Do Performance Trends Suggest Wide-spread Collaborative Cheating on Asynchronous Exams?

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

ABSTRACT

Using a data set from 29,492 asynchronous exams in an on-campus proctored computer-based testing facility (CBTF), we observed correlations between when a student chooses to take their exam within the exam period and their score on the exam. Somewhat surprisingly, instead of increasing throughout the exam period, which might be indicative of widespread collaborative cheating, we find that exam scores decrease throughout the exam period. While this could be attributed to weaker students putting off exams, this effect holds even when accounting for student ability as measured by a synchronous exam taken during the same semester. This suggests that precautions can be taken by a CBTF to maintain cheating at a low level (e.g., the level of proctored synchronous exams), in spite of the fact that students are taking their exams over a multi-day period.

Author Keywords

asynchronous exams; student performance; cheating; computerized testing.

INTRODUCTION

Exams are a frequently used method in college education to assess students' understanding of course material. However, running exams for a large class (e.g., 200+ students) can be a logistical nightmare [13, 18, 24]. It has been proposed that computerized exams in a face-to-face proctored environment can greatly reduce the overhead of running exams and broaden the kinds of questions that can be automatically graded [24]. Key to the efficient implementation of computer-based exams for large enrollment classes is running them *asynchronously* (e.g., allowing students to choose their exam time within a given time window), because it allows the testing center where the exams take place to be much smaller than the largest class and gracefully tolerates student conflicts [8, 25].

When faculty are invited to use asynchronous computerized exams in their courses, their almost universal first concern is 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 and resulting in increasing exam scores over the exam period. In fact, in a 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 [21].

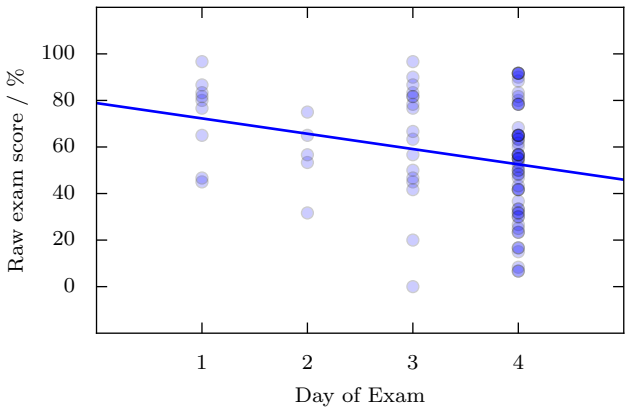

Figure 1. An example data set from one exam (Class D3, Exam 5) that was conducted over a 4 day period. Student raw scores on the exam are plotted against the day on which they took the exam, with each circle being a single student. The straight line is the OLS (ordinary least squares) regression line of the exam score against the day of exam, revealing in this case a large negative association between the day on which the student chose to take the exam and their score. This exam has one of the more negative slopes in our data set, and we chose it here because the highly negative slope is very easy to visualize. For the slopes of all exams, see Figure 6.

However, when we plot students' exam scores versus the day on which they took the exam (see Figure 1 for an example), we see that on average the scores are actually *decreasing* over the exam period. We characterize this effect by the slope β of the regression line in Figure 1 (for this exam, $\beta = -0.82$ standard deviations per exam period). The questions addressed by this paper are: (1) how robust is this negative-slope effect across courses and across semesters, (2) what factors might explain this phenomenon, and (3) what does this data suggest about cheating during in-person asynchronous exams?

L@S 2017, April 20-21, 2017, Cambridge, MA, USA

 ISBN 978-1-4503-4450-0/17/04…$15.00
DOI: http://dx.doi.org/10.1145/3051457.3051465

We report on two investigations into the relationship between when students elect to take their exams and their scores on those exams. In **Analysis 1**, we consider the asynchronous exam records of 9 courses over 3 semesters, demonstrating a trend of declining performance throughout the exam period in most of the 93 exams studied. In **Analysis 2**, we demonstrate that this effect remains even if we control for student ability, using a class that offered both a traditional (synchronous) written exam along with computerized exams in the same semester.

ANALYSIS 1 METHOD

The data collection took place in a large public research university during the Spring 2015, Fall 2015, and Spring 2016 semesters. The data was drawn from asynchronous exams that were held in the Computer-Based Testing Facility (CBTF) [24] and administrated via the PrairieLearn system [22]. All of the courses studied were undergraduate engineering subjects, ranging from introductory to advanced classes in computer science, mechanical engineering, and material science & engineering. With IRB approval and the consent from all of the relevant instructors, the CBTF administrators have shared this collection of data with us for this research.

Computer-Based Testing Facility

The CBTF is hosted in a converted computer lab with 49 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 [24]. During the period studied, the facility was open/proctored 10-12 hours a day, 7 days a week to accommodate one to two thousand exams per week [25]. Students were not permitted to take written notes, photos or other records into or out of the exam room.

Exams within the CBTF are typically administered as follows [25]: Courses typically assign a 3-5 day period for the students to take an exam depending on the class size; longer exam periods are used during finals week. Students are free to reserve any time during this exam period, provided that there are slots available at that time. Generally, the exam periods of exams from different classes overlap each other and the CBTF is almost always running a number of distinct exams concurrently. Sign-ups for exams typically begin 2 weeks before the exam period begins. Exam periods are scheduled so that the CBTF doesn't need to run at more than 85% capacity on any given day, to provide students many potential exam times to choose from and to be able to tolerate any operational problems. At their scheduled exam time, students have their identity checked by a proctor and are randomly assigned to a computer (to deter coordinated cheating).

PrairieLearn

PrairieLearn is an online problem posing system that permits the specification of *problem generators*, each of which is capable of generating a range of parameterized problem instances [22]. For exams, PrairieLearn can be configured to select random problem generators from a pool of questions and randomly generate problem instances from those generators to meet instructor-defined coverage and difficulty criteria [25]. Students sitting next to each other in the CBTF will typically be taking exams from different courses, but even if they are taking the same exam, they will generally have different sets of parameterized questions or the same set of questions with different parameters [25]. PrairieLearn also supports allowing students to have multiple attempts at each question with a partial-credit schedule controlled on a per question basis [25].

For each student taking an exam in the CBTF, PrairieLearn logs all the submissions the student makes during the exam period and calculates and stores the final score based on the instructor's multiple-attempts scoring scheme.

