker students tend to choose exam times later in the exam period while stronger students are more likely to choose earlier times.
- **Design/Method:** We collected scores from undergraduate engineering and computing courses that had both synchronous exams (all students at the same time) and asynchronous exams (students choose a time). We analyzed student exam-time choice and asynchronous exam scores, using synchronous exam scores within the same course as a control variable.
- **Results:** We find that students with lower scores on synchronous exams generally elect to take asynchronous exams later and that controlling for student ability (via synchronous exams) removes 70% of the observed decline in average asynchronous exam scores over the exam period, but does not eliminate the downward trend with time.
- **Conclusions:** We conclude that self-selection effects are primarily responsible for exam-score declines over time, that exam time selection is unlikely to be a useful target for interventions to improve performance, and that there is no evidence for wide-spread collaborative cheating in the dataset under consideration.
- **Keywords:** Asynchronous, Undergraduate, Test format [syn: Exam format], Automated grading

1 Introduction

There has been significant pressure on universities to increase the number of engineers graduated each year to meet workforce needs and maintain national competitiveness, and universities have responded to the call. From 2009 to 2016, the number of students awarded bachelor's degrees in engineering among major universities in the United States has increased by roughly 50% (Gibbons, 2009; Yoder, 2016), with the larger institutions growing disproportionately quickly. A result of this growth is that in 2017, 46.6% of bachelor's degrees in engineering were awarded by just 50 out of the 305 institutions tracked by the ASEE (Yoder, 2017). This growth and concentration of students necessitates teaching techniques and tools that can maintain excellence in the presence of scale.

One aspect of teaching where scale and excellence are frequently at odds is assessment. For the past 30 years, assessing student outcomes has been recognized as the centerpiece of evaluating an engineering education and hence plays a central role in both the accreditation of engineering programs via ABET and the feedback loop used to improve classes and programs (Olds et al., 2005; Shaeiwitz, 1996; Engineering Accreditation Commission, 1998). For content-based competencies (e.g., problem solving, interpreting data, applying knowledge/skills), the most commonly mentioned assessments in the literature are paper and pencil exams (Henri et al., 2017).

Running pencil-and-paper exams for large classes (e.g., 200+ students) presents management challenges that include requesting space, printing exams, proctoring, timely grading, and handling conflict exams (Muldoon, 2012; Lee et al., 2015; Zilles et al., 2015). These logistic burdens discourage faculty from using pedagogies that have been shown to improve student learning, such as frequent testing (Bangert-Drowns et al., 1991; Leeming, 2002) and mastery learning (Pennebaker et al., 2013; Kulik et al., 1990), and have led to over-use of multiple-choice exams (Scouller, 1998; Stanger-Hall, 2012).

Computer-based exams have been proposed as means of mitigating the tension between scale and excellence in assessment in engineering classes (DeMara et al., 2016; Shacham, 1998; Zilles et al., 2015). Computer-based exams allow a broad range of questions (e.g., numeric, graphical, symbolic, programming, and drawing) to be auto-graded and to provide students immediate feedback (Carrasquel, 1985; Rytkönen & Myyry, 2014; Shacham, 1998; West, Herman, & Zilles, 2015). Several studies have demonstrated the validity of computer-based testing across a broad range of subjects (Bodmann & Robinson, 2004; Boevé et al., 2015; Bugbee, 1996; Cagiltay & Ozalp-Yaman, 2013; McDonald, 2002; Prisacari & Danielson, 2017; Zandvliet & Farragher, 1997).

Computer-based testing is particularly well suited for courses in engineering and, more generally, STEM. Significant amounts of the material in these courses have two important properties: 1) students' responses are well-suited for digitization, and 2) these responses can be graded automatically (i.e., it is possible to write a computer program that can score a student's answer) (West, Herman, & Zilles, 2015). This is especially true in analysis classes (e.g., statics, thermodynamics), but some design tasks are also amenable to this evaluation, by creating tests that evaluate whether a student's solution exhibits desired properties (e.g., whether a beam designed in a CAD tool meets given stiffness and weight criteria, or whether a program computes the right answer for a variety of inputs). Importantly, the use of computer-based assessment doesn't preclude having non-autogradeable assessments in a course; in fact, the use of computer-based testing where appropriate can free up faculty/course staff time to include more tasks that benefit from expert input (e.g., projects, lab reports, etc.) in the course (West, Silva Sohn, & Herman, 2015; Essick et al., 2016; Sanders et al., 2016).

Much of the management overhead of running exams can be alleviated, while maintaining exam security, by running computer-based exams in a centralized proctored facility (Bugbee & Bernt, 1990; DeMara et al., 2016; Rytkönen & Myyry, 2014; Zilles et al., 2018). In order to handle the varied constraints of student schedules and handle classes with more students than seats in a proctored computer testing center, common practice is to offer computer-based exams *asynchronously* (i.e., allowing students to take their exam within a given time window, usually several days) (DeMara et al., 2016; Stehlik & Miller, 1985; Zilles et al., 2018).

Since allowing students to choose their exam time is an unusual feature in traditional university environments, it is important to study and understand students' behavior under such settings. Previous work (Chen et al., 2017) has studied a set of asynchronous computerized exams with randomized questions and found that students tend to choose later time slots and exam scores generally decline throughout the exam period. Figure 1 shows these two phenomena in one asynchronous computerized exam. Unfortunately, it is unclear what is the cause of these phenomena. One possible hypothesis that deserves particular attention is that stronger students choose to take asynchronous exams whenever they feel ready while weaker students choose to take asynchronous exams later[1]. This would be consistent with the finding of a robust negative correlation between students' measured procrastination and their academic achievement (Kim & Seo, 2015).

It is important to understand what is causing asynchronous exam scores to decline over time. If there is no separate mediating variable such as student ability, then it could mean that: (1) exam time choice alone can have a detrimental impact on students' performance and it is perhaps advisable to have some intervention in place to help students to overcome non-ideal exam time choices, (2) collaborative cheating, where students who have already taken the exam share the exam questions with other students, is either not wide-spread or is ineffective.

In fact, faculty who consider the use of asynchronous computerized exams in their courses often question the potential for collaborative cheating resulting from the asynchronous nature of the exams. It seems initially reasonable that students taking the exam on the first day would tell their friends about the exam questions, giving students later in the exam period an unfair advantage. In fact, in a previous survey of undergraduate students, the most-reported cheating mechanism was that they had "received answers to a quiz or test from someone who has already taken it" for face-to-face (i.e., non-online) classes (Watson & Sottile, 2010). If such cheating were effective and wide-spread, we would expect to see exam scores increase through the course of the exam period.

This paper thus aims to test the aforementioned hypothesis that stronger students tend to choose to take asynchronous exams earlier than weaker students, and that this is primarily responsible for the observed average score declines over the exam period for asynchronous exams. Previous work attempted to address this hypothesis, however the amount of data used in the analysis was not sufficient to draw any firm conclusion (Chen et al., 2017). In this paper, with a much larger data set consisting of 81 asynchronous exams and 15 synchronous exams, we will show that this self-selection effect indeed largely explains the observed decline in exam scores, although not all of it. The resulting slope is still negative (statistically significant with $p < 0.0001$), suggesting that the benefit of collaborative cheating is overwhelmed by other unexplored factors that make the slope negative.

The remainder of the paper is organized as follows. In Section 2, we briefly describe relevant studies. In Section 3, we introduce the setting under which the data was collected, and describe the analysis procedures. We then present the results in Section 4, discuss the implications in Section 5, and point out limitations of our study in Section 6. Finally, we summarize in Section 7.

2 Literature review

Kreiter et al. (2003) conducted a set of three asynchronous exams over two days in a clinical practice course of about 200 students at a midwestern medical college. Each student was randomly assigned to take each of the three exams on one of the two days. Each of the three exams was kept the same over the

[1] By stronger/weaker students we mean students who are observed to do well/poorly in synchronous exams.

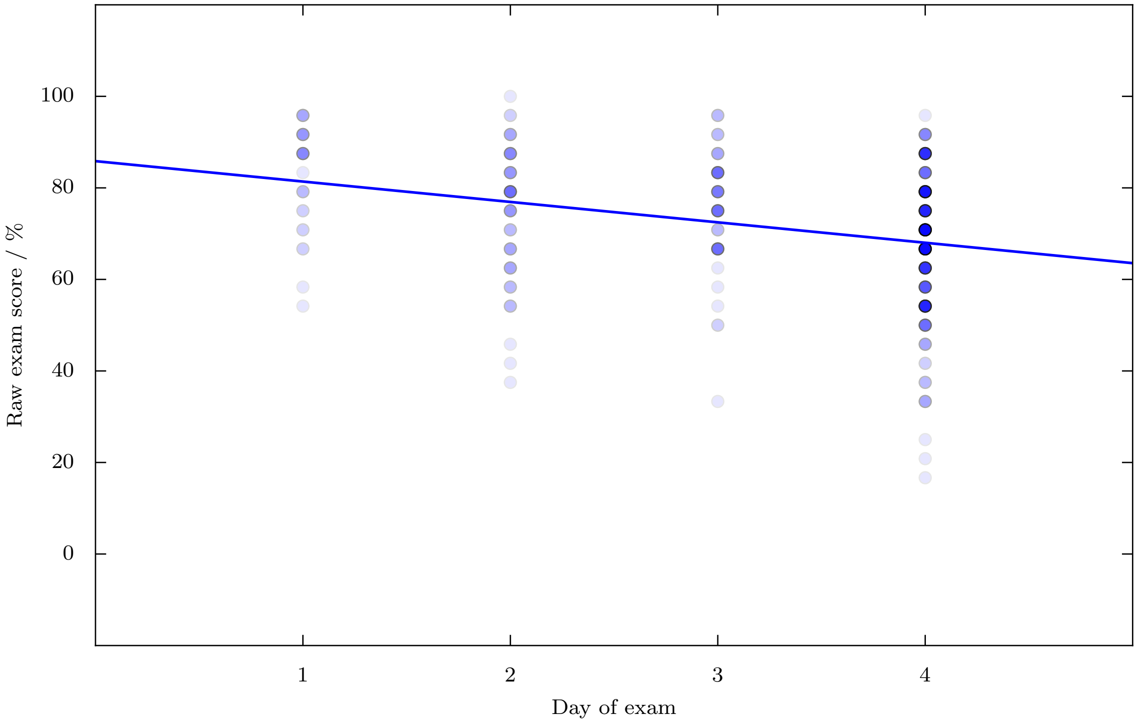

Figure 1: Example data from one asynchronous exam (Class C3, Asynchronous Exam 4) that was conducted over a 4-day period. Students' raw scores on the asynchronous exam are plotted against the day on which they took the asynchronous exam, with the intensity of each circle being proportional to the number of student with that score on that day. The straight line is the OLS (ordinary least squares) regression line of the exam score against the day of exam, revealing in this case a negative correlation between the day on which the student choose to take the asynchronous exam and their score. This asynchronous exam has one of the more negative slopes in our data set, and we chose it here because the highly negative slope is easy to discern.

two day period. Kreiter et al. (2003) reported no significant differences between students' performance on day 1 versus day 2 for each exam. This suggests that using a single test form for asynchronous exams spread over two days does not compromise the integrity of the exam results as long as students are randomly assigned when to take the exam. There are a few studies that describe experiments with computer-based exams in lab sections (i.e., asynchronous, but at times that students chose before the start of the class) for computer science courses, but none of them has reported score trends over time (Jacobson, 2000; Califf & Goodwin, 2002; Barros et al., 2003; Bennedsen & Caspersen, 2006).

Between 2003 and 2005, Burns et al. (2007) ran 13 asynchronous computer-based exams for a microscopic anatomy course with about 150 students for each exam. Each exam was identical for all students, and students were allowed to choose when to take their exams. Burns et al. (2007) found that in general students who choose to take exams early performed better than those who choose to take exams later.