Course and semester	Number of students	Number of exams	Average number of questions per exam
Class A2	180	5	22.6
Class A3	335	4	22.5
Class B2	576	5	3.0
Class B3	271	7	8.0
Class C1	482	2	25.0
Class C2	233	7	7.9
Class C3	453	7	9.7
Class D3	91	7	11.0
Class E3	75	5	14.8
Class F2	593	8	16.9
Class F3	587	9	15.4
Class G2	182	1	10.0
Class G3	250	1	15.0
Class H1	329	3	9.7
Class H2	362	4	11.2
Class H3	196	5	17.2
Class I2	246	7	4.7
Class I3	350	6	4.7

Table 1. Summary information for the 93 exams used in Analysis 1. Each course is indicated by a letter (A–I) and a number for the semester (1 = Spring 2015, 2 = Fall 2015, 3 = Spring 2016). Some courses only started using the CBTF/PrairieLearn environment in later semesters.

ANALYSIS 1 RESULTS

Data overiew

Our data set consists of 29,492 student records from 93 exams in 9 courses over 3 semesters, as listed in Table 1. To obtain this data we took all required exams[1] conducted using the CBTF/PrairieLearn system during these semesters, which yielded 106 exams in total. We then excluded 13 outlier exams with highly unusual score distributions (kurtosis more than 10), for which nearly all students received an identical score

[1]By only including required exams we excluded "second chance" exams that allowed students to optionally replace part or all of an exam score by taking a second equivalent exam at a later date [24]. Such optional exams introduce complex selection biases as they are taken by a non-random subset of students, so we excluded them from the analysis.

(e.g., nearly all students received 100%). All of the courses were undergraduate engineering subjects, ranging from introductory to advanced classes in computer science, mechanical engineering, and material science & engineering. The exam questions ranged from multiple choices, fill in the blank, and numerical calculations to vector drawing, finite state machine design, and coding.

For each of these 93 exams we obtained all student records, which are triples of the form **(day of exam, hour of exam, raw score)**. The day of exam ranged from 1 to the exam period length (variable across exams, generally 3 to 4 days, maximum of 8 days), the hour of exam ranged from 1 to 12, and the raw-score was on a 0% to 100% scale.

For each exam we excluded the student records for any student who completed less than 25% of the exams in the class to avoid including course staff members engaged in exam checking and students who dropped early in the semester. We also excluded 313 student records that occurred outside the official exam periods because the student was sick, on travel, or had some other excuse, and which would otherwise exert an artificially high influence on the score slope estimates.

Score standardization and distributions

We standardized each raw score to a **standardized score** on an exam-by-exam basis. That is, the standardized score is computed by subtracting the exam mean and dividing by the exam standard deviation, so the standardized score measures the number of standard deviations above or below the mean.

To understand the exam score distributions we computed the mean raw score, skewness, and kurtosis for each exam, as shown in Figure 2. These plots show that the score distributions are not normal (which would have zero skewness and a kurtosis of 3), but that they deviate from normal in a structured way. While non-normal distributions are pervasive [16], the particular form of non-normality in our score distributions has been observed since the middle of the last century [14, 5, 12]. As described by Lord [14], two key observed features are: (1) exams with mean above 50% generally have negative skew, and (2) exams with near symmetric distributions (skew near zero) generally have negative excess kurtosis (i.e., they are platykurtic, with kurtosis less than the normal distribution kurtosis of 3, so have light tails [23]). For exams with means well above 50%, we see that they are skewed left (negative skew) and have positive excess kurtosis (kurtosis more than 3, heavy tailed, leptokurtic).

The non-normality of exam scores can be explained by regarding the distribution as a limited (or censored) normal distribution, where scores that would be below 0% or above 100% are limited to these values. Figure 3 shows two example normal probability plots (Q-Q plots versus a normal) for representatives exams. We see in both cases that it is the limiting of scores at 0% and 100% that is causing non-normality.

There are many statistical techniques designed to either normalize non-normal data (e.g., the Box-Cox transform [19] or Item Response Theory (IRT) scoring and normalization [15]) before performing regressions or to perform a regression directly with a model designed for the non-normal data (e.g.,

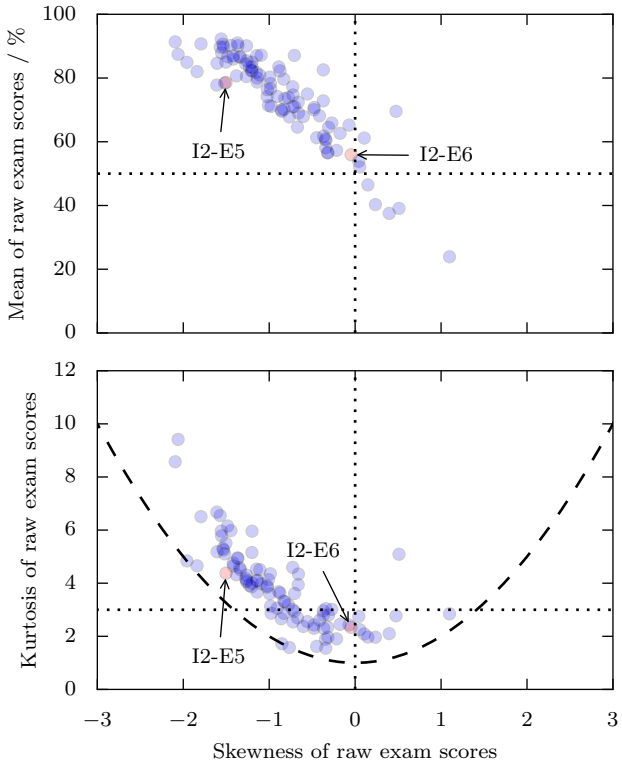

Figure 2. Summary statistics for the exams. Each data point is one exam. The dashed line in the bottom plot is the lower bound for kurtosis in terms of skewness. These plots show that the exam score distributions are non-normal in a way that is consistent with limiting (also called censoring) of the scores at 0% and 100%. Normal probability plots for the two representative labeled exams are shown in Figure 3 and demonstrate this limiting effect.