Chen et al. (2017) analyzed a set of 93 asynchronous computer-based exams among 9 engineering courses between 2015 and 2016 in a large public research university. In this dataset, most exams were slightly different for each student, and students were allowed to choose when to take their exams. Chen et al. (2017) reported that students tended to choose later exam time slots and that those students who elected to take the exam later performed worse on average than those who chose earlier exam time slots.

During Spring 2009, Wagner-Menghin et al. (2013) carried out an asynchronous pencil-and-paper exam consisting of only multiple-choice questions. A total of 671 students were initially assigned to four time slots and allowed to reschedule their exam time as they wished. By comparing the difficulty of reused items using the Rasch Model, Wagner-Menghin et al. (2013) observed that reused items became easier after their first use.

The relationship between procrastination and academic performance has been extensively studied (see Kim & Seo (2015) and studies cited therein). Procrastination is widespread among college students, with estimates of up to 80–90% of students engaging in procrastination (Steel, 2007). There has been debate as to whether procrastination should be regarded as a task-specific behavior or as a personality trait that is stable across time and context (Schouwenburg, 2004), although it is now more common to adopt the trait point of view (Kim & Seo, 2015). While there has been significant disagreement in the literature between studies finding that procrastination does not affect academic performance (e.g., Seo, 2011; Solomon & Rothblum, 1984) and those finding it does (e.g., Aremu et al., 2011; Balkis et al., 2013), meta-analyses show that many of these differences are due to underestimates of correlations from the use of self-reported data on both procrastination and performance (van Eerde, 2003; Kim & Seo, 2015). The best overall estimates show an average correlation of $r = -0.39$ (95% CI $[-0.65, -0.13]$) between measured procrastination and measured performance (Kim & Seo, 2015, Table 3).

3 Methods

3.1 Data collection

The data was collected in a large R1 university in the US during the Spring 2015, Fall 2015, Spring 2016, and Fall 2016 semesters. The asynchronous exam data was drawn from exams that were held in the Computer-Based Testing Facility (CBTF) (Zilles et al., 2018) and administrated via the PrairieLearn system (West, Herman, & Zilles, 2015). The synchronous exam data was provided by the corresponding instructors.

The CBTF is a computer lab with 85 seats for students and another 4 seats in a reduced-distraction environment for students registered with the disability resource center. Each of the computers is outfitted with a privacy screen that prevents test takers from reading off the screens of neighboring computers, and the networking and file system are strictly controlled (Zilles et al., 2018). During the period studied, the facility was open and proctored 10–12 hours a day, 7 days a week to accommodate two to four thousand exams per week. Students were not permitted to take written notes, photos, or other records into or out of the exam room. At their scheduled exam time, students had their identity checked by a proctor and were randomly assigned to a computer (to deter coordinated cheating).

Exams within the CBTF were typically administered as follows (Zilles et al., 2018): Classes assigned a 3–5 day period for the students to take an exam depending on the class size; longer exam periods were used during finals week. Students were free to reserve any time during the exam period, provided that there were slots available at that time. Sign-ups for exams typically began two weeks before the exam period began. Generally, the periods of exams from different classes overlapped each other and the CBTF was almost always running several distinct exams concurrently.

PrairieLearn is an online problem posing system that permits the specification of *automatic item generators (AIG)* (Attali, 2018), each of which is capable of generating a range of parameterized problem instances (West, Herman, & Zilles, 2015). A variety of problem types can be specified, including but not limited to numeric, graphical, symbolic, programming, and drawing problems. For exams, PrairieLearn selected random problem generators from a pool of available generators and randomly generated problem instances from those generators to meet instructor-defined coverage and difficulty criteria. Students sitting next to each other in the CBTF were typically taking exams from different courses, but even if they were taking the same exam, they generally had different sets of parameterized questions or the same set of questions with different parameters. PrairieLearn also supports allowing students to have multiple attempts at each question with a partial-credit schedule controlled on a per-question basis.

For each student taking an exam in the CBTF, PrairieLearn logged all the submissions the student made during the exam period and calculated and stored the final score based on the instructor's multiple-attempts scoring scheme.

3.2 Data description and preprocessing

The courses studied are drawn from the introductory sequences in Mechanical Engineering (statics, dynamics, and strength of materials) and Computer Science (intro to programming, computer organization, and system programming). The courses in Mechanical Engineering are primarily engineering sciences focused while the courses in Computer Science involve a combination of computing science and code writing. Since these courses are introductory, course material is mostly fixed with little variation, even when different instructors teach the courses. The asynchronous exam material is usually first developed by a single instructor over several semesters and then augmented by other instructors.

For each class in each semester, we have obtained the information of all of the asynchronous exams in the form of **(class id, exam id, start date, end date)**. The class id is a unique identifier to differentiate between each class in each semester. The exam id is a unique identifier for each exam. The start date is the first calendar day when students can take the exam. The end date is the last calendar day when students can take the exam. We refer the time period defined by the start date and the end date as the *exam period.* The synchronous exams' information is in the form of **(class id, exam id)**, where class id and exam id are the same as in the asynchronous case.

With IRB approval, we obtained all of the students' asynchronous exam records in the CBTF as well as their synchronous exam records outside the CBTF for each class in each semester. Each asynchronous exam record has the form **(exam id, student id, score, date, day, hour)**. The exam id is the same as defined above. The student id is a unique identifier for a student regardless of class. The score is a real number ranging from 0 to 100. The date is the calendar date when the student has taken the exam. The day is an integer ranging from 1 to the length of the exam period[2]. The hour is an integer ranging from 0 to 23. Each synchronous exam record has the form **(exam id, student id, score)** where they are the same as in the asynchronous case.

Given the raw asynchronous exam data, we filtered it as follows:

1. We excluded all optional second chance asynchronous exams that allowed students to replace part or all of an earlier asynchronous exam score by taking a second equivalent asynchronous exam at a later date.

2. We excluded students who have taken less than 50% of the non-second chance asynchronous exams in order to avoid including course staff members engaged in exam checking and students who dropped early in the semester.

3. We excluded students who do not have the corresponding synchronous exam records.

4. We excluded asynchronous exam records that are outside the corresponding exam periods[3].

5. We excluded asynchronous exams whose score distribution's kurtosis is more than 10. These exams have an unusually high number of scores that are more than several standard deviations away from the mean.

The first three mostly aim to filter out things that are irrelevant to the analysis. The forth filter is due to the fact that these asynchronous exam records are often outliers for the analysis where most students take exams within the exam period. The fifth filter is to avoid asynchronous exams that have many large deviations from the mean, which could have unstable effects on the regression coefficients. We also applied the fifth filter to the synchronous exam data, and none of the synchronous exams were excluded.

We examined the statistical characteristics of exam score distributions after the above filtering for both synchronous and asynchronous exam scores, and found that they are similar to each other and match the characteristics of typical exam score distributions reported in the literature. See Appendix A for more details.

The above filtering resulted in 26,139 exam records from 81 asynchronous exams and 5,534 exam records from 15 synchronous exams. A summary of the data is shown in Table 1. For courses with only one synchronous exam, these exams were either the final exam or a midterm towards the end of the semester. For the course with three synchronous exams, these exams were midterms and final.

Unfortunately the data collected does not contain demographic information for the students, thus our analysis will focus on the population as a whole. As a rough estimate of the demographic composition of the students in the data, we reported the demographic information of undergraduates who graduated with degrees in each discipline during the calendar year of 2018 in Table 2.

In order to analyze different exams with different score distributions together, we standardized all of the exam scores to z-scores on a per exam basis. Essentially, the standardized score measures how

[2] The exam period length varies across exams. It is generally 3 to 5 days, and up to 8 days for final exams.

[3] In exceptional circumstances such as long-term illness, students take an asynchronous exam outside the normal exam period.

Course and semester	Discipline	DFW Rate	Number of students	Number of asynchronous exams included	Number of asynchronous exams excluded	Number of asynchronous exam records excluded	Number of synchronous exams
Class A2	ME	12.6%	566	5	1	22	3
Class B2	ME	10.2%	230	7	0	0	1
Class B3	ME	11.6%	345	6	0	5	1
Class B4	ME	12.8%	181	7	0	3	1
Class C2	CS	20.0%	173	5	1	8	1
Class C3	CS	17.9%	325	4	3	0	1
Class C4	CS	18.8%	292	3	4	1	1
Class D1	ME	9.5%	477	2	0	0	1
Class E1	CS	18.8%	324	7	1	1	1
Class E2	CS	11.8%	352	8	0	24	1
Class E3	CS	14.9%	187	9	0	6	1
Class E4	CS	13.0%	369	8	1	0	1
Class F4	CS	2.0%	581	10	0	68	1
Total				81	11	138	15

Table 1: Summary information of the data used in the analysis. Each course is indicated by a letter (A–F) and a number for the semester (1 = Spring 2015, 2 = Fall 2015, 3 = Spring 2016, 4 = Fall 2016). For the discipline, CS stands for computer science and ME stands for mechanical engineering. Some courses only started using the CBTF/PrairieLearn environment in later semesters and some courses stopped running synchronous exams in later semesters. There is only one column for synchronous exams since none of them are excluded. This table only includes non-second chance exams. By "exams" in the header of the table we mean a unique exam available to all the students of the course. By "exam records" in the header of the table, we mean the record of an individual student taking an exam. See Section 3.2 for more details.

Discipline	Female	International	Hispanic	Asian American	Black	White	Other
Mechanical Engineering	15.1%	20.9%	7.5%	15.6%	0.9%	50.7%	4.4%
Computer Science	21.1%	33.8%	2.1%	35.1%	0.4%	26.9%	1.7%

Table 2: Demographic information of undergraduates who graduated with degrees in each discipline during the calendar year of 2018. The Other column in the table means people who are not Hispanic, Asian American, Black, or White.

many standard deviations a particular student's score is away from the mean. We refer to the exam scores after the standardization as the *standardized score*. Additionally, for each class and semester, we define the *synchronous score* to be the average of all of the standardized scores for synchronous exams. For example, if a class in a particular semester has 3 synchronous exams and a student's standardized score for the 3 synchronous exams are -0.5, 1.0, and 1.0, respectively, then the synchronous score of the student is 0.5 for that particular class in that semester.

Because asynchronous exams have different exam period lengths and the CBTF operation hours have changed slightly from semester to semester, scaling is necessary for the analysis to be meaningful. Specifically, we scaled the day of exam period to have the range [0, 1] where the first day of the exam period is represented by 0 and the last day of the exam period is represented by 1. We scaled the hour of day to have the range [0, 1], where the hour of the first asynchronous exam of each day is represented by 0 and the hour of the last asynchronous exam of each day is represented by 1. We refer to the day of exam period and hour of day after the scaling as *scaled day* and *scaled hour*, respectively.

3.3 Analysis

Our analysis consists of four parts. We first show the general trend of students' exam time choices in Section 4.1. We will provide the distribution of exam time choices for a typical asynchronous exam as well as the overall trend. We will then disaggregate students into three groups based on their synchronous score, to examine if there is any difference between students of different ability levels.

The second part of the analysis consists of studying the distribution of correlation coefficients between four pairs of measures. The first pair is between standardized scores of an asynchronous exam and corresponding synchronous score. The purpose of this pair is to understand how well asynchronous exam scores correlate with synchronous exam scores. To understand whether the correlations between the first pair is reasonable, the ideal comparison would be the correlation coefficients between scores of different synchronous exams for each class and each semester. Unfortunately most classes only had one synchronous exam as Table 1 shows. Instead, as the second pair, we compute correlations between the scores of asynchronous exams that belong to the same class and same semester. The third pair is between scaled day of an asynchronous exam and corresponding synchronous score. It is important to study this pair because our hypothesis suggests that this pair should be mostly negative for most asynchronous exams. For comparison, we will also study the correlation coefficients between standardized asynchronous score and scaled day of asynchronous exams as the forth pair. We present the result of this analysis in Section 4.2.