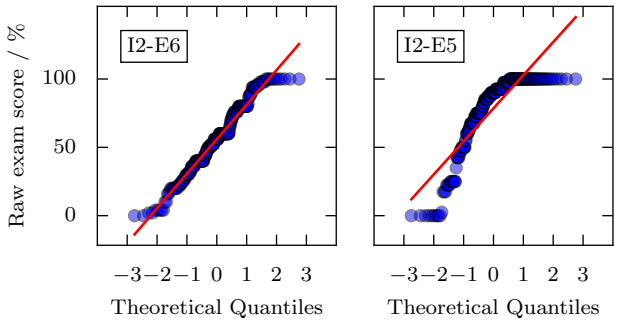

Figure 3. Normal probability plots for two representative exams (see Figure 2 for skew and kurtosis values of these exams). Left: I2-E6 (Class I2, Exam 6) has a nearly symmetric distribution (skew near zero) and negative excess kurtosis (kurtosis less than 3, so lighter-tailed than a normal distribution). Right: I2-E5 (Class I2, Exam 5) has a left-skewed distribution (negative skewness) with positive excess kurtosis (heavy-tailed relative to normal). In both cases we can see that these effects are due to the limiting of the distribution at 0% and 100%.

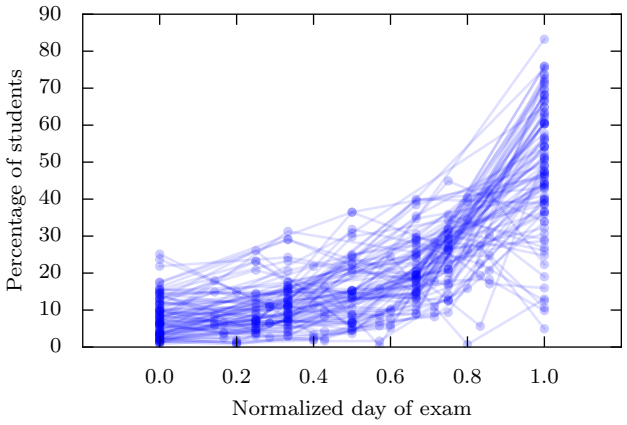

Figure 4. Fraction of students taking the exam on each day. Each exam is a single line on the plot. The horizontal axis shows the normalized day of exam, so 0 is the first day of each exam and 1 is the last day.

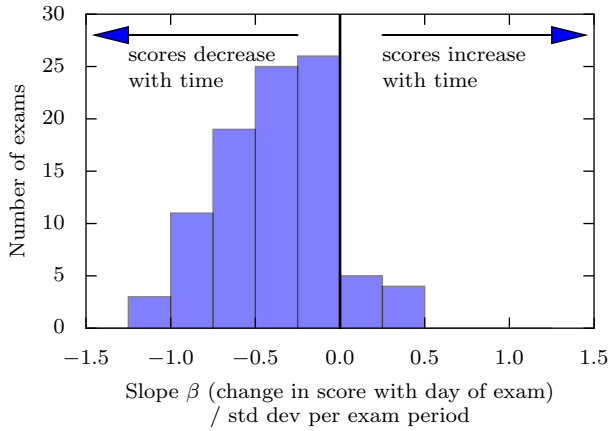

Figure 5. Histogram of exam slopes β_k (change in standardized exam score with normalized day of exam). The units of β are score standard deviations per exam period, so that a value of $\beta = -0.5$ means that scores decline on average by half a standard deviation from the start to the end of the exam.

Tobit models [20]). However, in many situations it may be difficult to avoid introducing other artifacts, such as highly discretized transformed distributions [12].

For this reason and to maintain simplicity of the analysis and clarity of interpretation, we calculate regressions with the standardized exam scores without any extra treatment for non-normality. This will have the effect of systematically underestimating regression slopes [9] because the limited scores have been capped at 0% and 100%, lessening their impact. This means that the effect sizes found in this work are actually underestimates of the true effect.

Day-of-exam student preferences

To compare day-of-exam values between exams of different period lengths, we normalized by dividing the day of exam by the exam period to get the **normalized day of exam** which ranges from 0 on the first day of the exam to 1 on the last day.

Using this scaling we can plot the fraction of students taking the exam on each day for all exams, as shown in Figure 4. Here we see that, when given agency in selecting their exam times, students overwhelmingly prefer to take exams toward the end of the exam period. Each series of connected line segments in Figure 4 represents the distribution of students for a single exam throughout its exam period. The trend in this data is almost exponential, with 40% to 80% of the students taking most exams on the last day of the exam period. While many student motivations could explain this data, two significant hypotheses are: 1) students are delaying their exams so as to have time to gather information from other students that took the exam early in the exam period, and 2) students are self-selecting later exam times so as to give themselves additional preparation time before the exam.

Score as a function of exam day

For each exam we used OLS (ordinary least squares) to fit a regression line of the form

$$z_{ik} = \alpha_k + \beta_k d_{ik}, \qquad (1)$$

where student i took exam k on normalized day of exam d_{ik} (0 to 1 for first day to last day) and received the standardized score z_{ik}. The fitted parameters are the **intercept** α_k and the **slope** β_k for exam k, and we also determine the sampling variance v_k of each slope β_k. Note that the regressions are performed for each exam independently.

The slope β has units of standard deviations per exam period, so a value of $\beta = -0.5$ would mean, roughly speaking, that the student exam scores decline by one half of a standard deviation from the first day to the last day of the exam.

A histogram of the slopes β_k for all exams is shown in Figure 5. The dominant feature is that most exams have negative slopes, meaning that student scores decline over the exam period. While this histogram captures the main effect that slopes are generally negative, it does not allow us to visualize the uncertainty in the slope estimates or to see the relationships between exams for the same class.

Meta-analysis of exam score slopes

To understand the average score trends for asynchronous exams we use the framework of meta-analysis [7] to find the average score slope. We begin by visualizing the slopes β_k together with their 95% confidence intervals on a forest plot in Figure 6. A forest plot is a standard meta-analysis visualization tool [7, 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 circles).

The two-tailed significance levels (p-values) for the slope being non-zero are shown on the right hand side of Figure 6. About half of the exam slopes are statistically significantly negative ($p < 0.05$), a majority of the remainder are non-significantly negative, and a small number are non-significantly positive. None of the exams have a slope that is statistically significantly positive ($p < 0.05$).

There is no clear consensus in the meta-analysis community on how to combine regression slopes in the general case [6].

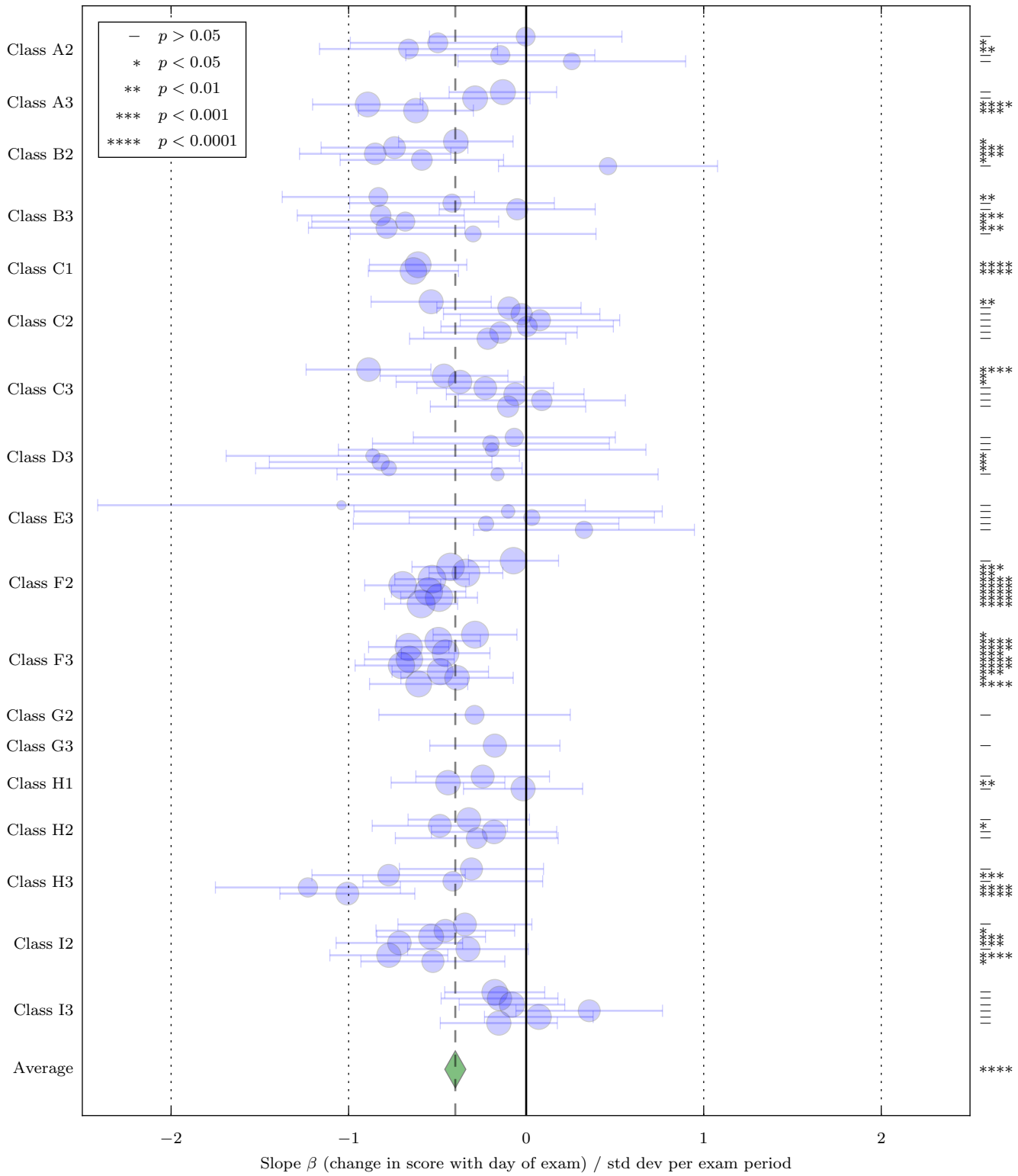

Figure 6. Forest plot shows the slopes β_k of standardized exam score versus normalized day of exam. Each circle represents the slope of one exam and they are grouped by the course and semester as shown on the left. The area of each circle is proportional to the weight $w_k = 1/v_k$ of the exam in the meta-analysis and the horizontal error bar is the 95% confidence interval for the slope. The diamond at the bottom of the figure represents the average population slope $\theta = -0.399$ (95% CI $[-0.458, -0.340]$) for all exams and its width specifies the 95% confidence interval (random-effects model, see Eq. (3)). The two-tailed significance levels of the exam slopes away from zero are shown on the right of the figure as a number of stars.

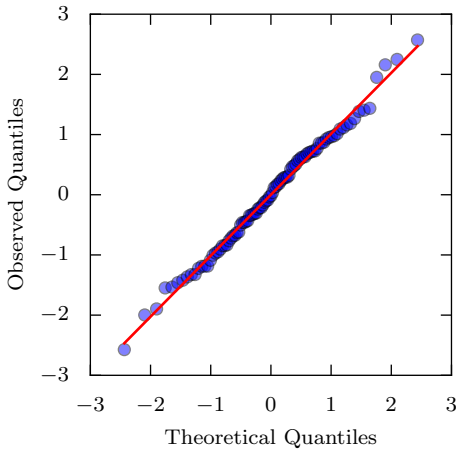

Figure 7. Normal probability plot for the slopes β_k from all exams. This shows that the slopes are approximately normally distributed.

However, both Becker and Wu [2] and Cooper [6] suggested that under the condition when both the response 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 below.

A normal probability plot for the slopes β_k (shown in Figure 7) reveals that they can be regarded as normally distributed. It is thus tempting to use a fixed-effects model of the form

$$\beta_k = \theta + e_k, \tag{2}$$

where θ is the common slope for all exams and $\text{var}(e_k) = v_k$ is the known sampling variance of the kth slope. We find, however, that such a model fails to account for the heterogeneity in the data. Homogeneity is rejected ($p < 0.0001$, $Q = 216.7$) by the standard homogeneity Q-test (Cochran's χ^2 test) [7, Chapter 14] for a common population effect size, and a more advanced measure of heterogeneity [11] finds that $I^2 = 57.5\%$ of the total variation is due to heterogeneity between exams (a medium degree of heterogeneity).

To account for exam heterogeneity we use a random-effects model of the form

$$\beta_k = \theta + u_k + e_k, \tag{3}$$

where θ is now the average population slope, $\text{var}(u_k) = \tau^2$ is the heterogeneity between exams, and $\text{var}(e_k)$ is the known sampling variance of the kth slope. Fitting this model yields an average population slope of $\theta = -0.399$ (95% CI $[-0.458, -0.340]$) which is negative ($p < 0.0001$). This average slope is plotted in Figure 6 as the diamond near the bottom.