Our third analysis quantifies the effect of when students choose to take asynchronous exams on their performance. We use the same analysis techniques as in previous work (Chen et al., 2017). Specifically, we regress the standardized score against scaled day and scaled hour as follows:

$$z_{ik} = \alpha_k + \beta_k d_{ik} + \gamma_k h_{ik}, \tag{1}$$

where z_{ik}, d_{ik}, h_{ik} are observed values from the data, defined as follows:

- z_{ik}: the standardized score of student i on asynchronous exam k,
- d_{ik}: the scaled day of student i taking asynchronous exam k,
- h_{ik}: the scaled hour of student i taking asynchronous exam k,

and $\alpha_k, \beta_k, \gamma_k$ are the regressors that we want to calculate, defined as follows:

- α_k: the intercept for asynchronous exam k,
- β_k: the coefficient that characterizes the effect of scaled day on scores for asynchronous exam k,
- γ_k: the coefficient that characterizes the effect of scaled hour on scores for asynchronous exam k.

The slope β has units of standard deviation per exam period, so a value of $\beta = -0.5$ would mean, roughly speaking, that the student asynchronous exam scores decline by one half of a standard deviation from the first day to the last day of the asynchronous exam. The slope γ has units of standard deviation per day, so a value of $\gamma = -0.1$ would mean, roughly speaking, that student asynchronous exam scores decline by one tenth of a standard deviation from the first hour of each day to the last hour of each day. We refer to this regression as the "uncontrolled regression" and report the results in Section 4.3.

Since there are multiple independent variables in the regression, we will report the maximum variance inflation factor (VIF) (Kutner et al., 2004) for each of the relevant regressors to examine if multicollinearity can undermine the interpretability of the coefficients. The VIF for a particular regressor δ_k is defined as:

$$\mathrm{VIF}_{\delta_k} = \frac{1}{1 - R^2_{\delta_k}}, \tag{2}$$

where $R^2_{\delta_k}$ is the coefficient of determination (correlation coefficient squared) for the regression of δ_k on the other regressors. VIF_{δ_k} is a multiplicative term in the calculation of $\hat{\sigma}^2_{\hat{\delta}_k}$, which essentially quantifies how much inflation in the observed variance of δ_k is contributed by correlation among regressors (O'Brien, 2007). The lower bound of VIF_{δ_k} is 1 when $R^2_{\delta_k} = 0$ and there is no upper bound. A large VIF for a regressor indicates that the observed variance of the coefficient of the regressor is inflated substantially and the resulting confidence interval of the coefficient of the regressor is much wider than the case when there is little correlation among regressors.

In the forth analysis, we add one additional regressor, δ_k, in the uncontrolled regression:

$$z_{ik} = \alpha_k + \beta_k d_{ik} + \gamma_k h_{ik} + \delta_k c_{ik}, \tag{3}$$

where c_{ik} is the observed synchronous score of student i corresponding to asynchronous exam k, and δ_k is the coefficient that quantifies the effect of synchronous score. The slope δ has units of standard deviation in asynchronous exam per standard deviation in synchronous exam, so a value of $\delta = 0.75$ would mean, roughly speaking, that a student whose synchronous score is 2 standard deviations higher than the average will get an asynchronous score which is 1.5 standard deviations higher than the average. This process of adding potential confounding factors to the regression formula to see if the coefficient of interest changes greatly is a standard procedure to verify confounding factors (Kleinbaum et al., 1998, 1982). We refer to this regression as the "controlled regression" and reported the results in Section 4.4.

4 Results

4.1 Students' exam time choices

We observed that more students choose to take asynchronous exams on later days of the exam period and at later hours of each day (especially on the last day), even though they are allowed to choose when

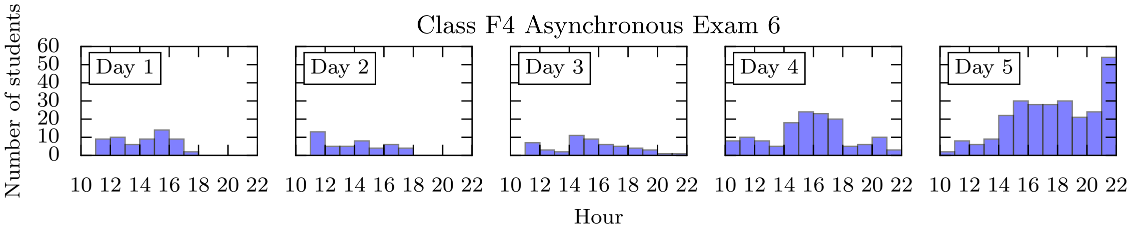

Figure 2: Example of exam record distributions during the exam period for one exam. Students choose to take exams on later days, especially the last day and on later hours of the last day.

to take their asynchronous exams up to 2 weeks before the exam period begins. We plotted an example of the day and hour distributions for one exam in Figure 2. As the figure shows, a large majority of the students took the exam on the last day. While the hour distribution of the first few days are somewhat spread out, the hour distribution of the last day is biased toward later hours, especially the last hour.

We plotted the distribution of the student asynchronous exam records of each asynchronous exam with respect to scaled day in Figure 3. As the figure suggests, students overwhelmingly choose to take asynchronous exams toward the end of the exam period. The few segments that drop at the end correspond to final exams where students may have wanted to leave the campus early.

To have a better understanding of how students' ability relates to their exam time choices for asynchronous exams, we separated students to "High", "Mid", and "Low" equal-sized groups on a per class basis based on their synchronous score, and plotted the distribution of student exam records of all asynchronous exams aggregated for each group in Figure 4. We aggregated different asynchronous exams by binning points on scaled day to intervals $[0, 0.25), [0.25, 0.50), [0.50, 0.75), [0.75, 1.00), [1.00, 1.00]$, and averaged their y-axis values. We then plotted the averaged value on the left side of each interval. As the result figure shows, "High" students' exam time choice is the most evenly distributed, and the distributions concentrate more at the end of the exam period as we move from "High" to "Low".

We also plotted the distribution of student asynchronous exam records for each asynchronous exam on the last day with respect to scaled hour in Figure 5. As the figure shows, there is a bias toward the later hours on the last day. These figures indicate that the example in Figure 2 is indeed representative.

4.2 Correlation analysis

We plotted a series of distributions of correlation coefficients and how significant they are in Figure 6. Specifically, we plotted the number of significant ($p < 0.05$) correlations in light green and the number of non-significant ones in dark blue as a stacked bar.

The first subplot shows the distribution of correlation coefficients between synchronous score and standardized score of asynchronous exams. Each correlation coefficient is calculated using standardized scores of one asynchronous exam and the corresponding synchronous scores. As the figure shows, all of the pairs are positively correlated and the coefficients center around 0.4 to 0.5 (mean $r = 0.432$, 95% CI $[0.401, 0.462]$). Almost all of them are significant at ($p < 0.05$) level. As a reference, the distribution of correlation coefficients between the scores of asynchronous exams that belong to the same class and same semester is plotted in the second subplot of Figure 6. As the subplot shows, all of the correlation coefficients are positive and centered around 0.3 to 0.4 (mean $r = 0.330$, 95% CI $[0.313, 0.346]$). Most of them are significant at ($p < 0.05$) level. We take this as positive evidence that the correlations between

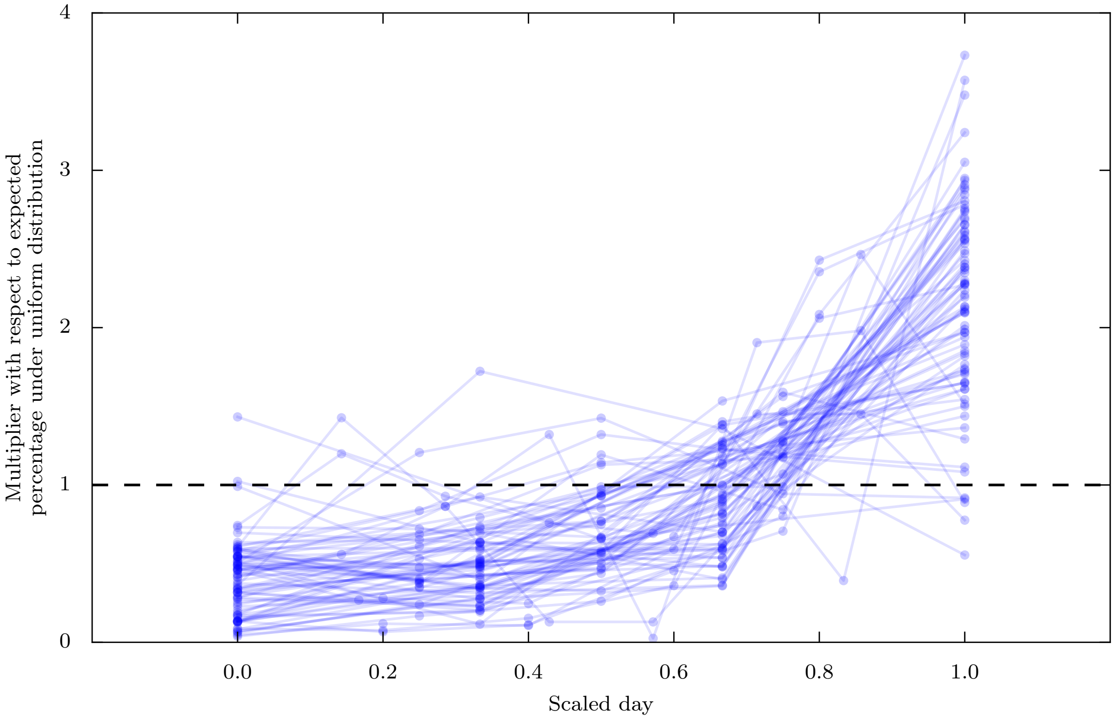

Figure 3: Distribution of student exam records over the exam period for all asynchronous exams. Each series of connected line segments represents the distribution for a single asynchronous exam. The horizontal axis shows the scaled day, so 0 is the first day of each exam and 1 is the last day. We scaled the vertical axis values so that all the line segments would overlap with the dashed line if student exam records were uniformly distributed over the exam period.

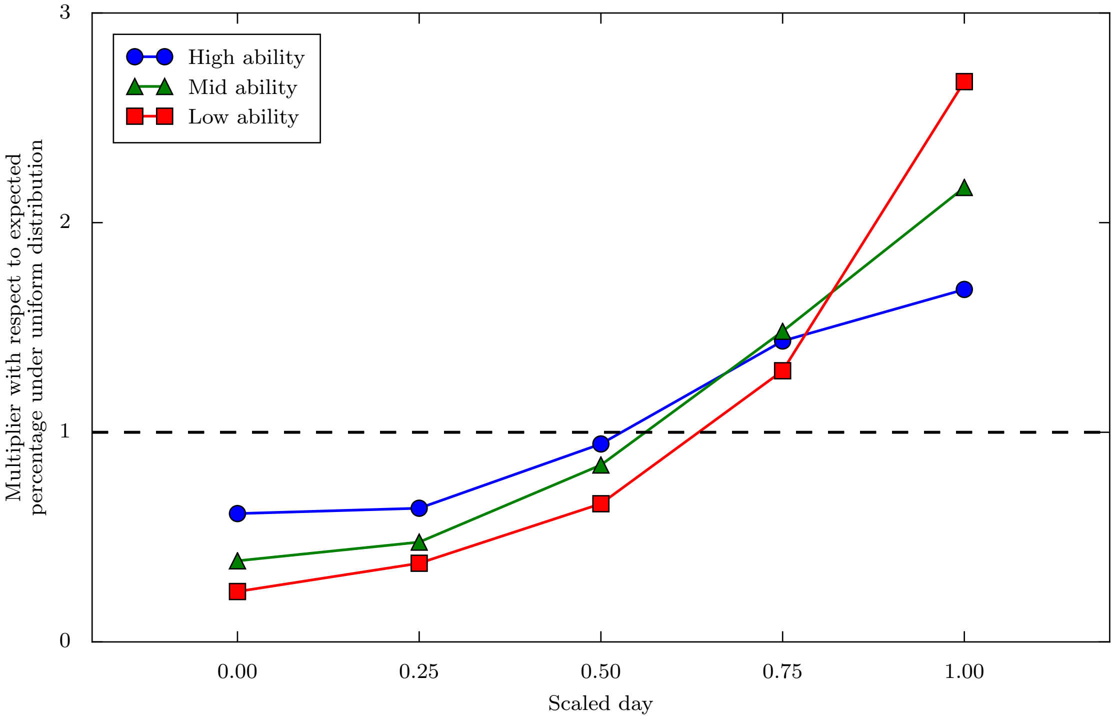

Figure 4: Distribution of student exam records over the exam period for all asynchronous exams for "High", "Mid", and "Low" students separately. Axes are the same as in Figure 3.

synchronous scores and standardized scores of asynchronous exams are reasonable.