ANALYSIS 1 DISCUSSION

While many instructors anecdotally expect that exam scores will rise with day of exam over the exam period either because students have longer time to prepare or because they are colluding with students who have taken the exam at an earlier date, our results show that in fact the opposite occurs and on average the exam scores decline by about 0.4 standard deviations over the period of the exam. In fact, we expect that we have underestimated this effect because we computed per-exam slopes with OLS (ordinary least squares) and ignored the non-normality introduced by the score limits at 0% and 100%.

From Figure 6 there are some indications that the effect size might vary between courses or between semesters. We have not attempted to isolate such variation in the analysis here, but this represents an interesting possibility for future work, perhaps by using three-level (or higher) models [4].

Although the Analysis 1 results do not seem to indicate the existence of widespread collaborative cheating, the possibility exists that stronger students are choosing to take the exam on earlier dates than weaker students. In this case an exam without cheating might actually have a steeper decline (say 1 standard deviation) and collaborative cheating might be assisting the weaker students later in the exam (say by 0.6 standard deviations), resulting in the net decline of 0.4 standard deviations that we observe. We investigate this possibility in Analysis 2 below.

ANALYSIS 2 METHOD

Although the above analysis demonstrated that students' exam scores decrease over time, it does not rule out the possibility that the phenomenon can be explained by good students selecting earlier exam dates while less capable students select later exam dates. To examine the influence of students' ability, we need to calibrate it in some way and take it into account in the regression model. A class that employs both synchronous and asynchronous exams in the same semester can provide such calibration.

In this regard, we exploit a natural experiment resulting from a class that adopted asynchronous computerized exams part way through the semester. In Class C1, a traditional pencil-and-paper synchronous exam (Exam 0) was held before the class changed to hold the remaining two exams (Exams 1 and 2) as asynchronous computerized exams in the CBTF. This exam structure allows us to use Exam 0 as a measure of student ability and investigate how controlling for student ability changes the observed trends of decreasing student performance from the beginning to the end of the exam period on Exams 1 and 2.

Exam Details

The three exams are detailed in Table 2 and the associated score distributions are shown in Figure 8.

Exam 0 was a pencil-and-paper multiple-choice exam in a format that had been traditionally used in Class C, covering material in Homeworks 1–4. All students took the exam at the same time in large classrooms proctored by course staff. The instructor of the course willingly shared the information of the first exam with us for this analysis.

Exams 1 and 2 were computer-based exams administered using PrairieLearn [22] in the CBTF [24], as described above in Analysis 1. Exam 1 was non-comprehensive, covering material from Homeworks 5–7, while Exam 2 was a comprehensive final exam. These exams used a fixed pool of questions for all students, with each student getting different parameterized versions of the same questions, with the question order

	Type	Purpose	Duration	Format	Questions	Mean	Std dev	Skew	Kurtosis
Exam 0	Synchronous	Midterm	2 h	Pencil and paper	20	63.9%	14.0%	−0.083	2.760
Exam 1	Asynchronous	Midterm	2 h	Computerized	20	60.7%	15.1%	−0.260	2.839
Exam 2	Asynchronous	Final	3 h	Computerized	30	73.6%	16.9%	−0.841	3.313

Table 2. Summary information for the three exams used in Analysis 2. Eleven students did not take all three exams during the official exam periods and their records were discarded from the data set. All statistics and analyses use the 469 common students who took all three exams. Exams 1 and 2 are also part of the Analysis 1 data set, where they are the two exams in Class C1. See Figure 8 for score distributions.

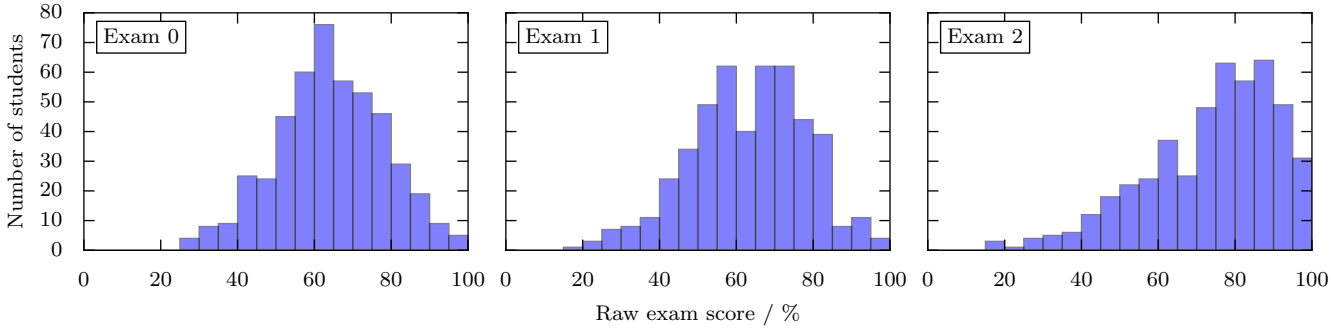

Figure 8. Histograms of exam scores used in Analysis 2. See Table 2 for summary statistics.

randomized. Exam 1 consisted of 10 questions previously assigned as homework and 10 questions that were new for the exam. Exam 2 was drawn exclusively from the pool of 199 PrairieLearn-based homework questions that were assigned throughout the semester. For Exam 1 students were given only one attempt at each question and were graded as correct/incorrect. For Exam 2 students were able to re-attempt questions for partial credit during the exam.

ANALYSIS 2 RESULTS

Slope of exam scores by day of exam

We follow the same procedure as in Analysis 1 and standardize the raw exam scores to standardized scores on an exam-by-exam basis. For Exams 1 and 2, the day of exam is normalized by the exam period to give the normalized day of exam that runs from 0 to 1. The number of students taking the exams on each day is shown in Figure 9, where we see the same general upward trend for Exam 1 as in Figure 4. The drop on the last day was atypical; only 13 out of 93 exams studied in Analysis 1 have this phenomenon. We believe that the behavior observed for this exam is due to the fact that about 40% of the students were enrolled in another class that held an exam immediately after the last day of the exam period. The upward trend was less pronounced for Exam 2 because it was held in the final week of the semester when many students wanted to leave campus early. The drop in the middle is caused by the same co-enrollment class holding a final exam on day 5.

Figure 10 shows the raw exam score distributions by day of exam for Exams 1 and 2 (Exam 0 was held at a single time, so does not have a corresponding plot). The OLS models from Analysis 1 for these standardized exam scores versus the normalized day of exam is shown in Table 4, where the model variables are listed in Table 3. These slopes are shown graphically in the upper half of Figure 11, and an average slope of $\beta_{\text{day}} = -0.607$ (95% CI $[-0.794, -0.419]$) was computed using a fixed-effects model of the form shown in Equation (2).

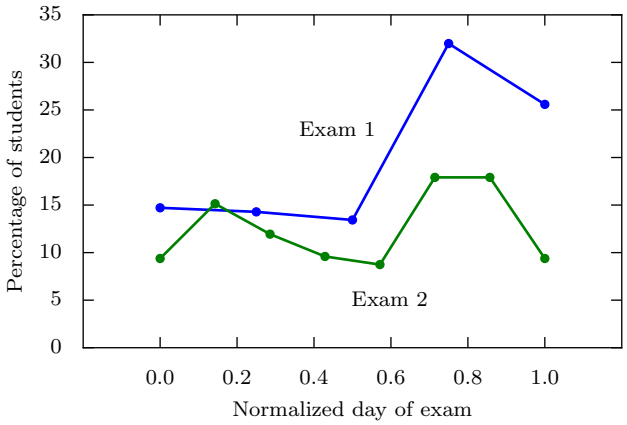

Figure 9. Number of students per exam day for Analysis 2.

Exam scores controlling for Exam 0 score

To investigate the hypothesis that the negative slopes observed in Analysis 1 are due to higher-ability students preferentially taking the exam early, we take Exam 0 scores as a proxy for ability and use it as a control when regressing Exam 1 and 2 scores against the day of exam.

We first checked the correlation of Exam 0 scores with the scores and days of Exams 1 and 2, as shown in Table 5. As expected, Exam 0 scores are significantly positively correlated with both Exam 1 scores and Exam 2 scores. Students who performed well on Exam 0 are also likely to have taken Exam 1 earlier in the exam period (statistically significant negative correlation), but there is no statistically significant relationship between Exam 0 scores and the day of Exam 2. This is consistent with stronger students choosing to take Exam 1 somewhat earlier, but there being no clear preference for scheduling times for Exam 2 which was held at the end of semester.

The lower two models in Table 4 show the regression of the Exam 1 and 2 scores against the day of exam and Exam 0

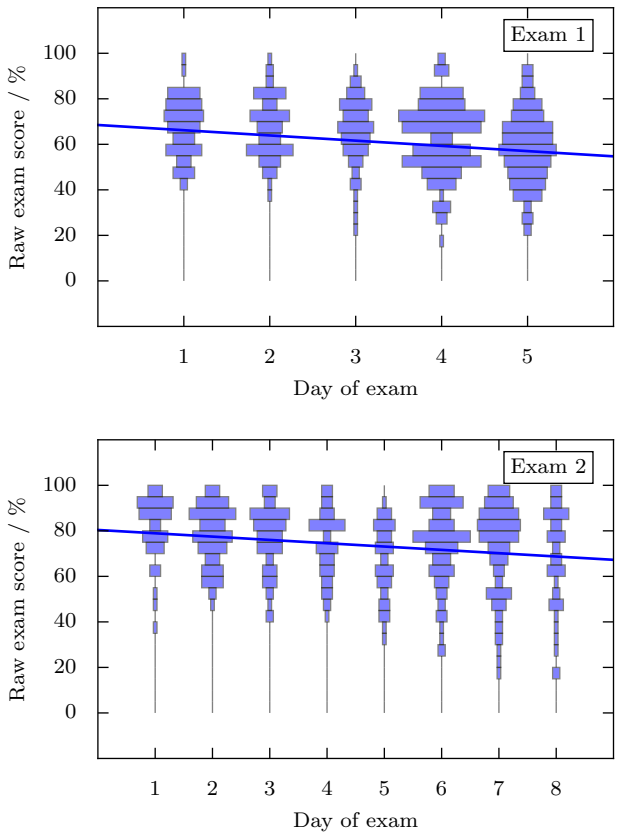

Figure 10. Raw scores versus day of exam for the two asynchronous exams in Analysis 2. The trend line is the OLS regression, as in Analysis 1, and in both cases it shows a negative slope.

Variable	Description
score0	Standardized Exam 0 score
score1	Standardized Exam 1 score
score2	Standardized Exam 2 score
day1	Normalized day of exam for Exam 1
day2	Normalized day of exam for Exam 2

Table 3. Variables used in Tables 4 and 5.

Dep var	Indep	Coef	95% CI		R^2
score1	const	0.364	0.187	0.542	0.045
	day1	−0.609	−0.865	−0.352	
score2	const	0.315	0.145	0.485	0.038
	day2	−0.604	−0.881	−0.327	
score1	const	0.257	0.105	0.409	0.303
	score0	0.513	0.436	0.589	
	day1	−0.429	−0.650	−0.209	
score2	const	0.306	0.157	0.455	0.266
	score0	0.478	0.399	0.556	
	day2	−0.586	−0.829	−0.343	

Table 4. Linear models for Exam 1 and Exam 2 standardized scores in terms of the normalized day of exam (top two model rows), and in terms of the normalized day of exam and the standardized Exam 0 score (bottom two rows). See Table 3 for variable descriptions. Figure 11 shows a graphical representation of the dependence of each model on the day of exam variables.

Correlation	r	p	95% CI	
score0 with score1	0.531	0.000	0.462	0.593
score0 with score2	0.480	0.000	0.407	0.547
score0 with day1	−0.121	0.009	−0.210	−0.031
score0 with day2	−0.012	0.793	−0.103	0.078

Table 5. Correlation coefficients between student scores on Exam 0 and both the scores and the day of exam for Exams 1 and 2. The scores are positively correlated, and there is a weak negative correlation between Exam 0 score and the day of exam for Exam 1. See Table 3 for variable definitions.

scores, where we are regarding the Exam 0 scores as a proxy for student ability. Because the independent variables are correlated, we must be careful of multicollinearity in performing these regressions. We computed the variance inflation factor [1, Chapter 13] and found VIF < 1.02 for all independent variables, indicating that multicollinearity is very low and not a concern. The regression coefficients of Exam 1 and 2 scores with day of exam in the models with Exam 0 as a control are plotted in the lower half of Figure 11, and the average fixed-effects slope is $\beta_{\text{day,Exam0}} = -0.500$ (95% CI $[-0.663, -0.337]$).