The distribution of correlation coefficients between synchronous scores and scaled days of asynchronous exams is plotted in the third subplot of Figure 6. As the subplot suggests, they are mostly negatively correlated at around -0.3 to -0.2 (mean $r = -0.215$, 95% CI $[-0.233, -0.198]$). A few of the correlation coefficients are actually positive but relatively close to 0. Most of the correlation coefficients are significant ($p < 0.05$) except those near 0. This is consistent with our hypothesis that weaker students choose to take asynchronous exams on later days of the exam period. For comparison, we plotted the distribution of correlation coefficients between standardized asynchronous scores and scaled days of asynchronous exams on the last subplot of Figure 6. Most correlation coefficients are significant ($p < 0.05$) centering around -0.2 to -0.1 (mean $r = -0.128$, 95% CI $[-0.149, -0.107]$) and a slightly higher number of correlation coefficients are positive compared to the previous case. Overall, all the observations in the figure are consistent with our hypothesis and previous observations.

4.3 Uncontrolled regression

The maximum VIF for both β_k and γ_k among all the asynchronous exams is 1.295. These VIFs are reasonably small, and thus multicollinearity is not a concern.

We visualize each asynchronous exam's β_k and γ_k with their 95% confidence intervals on a pair of forest plots in the upper parts of Figures 7 and 8, respectively. A forest plot is a standard meta-analysis visualization tool (Cooper et al., 2009, Chapter 26) that shows effect sizes for many different studies together with their confidence intervals (horizontal bars) and an indicator of study reliability (area of

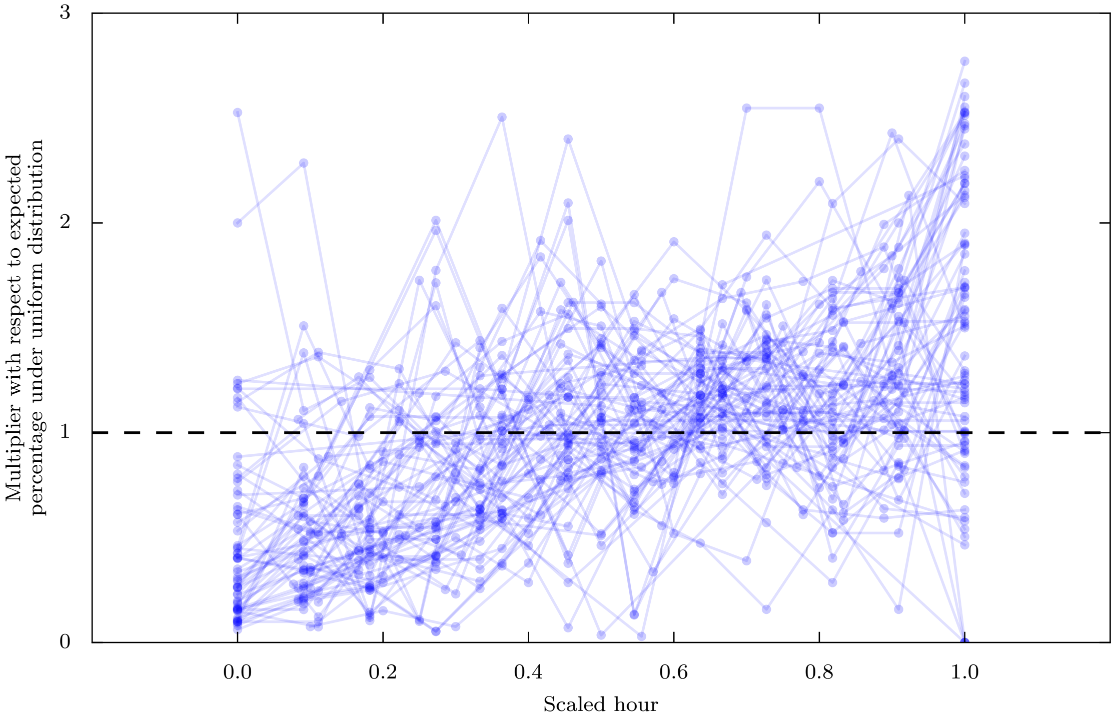

Figure 5: Distribution of student exam records over the operation hours of the last day for all of the asynchronous exams. Each series of connected line segments represents the distribution for a single asynchronous exam. The horizontal axis shows the scaled hour, so 0 is the hour of the first exam record on the last day of each exam and 1 is the hour of the last exam record on the last day. We scaled the vertical axis values so that all the line segments would overlap with the dashed line if student exam records were uniformly distributed over the operation hours of the last day of each exam.

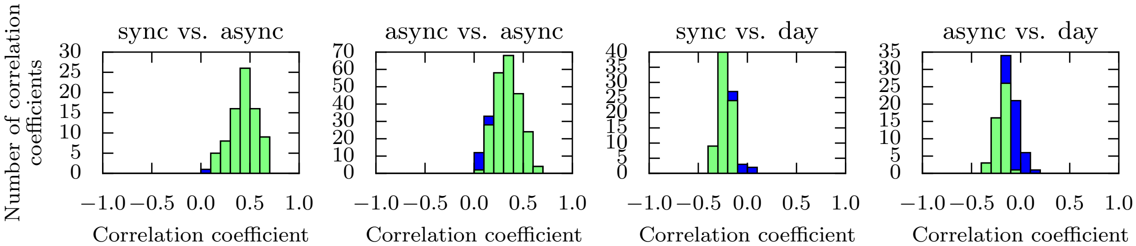

Figure 6: Distribution of correlation coefficients between synchronous score (sync), standardized score of asynchronous exams (async), and scaled day of asynchronous exams (day). Note that the third subplot involves *synchronous* scores correlated with the scaled day of the corresponding *asynchronous* exams. We plotted the total number of correlation coefficients in each bin in dark blue and the number of significant ($p < 0.05$) correlation coefficients in each bin in light green.

circles).

The two-tailed significance levels (p-values) for the slopes being non-zero are shown on the right hand side of the figures. For the effect of scaled day (β_k), none (0%) of the slopes are statistically significantly positive ($p < 0.05$), 6 (7%) of the slopes are non-significantly positive ($p > 0.05$), 46 (57%) of the slopes are statistically significantly negative and 29 (36%) are non-significantly negative. For the effect of scaled hour (γ_k), 2 (2%) of them are significantly positive, 17 (21%) are non-significantly positive, 18 (22%) are significantly negative and 44 (54%) are non-significantly negative. To obtain an aggregated measure of β and γ, we adopted the standard meta-analysis techniques described in Cooper et al. (2009, Chapter 14). Though there is no clear consensus in the meta-analysis community on how to combine regression slopes in the general case (Cooper, 2016), previous works (Becker & Wu, 2007; Cooper, 2016) suggest that under the condition when both the dependent and independent variables are measured similarly across studies, the regression slopes can be safely combined by treating them as a simple effect. This is the approach that we adopted. Details about these techniques are included in Appendix B. One of the important assumptions that the techniques make is that the data is normally distributed. We plotted the normal probability plots of β_k and γ_k for the uncontrolled regression at the top of Figure 9. As the figure suggests, they are both approximately normally distributed. Using meta-analysis techniques (see Appendix B for more details), we obtained aggregate $\beta = -0.390$ (95% CI $[-0.453, -0.328]$) which is significantly negative ($p < 0.0001$) and aggregate $\gamma = -0.181$ (95% CI $[-0.240, -0.121]$) which is significantly negative ($p < 0.0001$). The aggregate β and γ are plotted as diamonds at the bottom of the upper parts of Figures 7 and 8. We have also computed the R^2 for each asynchronous exam, and the average R^2 for the uncontrolled regression is 0.035.

To understand the impact of filtering students who have dropped the course, we calculated the aggregate β and γ with students who have dropped the course. In this case, $\beta = -0.397$ (95% CI $[-0.462, -0.332]$) and $\gamma = -0.193$ (95% CI $[-0.256, -0.130]$). Both values become slightly more negative as compared to those in the previous paragraph, calculated without students who have dropped the course. However, they are still well within each other's confidence intervals.

4.4 Controlled regression

The maximum VIFs for β_k, γ_k, and δ_k are 1.366, 1.335, and 1.237 respectively. These VIFs are again reasonably small, and thus multicollinearity is not a concern. The distributions of β_k and γ_k are shown in the lower part of Figures 7 and 8, respectively, where each β_k and γ_k is plotted with its 95% confidence interval.

We again show the two-tailed significance levels (p-values) for the slopes being non-zero on the right hand side of the figures. For the effect of scaled day (β_k), 5 (6%) of them are statistically significantly positive ($p < 0.05$), 17 (21%) are non-significantly positive ($p > 0.05$), 13 (16%) are significantly negative and 46 (57%) are non-significantly negative. For the effect of scaled hour (γ_k), 6 (7%) are significantly positive, 38 (47%) are non-significantly positive, 4 (5%) are significantly negative and 33 (41%) are non-significantly negative. We plotted the normal probability plots of β_k, γ_k and δ_k for the controlled regression at the bottom of Figure 9. As the figure suggests, they are approximately normally distributed. The same methodology as in the uncontrolled regression was used to calculate the aggregate β and γ, and we obtained $\beta = -0.115$ (95% CI $[-0.168, -0.063]$) which is significantly negative ($p < 0.0001$) and $\gamma = 0.019$ (95% CI $[-0.033, 0.071]$) which is non-significantly positive ($p > 0.1$). The aggregate β and γ are plotted as diamonds at the bottom of the lower parts of Figures 7 and 8. We have also computed the R^2 for each asynchronous exam, and the average R^2 for the controlled regression is 0.217.