From Figure 11 we can also see that controlling for student ability via the Exam 0 score reduced the magnitude of the average slope by about 18%, from -0.607 to -0.500 standard deviations per day (not statistically significant, $p = 0.402$).

ANALYSIS 2 DISCUSSION

The magnitude of the negative exam score slope over the exam period is only mildly reduced (about 20%), and not statistically significantly, when controlling for student ability as measured by the synchronous Exam 0 scores. This suggests that the negative slope in scores is not primarily due to stronger students taking the exam earlier, although this does happen to a small extent.

Because Exam 0 is only a proxy for student ability and only moderately correlated with Exam 1 and 2 scores, we cannot rule out the possibility that early exam taking by stronger students is in fact responsible for a larger (or even entire) proportion of the declining-scores effect. One potential approach to more definitively resolve this question might be to use student performance data from other courses as a control.

CONCLUSIONS AND FUTURE WORK

In this paper we examined approximately 30,000 asynchronous exam records from 93 exams in 9 courses over 3 semesters to test the hypothesis that collaborative cheating would inflate student scores in asynchronous exams held in a face-to-face, proctored testing center. We did not find any significant evidence that collaborative cheating was inflating student scores later in the exam period. In fact, we found that student scores decreased substantially over the course of the exam period (by about 0.4 standard deviations), even when controlling for student ability as measured by a synchronous exam scores. In the engineering and computer science courses that formed our data set, 0.4 standard deviations typically corresponds to about half of a letter grade, making this a sizable

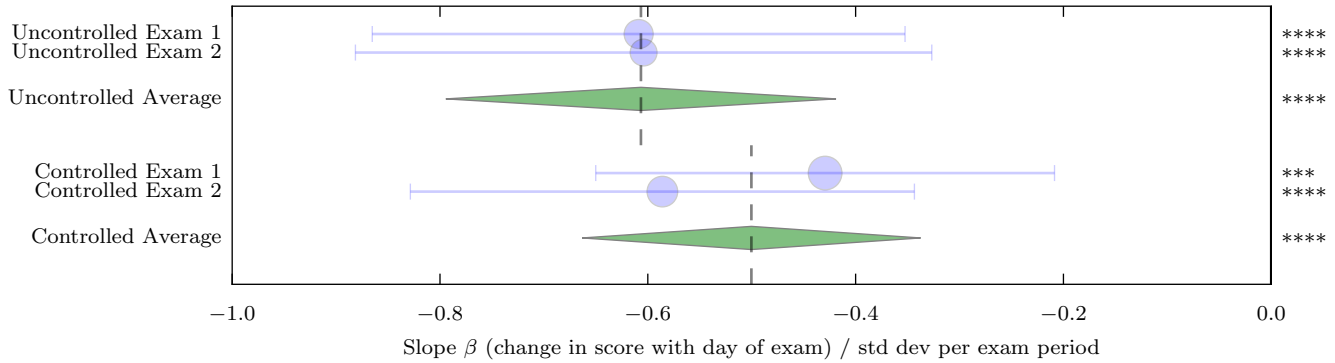

Figure 11. Forest plot for the regressions against day of exam in Analysis 2, as listed in Table 4, with error bars indicating 95% confidence intervals. The "Uncontrolled" regressions fit Exam 1 and 2 scores as functions of the day of exam, while the "Controlled" regressions fit as functions of day of exam and Exam 0 scores, which are a proxy for student ability. The average slope diamonds are the fixed-effects average of the corresponding two exam slopes, and their widths are the 95% confidence interval. We see that controlling for ability reduces the average slope magnitude from -0.607 **(95% CI** $[-0.794, -0.419]$**) to** -0.500 **(95% CI** $[-0.663, -0.337]$**) (not statistically significant).**

effect. This decline in scores over the exam period seems to indicate that student choice of exam date is revealing additional information about their preparedness or ability level, above and beyond that measured by their scores on the traditional synchronous exam.

In addition, our initial investigations suggest that there also is a time-of-day effect, with similar structure to the day-of-exam effect that is the focus of this paper. Students were able to choose what time they take their exams, and they predominantly preferred exam times later in the day, especially near the end of an exam period. Furthermore, exam scores are negatively associated with the time of day of the exam, although the effect is considerably weaker than the day-of-exam effect. The consistency of these results support the hypothesis that students are signaling their lack of preparation for an exam by selecting later slots in the exam schedule. This observation opens up the possibility that interventions could be targeted to these students.

Our results also support an alternative hypothesis for the study reported by Brothen and Peterson [3]. Their work reports a natural experiment that occurred during a proctored asynchronous computerized final exam where students could elect the day on which to take their exam. Computer problems interrupted the exam in the middle of the first exam period, forcing the faculty to provide the first cohort the opportunity to take the exam online and unproctored later in the week, as some students' travel arrangements prevented them from completing a proctored make-up later in the exam period. When this cohort out-performed the rest of the class by 0.63 standard deviations, the authors found that cheating was the most likely explanation. Our data suggests that a non-trivial portion of this difference can be explained because the experimental cohort was the one that *chose* to sign up for the first time slot. That cheating may not have been the dominant effect is also supported by the partial results from the aborted exams, which actually projected higher average scores than the cohort achieved on their unproctored re-take.

While our findings suggest that collaborative cheating is not the dominant effect in these asynchronous exams, we certainly cannot conclude that no collaborative cheating is occurring from our results. Nevertheless, it gives some confidence that the precautions taken by the CBTF (e.g., proctoring, question randomization, preventing notes from being brought in or removed from the CBTF) are sufficient to prevent widespread collaborative cheating. Therefore in-person asynchronous environments are encouraged to adopt similar strategies to prevent cheating. Our results, however, offer no direct insight into cheating in the context of unproctored, online exams [10, 17, 21]. Regardless of this, it would be interesting to repeat the analysis with online testing data both for remote-proctored and unproctored online exams, to see whether similar effects are seen in those environments and whether there is evidence of cheating.

The trend of decreasing performance throughout the exam period is well supported by the data, but there is a lot of variation between classes and exams that remains unexplained. Our current hypothesis is that this variation is derived from variations in exam construction (e.g., exam length, exam difficulty, drawing questions from a pool versus using a fixed set of problems for all students). In particular, we believe it is important to understand the degree to which any of these characteristics contribute to deterring collaborative cheating.