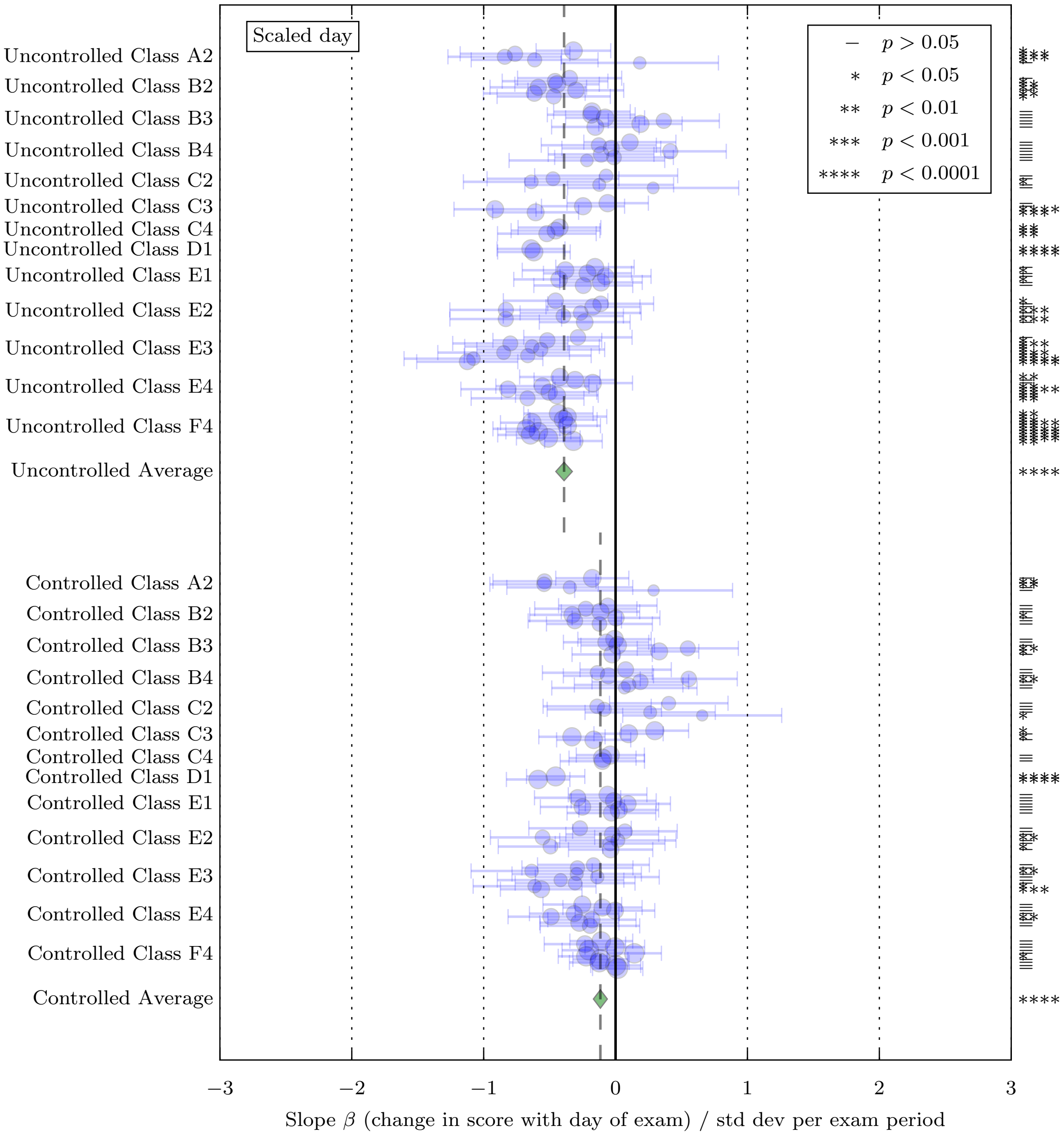

Figure 7: Joint forest plot that compares the slopes β_k of standardized asynchronous score versus scaled day under uncontrolled (top) and controlled (bottom) conditions. Each circle represents the slope of one asynchronous exam and they are grouped by course and semester as shown on the left. The area of each circle is proportional to the weight $w_k = 1/v_k$ of the corresponding exam in the meta-analysis, and the horizontal error bar is the 95% confidence interval for the slope. The diamond at the bottom of the upper part of the figure represents the aggregate population slope $\beta = -0.390$ (95% CI $[-0.453, -0.328]$) for all of the exams under the uncontrolled analysis and the diamond at the bottom of the lower part of the figure represents the aggregate population slope $\beta = -0.115$ (95% CI $[-0.168, -0.063]$) for all of the exams under the controlled analysis. The width of the diamonds specify the 95% confidence intervals of the estimates. The two-tailed significance levels of the slopes away from zero are shown on the right of the figure as a number of stars.

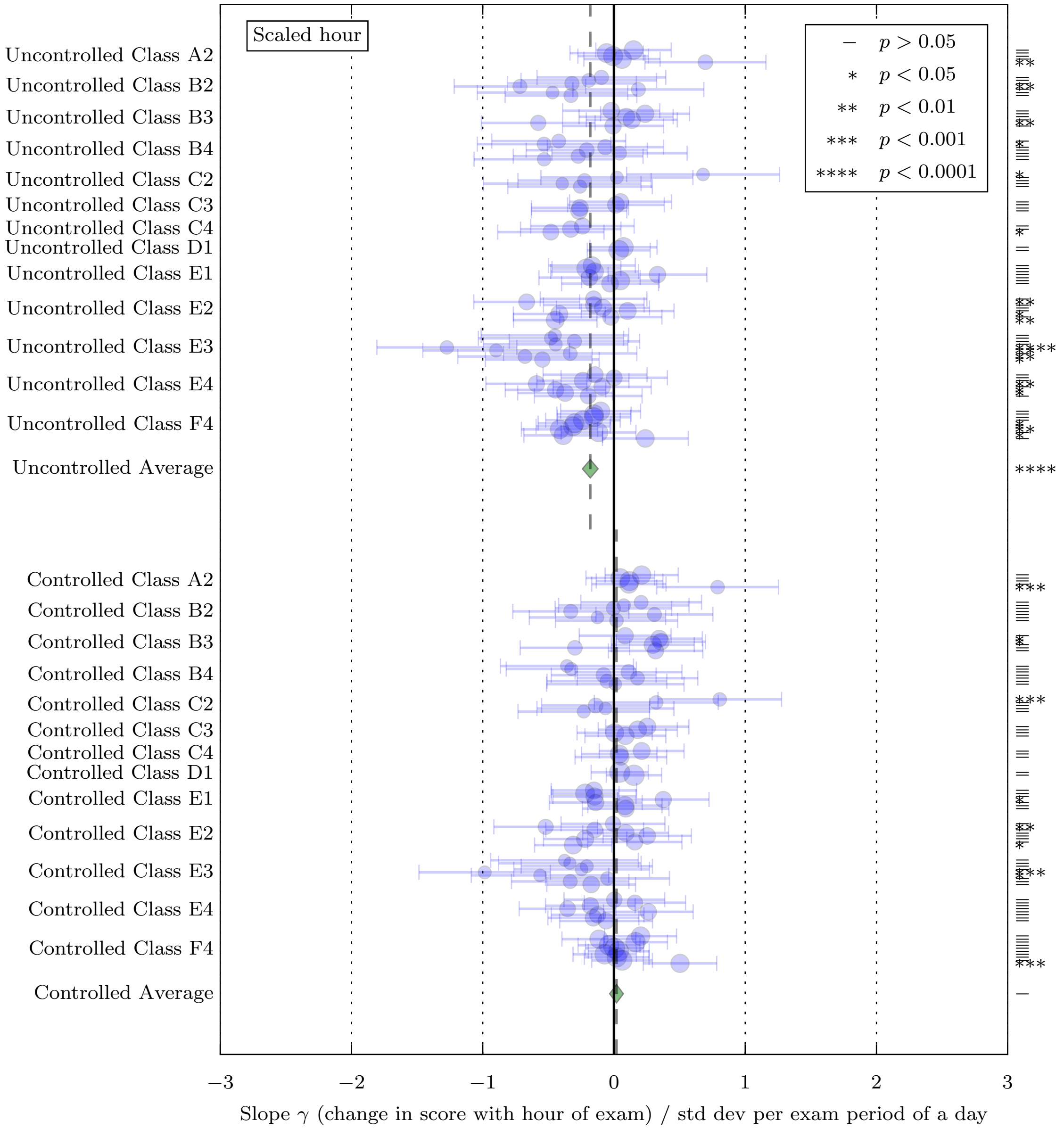

Figure 8: Joint forest plot that compares the slopes γ_k of standardized asynchronous score versus scaled hour under uncontrolled (top) and controlled (bottom) conditions. The aggregate population slope under the uncontrolled analysis is $\gamma = -0.181$ (95% CI $[-0.240, -0.121]$) while the aggregate population slope under the controlled analysis is $\gamma = 0.019$ (95% CI $[-0.033, 0.071]$). See Figure 7 for a description of the figure format.

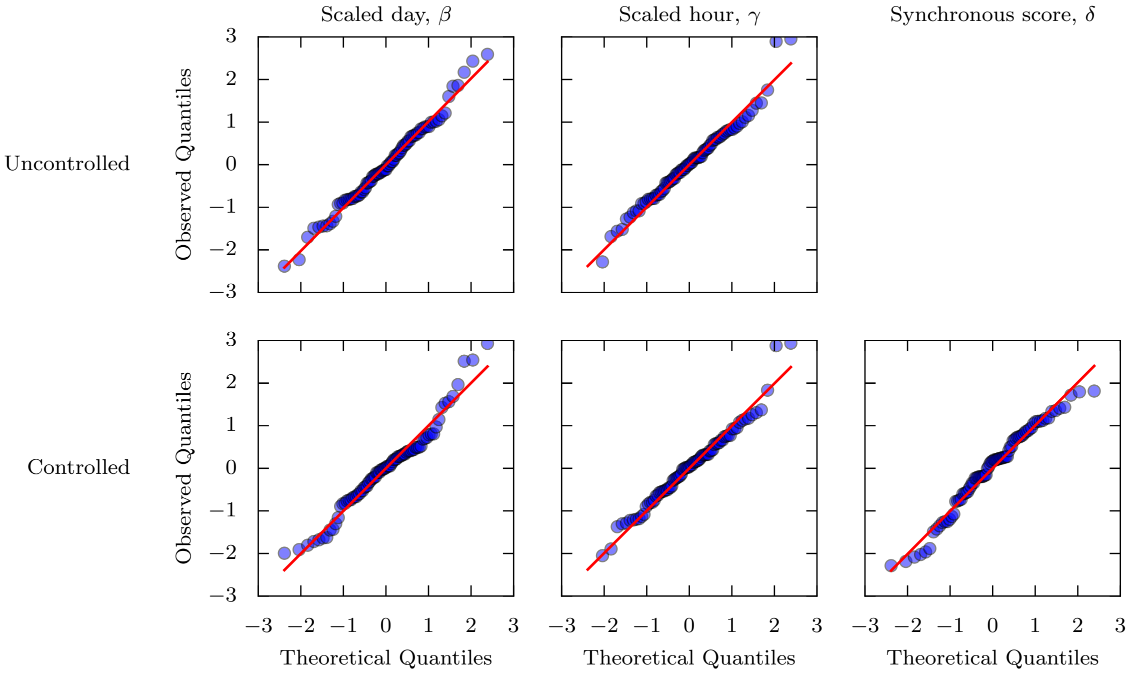

Figure 9: Normal probability plots for the slopes β_k, γ_k and δ_k from all exams under uncontrolled and controlled settings. These plots show that the slopes are approximately normally distributed.

5 Discussion

The purpose of this paper is to understand the previously-observed phenomenon where, when given a choice of when to take an exam, many students choose to take an exam toward the end of the exam period and on average perform worse compared to students who choose to take the exam earlier (Chen et al., 2017). Our hypothesis for the cause of this phenomenon is that weaker students tend to put off the exam while stronger students tend to take the exam with a more uniform distribution of times. That is, we hypothesize that weaker students procrastinate more.

To test our hypothesis, we investigated data from courses that have run both synchronous exams (all students take the exam at the same time) and asynchronous exams (students can choose when to take the exam within a short time period, usually 3–5 days) in the same semester. The synchronous exams are typically midterms and finals, which have more weight in the course grades than asynchronous exams. The synchronous exams typically occurred chronologically in the middle or after the asynchronous exams. We found that students' choices of exam time negatively correlate with their scores on synchronous exams ($r = -0.215$, 95% CI $[-0.233, -0.198]$), meaning that students with lower scores on synchronous exams tend to choose to take asynchronous exams later. This is consistent with the best estimates of the correlation between measured procrastination and measured academic achievement of $r = -0.39$ (95% CI $[-0.65, -0.13]$) (Kim & Seo, 2015). This suggests that student ability may in fact be the mediating variable between the exam time choice and exam performance.

By using students' scores on synchronous exams as a control (Equation 3), we found that the magnitude of the observed decline in asynchronous exam scores throughout the exam period reduces considerably. As Figures 7 and 8 show, when synchronous score is added to the regression, both β and γ shift substantially to the right. Specifically, β moves from -0.390 to -0.115, which corresponds to about a 70% reduction in its value while remaining significantly below zero; γ changes from -0.181 to 0.019, which is a change away from significantly negative to non-significance. These drastic changes suggest that the observed decline of asynchronous exam score over the exam period can be largely attributed to the confounding factor of synchronous exam score. In other words, students' ability, as measured by synchronous exam score, explains the majority of the declining trend. However, the coefficient corresponding to the decline of student scores over time is still statistically-significantly negative even when students' ability is taken into account, suggesting that there are other factors causing students' score to decline over time and that there are no countervailing effects (such as widespread collaborative cheating) that are large enough to cause average scores to rise over time.

While our results suggest that students' performance in synchronous exams play an important role in choosing exam times in the asynchronous setting, our data does not reveal why weaker students choose these later time slots. One hypothesis is that they choose later times because it gives them the most time for studying, although longer study time does not necessarily result in better performance (Kember et al., 1995; Plant et al., 2005). Another hypothesis is that they procrastinate and end up with an equal amount of study time as for a synchronous exam, since it is well known that procrastination negatively correlates with academic performance (Richardson et al., 2012). Either way, our results suggest that interventions trying to directly prevent students from scheduling their exams later will not necessarily improve students' performance as their choice of exam time is closely related to their performance in synchronous exams.

Importantly, our study also provides two pieces of evidence supporting the use of asynchronous computerized exams as a viable alternative for synchronous pencil-and-paper exams. The first is the observed positive correlations between asynchronous exam scores and synchronous exam scores. The second is that the decline of scores over time is not fully neutralized even when synchronous scores are controlled for, suggesting that wide-spread collaborative cheating is not present in the asynchronous setting. These two observations support the hypothesis that asynchronous computerized exams with proper proctoring and randomization can achieve integrity similar to synchronous pencil-and-paper exams.

Our study suggests that the particular setup of the CBTF and precautions taken for CBTF exams are sufficient for successful asynchronous computerized exams. To summarize, the CBTF is a normal computer lab converted for testing purpose, where its file system and network are restricted. The CBTF is proctored while exams are running. Students are not allowed to take any notes in or out of the CBTF. Exam questions are randomly selected and parameterized. We encourage similar strategies to be adopted if any institution wishes to have their own asynchronous computerized environments.

There are a few obvious benefits of asynchronous randomized computerized exams. The computerized format allows questions with more sophisticated formats to be automatically graded, thus allowing class sizes to scale and reducing grading time. The randomized questions reduce exam development time in the long run since items built with randomization can be reused from semester to semester. The asynchronous scheduling virtually eliminates the need for conflict exams and greatly simplifies the handling of exceptions such as student illness. These three benefits facilitate frequent testing for large classes, thus enabling the use of the well-known testing effect at scale (Roediger & Karpicke, 2006; Phelps, 2012) and potentially leading to improved student learning outcomes (Nip et al., 2018).

6 Limitations

As the study in this paper used existing data from real courses, it has a number of limitations as discussed below.

First, we did not have access to demographic information corresponding to individual student records, so we were unable to test for demographic correlates with student behavior or results. In future work, it will be important to understand whether there are specific subgroups of students with different outcomes in the asynchronous computerized exam system. It is easy to imagine that there could be multiple complex and interacting effects, such as asynchronous exams benefiting nontraditional students with family or work obligations, or asynchronous exams disadvantaging nontraditional students without well-established study habits. If we had per-student demographic information, we could examine this by repeating our analysis disaggregated by subgroup.

Second, it is unclear the extent to which our results generalize beyond the setting of engineering courses at a large R1 university in the US. The student body in this study is likely only representative for similarly situated universities and for similar engineering and perhaps STEM courses. As compared to national numbers for the engineering student population, the population in the university where the data collected has slightly fewer females (19.5% / 21.3%), fewer White (41.5% / 62.3%), Black (1.8% / 4.1%), and Hispanic (5.8% / 11.1%) students, but more American Asian (22.1% / 14.6%) and foreign (25.6% / 3.8%) students (Yoder, 2017). It would be very interesting to compare data from other environments.

Third, our data was limited in the number of synchronous exam controls we had for each course. Most of the courses for which we had data only offered a single synchronous exam during the semester. While we have no reason to believe that this would cause a consistent bias in using the synchronous exam as an estimate of student ability in the course, it is likely that the synchronous exam is not measuring exactly the same skills as the asynchronous exams. This would tend to lead to our analysis underestimating the extent to which student score declines over the exam period are due to student ability correlation with exam time selection. Access to other measures of student ability would offer more control that could improve our analysis.

Fourth, we did not have detailed information about the exams in each course. This means that we were unable to investigate the effect of different question types (e.g., multiple choice versus writing code), different exam purposes (e.g., primarily low-stakes formative feedback versus high-stakes summative assessments), or the amount of variation in questions given to different students. Future work with access to per-exam and per-question details could help to elucidate the extent to which these and other factors alter the effects analyzed in this work.

7 Conclusion

We examined 26,139 asynchronous exam records from 81 asynchronous exams and 5,534 exam records from 15 synchronous exams, all gathered over 4 semesters from 6 engineering and computer science courses. We tested the hypothesis that the observed score decline, where students' average performance drops over the exam period in asynchronous exams, can be attributed to weaker students electing to take exams later in the exam period. We found that students' choices of exam time negatively correlate with their scores on synchronous exams, meaning that students with lower scores on synchronous exams tend to choose to take asynchronous exams later. Furthermore, we found that students' performance in synchronous exams explains about 70% of the observed decline in student scores over the exam period, where the observed decline is characterized by β in the uncontrolled regression (Equation 1). This observation indicates that the majority of the observed decline in student scores over the exam period is due to students with different levels of ability choosing to schedule their exams earlier or later in the exam period.

Appendix A: Characteristics of exam score distributions

We will show that the exam score distributions of asynchronous exams do not differ much from that of synchronous exams after the filtering. This is important as discrepancies between the two might indicate something unusual is going on in asynchronous exams. We use *skewness* and *kurtosis* to describe the distributions in addition to the mean of the scores, as in the literature (Lord, 1955; Cook, 1959; Ho & Yu, 2015). The skewness is a measure of the asymmetry of a distribution where negative skewness means longer left tail and positive skewness means longer right tail. The kurtosis is a measure of "tailedness" of a distribution where larger kurtosis indicates there are more values at the tail, and normal distributions have kurtosis 3. The squared skewness plus one is a lower bound of kurtosis, since knowing a distribution is skewed already puts some constraints on the tailedness of the distribution. We plotted the overall distribution of kurtosis versus skewness of asynchronous exams in the top left subplot of Figure 10. To give a concrete sense of what score distributions lead to the different skewness/kurtosis, we plotted score distributions of three selected asynchronous exams around the scatter plot.

We plotted the mean, skewness, and kurtosis of asynchronous exams and synchronous exams side by side in Figure 11 to compare them. As the figure shows, there is no distinctive difference between the two types of exams, and there are two noticeable trends in the distributions: (1) exams with mean above 50% tend to have negative skewness, and (2) exams with near symmetric distributions tend to have negative excess kurtosis as compared to normal distribution, which has kurtosis 3. These two trends for exams have been observed since the middle of the last century (Lord, 1955; Cook, 1959; Ho & Yu, 2015) and indicate that these exams have typical score distributions.

To help understand what kind of asynchronous exams are excluded, we plotted the exam score distributions of two excluded asynchronous exams due to large kurtosis in Figure 12. The main feature of these excluded asynchronous exams are the presence of exam scores that are far away from the mean in terms of number of standard deviations.

Appendix B: Meta-analysis

Meta-analysis techniques essentially deal with cases where it is not desirable to directly average effect sizes to compute the mean. Specifically, the task we are trying to solve is: given a set of k observed effect sizes $T_1, \ldots, T_k$, each with their unknown true effect sizes $\theta_1, \ldots, \theta_k$, find an estimate of θ which is the average of all the θ_is. As an example, consider the effect size to be the length of manufactured rods, and the T_is to be k measurements of the length of k different rods. In the simplest case, we assume that all rods have exactly the same length ($\theta_1 = \ldots = \theta_k = \theta$), and we use a single ruler to measure their lengths so that the variance σ^2 of each measurement is the same. In this case, computing an estimate of θ and its variance is straightforward with

$$\widehat{\theta} = \overline{T} = \frac{\sum_{i=1}^{k} T_i}{k} \tag{4}$$

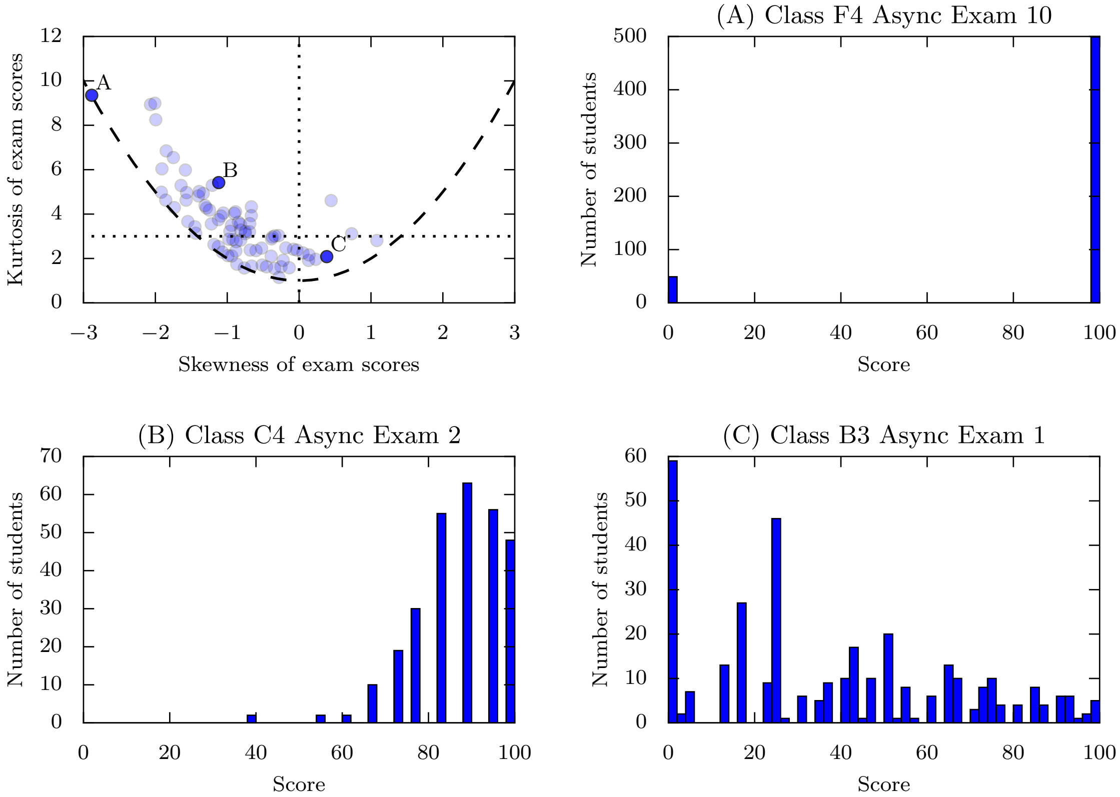

Figure 10: The subplot on the top left is the distribution of kurtosis versus skewness of all the asynchronous exams after the filtering. Each data point in the scatter plot represents a single asynchronous exam. The dashed curve is the lower bound for kurtosis in terms of skewness. The other three subplots are the exam score distributions of the highlighted asynchronous exams in the skewness-kurtosis subplot.