Finally, our analysis largely treats each student taking each exam as independent events. We believe that future work that exploits the structure in the data could more clearly eludicate these effects. A few extensions are obvious. First, because our anonymized data set retains the association for all exams taken by a given student, we can study if individual students perform better when they choose to take exams early in the exam period. Second, we can use multi-level models to recognize that, for example, Class H1 Exam 1 is basically the same exam as Class H2 Exam 1 that is taken by a different student cohort. Lastly, we can explore the degree to which our results are being unduly affected by a certain portion of the student distribution, for example the very weakest students choosing to postpone the exam as long as possible.

ACKNOWLEDGEMENTS

This work was partially supported by NSF DUE-1347722, NSF CMMI-1150490, and the College of Engineering at the University of Illinois at Urbana-Champaign under the Strategic Instructional Initiatives Program (SIIP).

REFERENCES

1. R. Azen and D. Budescu (Eds.). 2009. *Applications of Multiple Regression in Psychological Research*. SAGE Publications. DOI: http://dx.doi.org/10.4135/9780857020994
2. B. J. Becker and M.-J. Wu. 2007. The synthesis of regression slopes in meta-analysis. *Statist. Sci.* (2007), 414–429.
3. T. Brothen and G. Peterson. 2012. Online exam cheating: A natural experiment. *International Journal of Instructional Technology and Distance Learning* 9, 2 (2012), 15–20.
4. M. W.-L. Cheung. 2014. Modeling dependent effect sizes with three-level meta-analyses: a structural equation modeling approach. *Psychol. Methods* 19, 2 (2014), 211.
5. D. L. Cook. 1959. A replication of Lord's study of skewness and kurtosis of observed test-score distributions. *Educational and Psychological Measurement* 19, 1 (1959), 81–87. DOI: http://dx.doi.org/10.1177/001316445901900109
6. H. Cooper. 2016. *Research synthesis and meta-analysis: A step-by-step approach*. Vol. 2. SAGE publications.
7. H. Cooper, L. V. Hedges, and J. C. Valentine. 2009. *The handbook of research synthesis and meta-analysis* (2nd ed.). Russell Sage Foundation.
8. R. F. DeMara, N. Khoshavi, S. Pyle, J. Edison, R. Hartshorne, B. Chen, and M. Georgiopoulos. 2016. Redesigning Computer Engineering Gateway Courses Using a Novel Remediation Hierarchy. In *2016 ASEE Annual Conference & Exposition*. ASEE Conferences, New Orleans, Louisiana.
9. W. H. Greene. 1981. On the asymptotic bias of the ordinary least squares estimator of the Tobit model. *Econometrica* 49, 2 (1981), 505–513. DOI: http://dx.doi.org/10.2307/1913323
10. O. R. Harmon and J. Lambrinos. 2008. Are online exams an invitation to cheat? *The Journal of Economic Education* 39, 2 (2008), 116–125.
11. J. P. T. Higgins and S. G. Thompson. 2002. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 21 (2002), 1539–1558. DOI: http://dx.doi.org/10.1002/sim.1186
12. A. D. Ho and C. C. Yu. 2015. Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement* 75, 3 (2015), 365–388. DOI: http://dx.doi.org/10.1177/0013164414548576
13. E. Lee, N. Garg, C. Bygrave, J. Mahar, and V. Mishra. 2015. Can University Exams be Shortened? An Alternative to Problematic Traditional Methodological Approaches. In *Proceedings of the 14th European Conference on Research Methods*. Valletta, Malta.
14. F. M. Lord. 1955. A survey of observed test-score distributions with respect to skewness and kurtosis. *Educational and Psychological Measurement* 15, 4 (1955), 383–389. DOI: http://dx.doi.org/10.1177/001316445501500406
15. F. M. Lord. 1980. *Applications of item response theory to practical testing problems*. Erlbaum.
16. T. Micceri. 1989. The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin* 105, 1 (1989), 156–166. DOI: http://dx.doi.org/10.1037/0033-2909.105.1.156
17. A. Miller and A. D. Young-Jones. 2012. Academic integrity: Online classes compared to face-to-face classes. *Journal of Instructional Psychology* 39, 3/4 (2012), 138.
18. R. Muldoon. 2012. Is it time to ditch the traditional university exam? *Higher Education Research & Development* 31, 2 (2012), 263–265. DOI: http://dx.doi.org/10.1080/07294360.2012.680249
19. R. M. Sakia. 1992. The Box-Cox Transformation Technique: A Review. *Journal of the Royal Statistical Society. Series D (The Statistician)* 41, 2 (1992), 169–178. DOI:http://dx.doi.org/10.2307/2348250
20. W. Schnedler. 2005. Likelihood Estimation for Censored Random Vectors. *Econometric Reviews* 24, 2 (2005), 195–217. DOI:http://dx.doi.org/10.1081/ETC-200067925
21. G. Watson and J. Sottile. 2010. Cheating in the Digital Age: Do Students Cheat More in Online Courses. *Online Journal of Distance Learning Administration* 13, 1 (2010). http://www.westga.edu/~distance/ojdla/spring131/watson131.html
22. M. West, G. L. Herman, and C. Zilles. 2015. PrairieLearn: Mastery-based Online Problem Solving with Adaptive Scoring and Recommendations Driven by Machine Learning. In *Proceedings of the 2015 ASEE Annual Conference and Exposition (ASEE 2015)*. 26.1238.1–26.1238.14. DOI: http://dx.doi.org/10.18260/p.24575
23. P. H. Westfall. 2014. Kurtosis as Peakedness, 1905–2014. *R.I.P. American Statistician* 68, 3 (2014), 191–195. DOI: http://dx.doi.org/10.1080/00031305.2014.917055
24. C. Zilles, R. T. Deloatch, J. Bailey, B. B. Khattar, W. Fagen, C. Heeren, D. Mussulman, and M. West. 2015. Computerized Testing: A Vision and Initial Experiences. In *Proceedings of the 2015 ASEE Annual Conference and Exposition (ASEE 2015)*. 26.387.1–26.387.13. DOI: http://dx.doi.org/10.18260/p.23726
25. C. Zilles, M. West, and D. Mussulman. 2016. Student Behavior in Selecting an Exam Time in a Computer-Based Testing Facility. In *2016 ASEE Annual Conference & Exposition*. ASEE Conferences, New Orleans, Louisiana. DOI: http://dx.doi.org/10.18260/p.25896