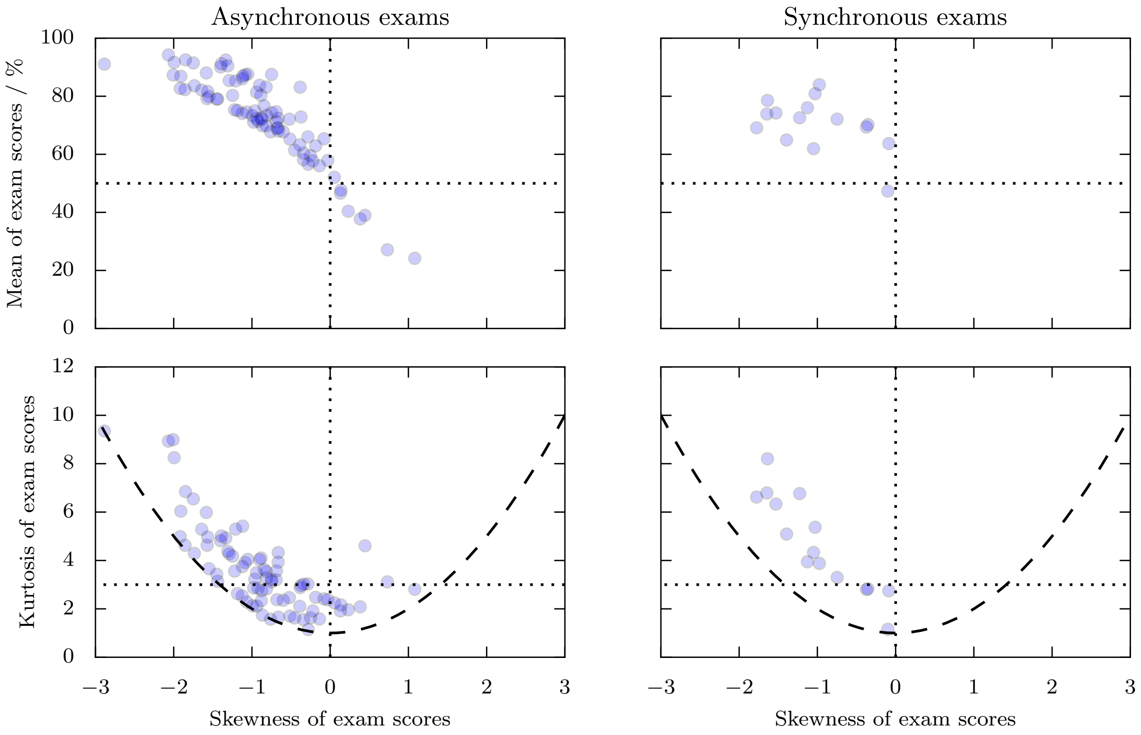

Figure 11: Summary statistics for both the asynchronous exams and the synchronous exams. Each data point represents one exam. The dashed curves in the bottom plots are the lower bound for kurtosis in terms of skewness.

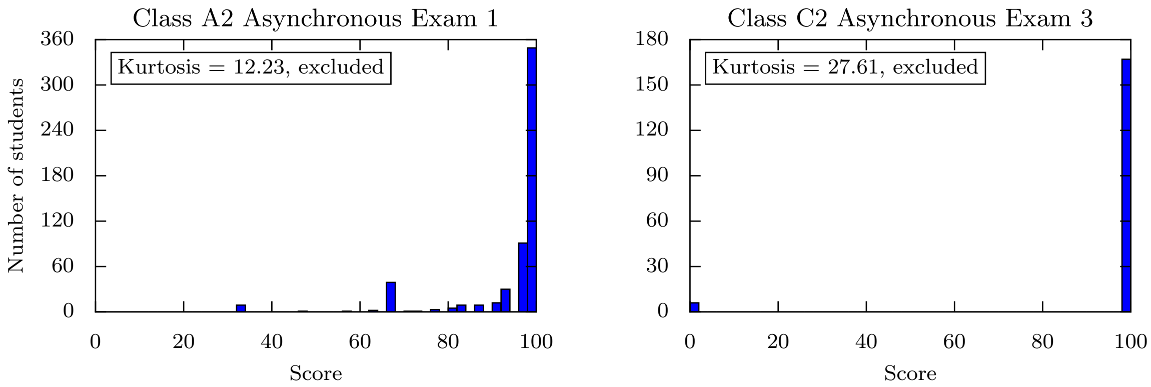

Figure 12: Examples of score distributions of excluded asynchronous exams whose kurtosis is larger than 10. These asynchronous exams generally have scores that are far away from the mean in terms of standard deviation.

and

$$v = \frac{\sigma^2}{k}. \tag{5}$$

Now consider the case when the k measurements are performed with different rulers that have different standard errors $\sigma_1, \ldots, \sigma_k$. We can no longer directly average the T_is to obtain an estimate of θ because they are not equally valuable. In this case, we weight each T_i with $w_i = 1/\sigma_i^2$, so low variance measurements have higher weights (Cooper et al., 2009, Chapter 14). We then compute the estimate of θ and its variance as

$$\widehat{\theta} = \overline{T}_{\cdot} = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i} \tag{6}$$

and

$$v_{\cdot} = \frac{1}{\sum_{i=1}^{k} w_i}. \tag{7}$$

This case is often referred as a fixed effect model characterized by

$$T_i = \theta + e_i, \tag{8}$$

where $\mathrm{var}(e_i) = \sigma_i^2$ is the variance of the ith effect size due to estimation error. The fixed effect model assumes that e_i is normally distributed with mean 0 and variance σ_i^2 ($e_i \sim N(0, \sigma_i^2)$), and the covariance is 0 between e_i and e_j for all $i \neq j$ ($\mathrm{Cov}(e_i, e_j) = 0$ for $i \neq j$).

So far we have assumed that all rods are identical in length in which case the fixed effect model is sufficient. But what if the actual lengths $\theta_1, \ldots, \theta_k$ are normally distributed around the θ that we want to estimate? To deal with this, we need to first determine if there is enough evidence to invalidate the hypothesis that they are created equal. We compute the following homogeneity test statistic for this purpose (Cooper et al., 2009, Chapter 14):

$$Q = \sum_{i=1}^{k} \left[\left(T_i - \overline{T}_{\cdot} \right)^2 / \sigma_i^2 \right] = \sum_{i=1}^{k} w_i \left(T_i - \overline{T}_{\cdot} \right)^2. \tag{9}$$

If Q exceeds the upper-tail critical value of chi-square at $k-1$ degrees of freedom, the observed variance in effect sizes is significantly greater than what we would expect by chance if all studies shared a common population effect size and therefore we reject the null hypothesis (Cooper et al., 2009, Chapter 14). Otherwise there is not enough evidence to reject the null hypothesis and the fixed effect model is the correct choice. In the case of rejection, we need to use a random effect model characterized by

$$T_i = \theta + u_i + e_i, \tag{10}$$

where $\mathrm{var}(u_i) = \sigma_\theta^2$ is the variance of the effect size due to heterogeneity and $\mathrm{var}(e_i) = \sigma_i^2$ is the variance of the ith effect size due to estimation error. The random effect model assumes that $e_i \sim N(0, \sigma_i^2)$, $u_i \sim N(0, \sigma_\theta^2)$ and $\mathrm{Cov}(e_i, e_j) = 0$ for $i \neq j$, $\mathrm{Cov}(e_i, u_j) = 0$ for all i and j. In the case of our example, the existence of u_i could be due to the fact that these rods are not created equal. We need to take σ_θ^2 into account while computing an estimate of θ, therefore we need to estimate σ_θ^2. There are several estimators for this purpose, including the Hedges estimator

$$\widehat{\sigma}_{\theta,H}^2 = \frac{1}{k-1} \sum_{i=1}^{k} \left(T_i - \overline{T} \right)^2 - \frac{1}{k} \sum_{i=1}^{k} \sigma_i^2 \tag{11}$$

and the DerSimonian–Laird estimator

$$\widehat{\sigma}^2_{\theta,DSL} = \frac{Q-(k-1)}{\sum_{i=1}^{k} w_i - \sum_{i=1}^{k} w_i^2 / \sum_{i=1}^{k} w_i}, \tag{12}$$

where Q is as defined in Equation 9. The specific conditions under which each estimator should be used are (Cooper et al., 2009, Chapter 14)

$$\widehat{\sigma}^2_{\theta} = \begin{cases} \max(0, \widehat{\sigma}^2_{\theta,DSL}) & \text{if homogeneity test is not rejected} \\ \widehat{\sigma}^2_{\theta,DSL} & \text{if } \big(\text{homogeneity test is rejected, } \widehat{\sigma}^2_{\theta,H} < 0,\ \widehat{\sigma}^2_{\theta,DSL} > 0\big) \\ & \text{or } \big(\text{homogeneity test is rejected, } \widehat{\sigma}^2_{\theta,H} > 0,\ \widehat{\sigma}^2_{\theta,DSL} > 0,\ Q \leq 3(k-1)\big) \\ \widehat{\sigma}^2_{\theta,H} & \text{otherwise.} \end{cases}$$

After the estimation of σ^2_θ, we then compute weights using $w_i = 1/\left(\sigma^2_\theta + \sigma^2_i\right)$ and use the new w_is in Equations 6 and 7.

So far we have discussed the fixed effect model and the random effect model using the example of measuring rods, but how does this relate to the analysis in the paper? In the paper, we assumed that there exists some value corresponding θ which describes how exam scores would behave on average with respect to time in asynchronous exams. Since we don't know θ, we need to estimate it from the data. We assumed that each asynchronous exam has some slope θ_k that is normally distributed around θ but cannot be observed directly. We thus need to obtain estimates of the θ_ks, which we do so by first having a group of students take the exam asynchronously and then performing OLS regression of scores with respect to time. From the OLS regression, we obtain estimates of the θ_ks as T_ks with standard errors σ_k. The OLS regression can be seen as a ruler, since we used it to obtain T_k with its corresponding error σ_k. This data is exactly what we need to estimate θ.

Appendix C: *

References

Aremu, A. O., Williams, T. M., & Adesina, F. T. (2011). Influence of academic procrastination and personality types on academic achievement and efficacy of in-school adolescents in Ibadan. *IFE PsychologIA*, *19*(1), 93–113. doi: 10.4314/ifep.v19i1.64591

Attali, Y. (2018). Automatic item generation unleashed: An evaluation of a large-scale deployment of item models. In *International conference on artificial intelligence in education* (pp. 17–29). doi: 10.1007/978-3-319-93843-1_2

Balkis, M., Duru, E., & Bulus, M. (2013, Sep 01). Analysis of the relation between academic procrastination, academic rational/irrational beliefs, time preferences to study for exams, and academic achievement: a structural model. *European Journal of Psychology of Education*, *28*(3), 825–839. doi: 10.1007/s10212-012-0142-5

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1991). Effects of frequent classroom testing. *The Journal of Educational Research*, *85*(2), 89–99. doi: 10.1080/00220671.1991.10702818

Barros, J. P., Estevens, L., Dias, R., Pais, R., & Soeiro, E. (2003). Using lab exams to ensure programming practice in an introductory programming course. *ACM SIGCSE Bulletin*, *35*(3), 16–20. doi: 10.1145/961290.961519

Becker, B. J., & Wu, M.-J. (2007). The synthesis of regression slopes in meta-analysis. *Statistical Science*, 414–429. doi: 10.1214/07-STS243

Bennedsen, J., & Caspersen, M. E. (2006). Assessing process and product-a practical lab exam for an introductory programming course. In *Frontiers in education conference* (pp. 16–21). doi: 10.1109/FIE.2006.322434

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Research*, *31*(1), 51-60. doi: 10.2190/GRQQ-YT0F-7LKB-F033

Boevé, A. J., Meijer, R. R., Albers, C. J., Beetsma, Y., & Bosker, R. J. (2015, 12). Introducing computer-based testing in high-stakes exams in higher education: Results of a field experiment. *PLOS ONE*, *10*(12), 1-13. doi: 10.1371/journal.pone.0143616

Bugbee, A. C., Jr. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education*, *28*(3), 282–299. doi: 10.1080/08886504.1996.10782166

Bugbee, A. C., Jr., & Bernt, F. M. (1990). Testing by computer: Findings in six years of use 1982-1988. *Journal of Research on Computing in Education*, *23*(1), 87-100. doi: 10.1080/08886504.1990.10781945

Burns, E. R., Garrett, J. E., & Childs, G. V. (2007). A study of student performance on self-scheduled, computer-based examinations in a medical histology course: is later better? *Medical Teacher*, *29*(9-10), 990–992. doi: 10.1080/01421590701477365

Cagiltay, N., & Ozalp-Yaman, S. (2013). How can we get benefits of computer-based testing in engineering education? *Computer Applications in Engineering Education*, *21*(2), 287-293. doi: 10.1002/cae.20470

Califf, M. E., & Goodwin, M. (2002). Testing skills and knowledge: Introducing a laboratory exam in cs1. *ACM SIGCSE Bulletin*, *34*(1), 217–221. doi: 10.1145/563517.563425

Carrasquel, J. (1985). Competency testing in introductory computer science: the mastery examination at carnegie-mellon university. In *Acm sigcse bulletin* (Vol. 17, p. 240). doi: 10.1145/323275.323387

Chen, B., West, M., & Zilles, C. (2017). Do performance trends suggest wide-spread collaborative cheating on asynchronous exams? In *Learning at scale.* doi: 10.1145/3051457.3051465

Cook, D. L. (1959). A replication of lord's study on skewness and kurtosis of observed test-score distributions. *Educational and Psychological Measurement*, *19*(1), 81-87. doi: 10.1177/001316445901900109

Cooper, H. (2016). *Research synthesis and meta-analysis: A step-by-step approach* (Vol. 2). Sage Publications.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis.* Russell Sage Foundation.

DeMara, R. F., Khoshavi, N., Pyle, S. D., Edison, J., Hartshorne, R., Chen, B., & Georgiopoulos, M. (2016, June). Redesigning computer engineering gateway courses using a novel remediation hierarchy. In *2016 asee annual conference & exposition.* New Orleans, Louisiana: ASEE Conferences. doi: 10.18260/p.26063

Engineering Accreditation Commission. (1998, January). *Engineering criteria 2000: Criteria for accrediting programs in engineering in the united states, 2nd edition* (Tech. Rep.). Accreditation Board for Engineering and Technology, Inc.

Essick, R., West, M., Silva, M., Herman, G. L., & Mercier, E. (2016). Scaling-up collaborative learning for large introductory courses using active learning spaces, TA training, and computerized team management. In *Proceedings of the 123rd american society for engineering education annual conference and exposition (asee 2016).* doi: 10.18260/p.27342

Gibbons, M. T. (2009). *Engineering by the numbers* (Tech. Rep.). American Society for Engineering Education.

Henri, M., Johnson, M. D., & Nepal, B. (2017). A review of competency-based learning: Tools, assessments, and recommendations. *Journal of Engineering Education*, *106*(4), 607-638. doi: 10.1002/jee.20180

Ho, A. D., & Yu, C. C. (2015). Descriptive statistics for modern test score distributions. *Educational and Psychological Measurement*, *75*(3), 365-388. doi: 10.1177/0013164414548576

Jacobson, N. (2000). Using on-computer exams to ensure beginning students' programming competency. *ACM SIGCSE Bulletin*, *32*(4), 53–56. doi: 10.1145/369295.369324

Kember, D., Jamieson, Q. W., Pomfret, M., & Wong, E. T. (1995). Learning approaches, study time and academic performance. *Higher Education*, *29*(3), 329–343. doi: 10.1007/BF01384497

Kim, K. R., & Seo, E. H. (2015). The relationship between procrastination and academic performance: A meta-analysis. *Personality and Individual Differences*, *82*, 26–33. doi: 10.1016/j.paid.2015.02.038

Kleinbaum, D. G., Kupper, L. L., & Morgenstern, H. (1982). *Epidemiologic research: Principles and quantitative methods.*

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., & Nizam, A. (1998). *Applied regression analysis and other multivariable methods.* Thomson Brooks/Cole Publishing Co.

Kreiter, C., Peterson, M., Ferguson, K., & Elliott, S. (2003). The effects of testing in shifts on a clinical in-course computerized exam. *Medical Education*, *37*(3), 202–204. doi: 10.1046/j.1365-2923.2003.01435.x

Kulik, C.-L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research*, *60*(2), 265–299. doi: 10.2307/1170612

Kutner, M. H., Nachtsheim, C., & Neter, J. (2004). *Applied linear regression models.* McGraw-Hill/Irwin.

Lee, E., Garg, N., Bygrave, C., Mahar, J., & Mishra, V. (2015, 06). Can university exams be shortened? an alternative to problematic traditional methodological approaches. In *European conference on research methods 2015.* Academic Conferences International Limited.

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology*, *29*(3), 210–212. doi: 10.1207/s15328023top2903_06

Lord, F. M. (1955). A survey of observed test-score distributions with respect to skewness and kurtosis. *Educational and Psychological Measurement*, *15*(4), 383-389. doi: 10.1177/001316445501500406

McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers & Education*, *39*(3), 299 - 312. doi: 10.1016/s0360-1315(02)00032-5

Muldoon, R. (2012). Is it time to ditch the traditional university exam? *Higher Education Research and Development*, *31*(2), 263-265. doi: 10.1080/07294360.2012.680249

Nip, T., Gunter, E., Herman, G. L., Morphew, J., & West, M. (2018). Using a computer-based testing facility to improve student learning in a programming languages and compilers course. In *Proceedings of the acm special interests group on computer science education (sigcse '18).* doi: 10.1145/3159450.3159500

Olds, B. M., Moskal, B. M., & Miller, R. L. (2005). Assessment in engineering education: Evolution, approaches and future collaborations. *Journal of Engineering Education*, *94*(1), 13-25. doi: 10.1002/j.2168-9830.2005.tb00826.x

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, *41*(5), 673–690. doi: 10.1007/s11135-006-9018-6

Pennebaker, J. W., Gosling, S. D., & Ferrell, J. D. (2013). Daily online testing in large classes: Boosting college performance while reducing achievement gaps. *PLOS One*, *8*(11), e79774. doi: 10.1371/journal.pone.0079774

Phelps, R. P. (2012). The effect of testing on student achievement, 1910–2010. *International Journal of Testing*, *12*(1), 21–43. doi: 10.1080/15305058.2011.602920

Plant, E. A., Ericsson, K. A., Hill, L., & Asberg, K. (2005). Why study time does not predict grade point average across college students: Implications of deliberate practice for academic performance. *Contemporary Educational Psychology*, *30*(1), 96–116. doi: 10.1016/j.cedpsych.2004.06.001

Prisacari, A. A., & Danielson, J. (2017). Rethinking testing mode: Should I offer my next chemistry test on paper or computer? *Computers & Education*, *106*, 1 - 12. doi: 10.1016/j.compedu.2016.11.008

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychological Bulletin*, *138*(2), 353. doi: 10.1037/a0026838

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. doi: 10.1111/j.1745-6916.2006.00012.x

Rytkönen, A., & Myyry, L. (2014). Student experiences on taking electronic exams at the university of helsinki. In *World conference on educational multimedia, hypermedia and telecommunications* (pp. 2114–2121).

Sanders, J. W., West, M., & Herman, G. L. (2016). Scaling up project-based learning for a large introductory mechanics course using mobile phone data capture and peer feedback. In *Proceedings of the 123rd american society for engineering education annual conference and exposition (asee 2016).* doi: 10.18260/p.27341

Schouwenburg, H. C. (2004). Procrastination in academic settings: General introduction. In H. C. Schouwenburg, C. H. Lay, T. A. Pychyl, & J. R. Ferrari (Eds.), *Counseling the procrastinator in academic settings* (pp. 3–17). Washington, DC: American Psychological Association. doi: 10.1037/10808-001

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education*, *35*(4), 453–472. doi: 10.1023/A:1003196224280

Seo, E. H. (2011). The relationships among procrastination, flow, and academic achievement. *Social Behavior and Personality: An International Journal*, *39*(2), 209-217. doi: 10.2224/sbp.2011.39.2.209

Shacham, M. (1998). Computer-based exams in undergraduate engineering courses. *Computer Applications in Engineering Education*, *6*(3), 201–209. doi: 10.1002/(sici)1099-0542(1998)6:3<201::aid-cae9>3.0.co;2-h

Shaeiwitz, J. A. (1996, 7). Outcomes assessment in engineering education. *Journal of Engineering Education*, *85*(3), 239–246. doi: 10.1002/j.2168-9830.1996.tb00239.x

Solomon, L. J., & Rothblum, E. D. (1984). Academic procrastination: Frequency and cognitive-behavioral correlates. *Journal of Counseling Psychology*, *31*(4), 503–509. doi: 10.1037/0022-0167.31.4.503

Stanger-Hall, K. F. (2012). Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes. *CBE—Life Sciences Education*, *11*(3), 294–306. doi: 10.1187/cbe.11-11-0100

Steel, P. (2007). The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin*, *133*(1), 65–94. doi: 10.1037/0033-2909.133.1.65

Stehlik, M. J., & Miller, P. L. (1985). *Implementing a mastery examination in computer science* (Tech. Rep. No. CMU-CS-85-175). Carnegie Mellon University.

van Eerde, W. (2003). A meta-analytically derived nomological network of procrastination. *Personality and Individual Differences*, *35*(6), 1401 - 1418. doi: 10.1016/S0191-8869(02)00358-6

Wagner-Menghin, M., Preusche, I., & Schmidts, M. (2013). The effects of reusing written test items: A study using the rasch model. *ISRN Education*, *2013*. doi: 10.1155/2013/585420

Watson, G., & Sottile, J. (2010). Cheating in the digital age: Do students cheat more in online courses? *Online Journal of Distance Learning Administration*, *13*(1).

West, M., Herman, G. L., & Zilles, C. (2015, June). Prairielearn: Mastery-based online problem solving with adaptive scoring and recommendations driven by machine learning. In *2015 annual conference & exposition.* Seattle, Washington: ASEE Conferences. doi: 10.18260/p.24575

West, M., Silva Sohn, M., & Herman, G. L. (2015). Sustainable reform of an introductory mechanics course sequence driven by a community of practice. In *Proceedings of the asme 2015 international mechanical engineering congress & exposition (imece 2015)* (p. IMECE2015-51493). doi: 10.1115/IMECE2015-51493

Yoder, B. L. (2016). *Engineering by the numbers* (Tech. Rep.). American Society for Engineering Education.

Yoder, B. L. (2017). *Engineering by the numbers* (Tech. Rep.). American Society for Engineering Education.

Zandvliet, D., & Farragher, P. (1997). A comparison of computer-administered and written tests. *Journal of Research on Computing in Education*, *29*(4), 423-438. doi: 10.1080/08886504.1997.10782209

Zilles, C., Deloatch, R. T., Bailey, J., Khattar, B. B., Fagen, W., Heeren, C., ... West, M. (2015, June). Computerized testing: A vision and initial experiences. In *2015 asee annual conference & exposition.* Seattle, Washington: ASEE Conferences. doi: 10.18260/p.23726

Zilles, C., West, M., Mussulman, D., & Bretl, T. (2018). Making testing less trying: Lessons learned from operating a computer-based testing facility. In *2018 ieee frontiers in education conference (fie)* (pp. 1–9). doi: 10.1109/fie.2018.8658551