Online teaching: a study for the effectiveness of randomized exams

Michael R. Santos
Sonoma State University

Vincent Richman
Sonoma State University

Jinglin Jiang
Sonoma State University

ABSTRACT

Online teaching is growing in popularity. This paper presents a case study for an online business course investigating student exam scores obtained from non-randomized and randomized exams. We find that the non-randomized exam scores of the students are suspicious because there are statistically significant disparities from the randomized exam scores. Further, the paper explores the constraints placed on instructors to construct randomized exams due to an insufficient number of questions in test banks provided by textbook publishers.

Keywords: online teaching, cheating, randomized exams, ANOVA analysis, F-tests

INTRODUCTION

Online course offerings are increasing worldwide perhaps due to locational convenience and touted as student-centric because instructors are facilitators in online teaching. However, there are serious concerns among instructors about the integrity of the online teaching about cheating and technological limitations to instructor-student interaction.

There are numerous published articles in the literature about online teaching relating to suspicious student activities. Majority of these publications focus on two themes: (1) ethical reasoning of students for cheating, and (2) tools to minimize cheating. Many researchers use survey responses from students and instructors as their method of analyses. For example, Simkin & Mcleod (2009) survey 144 students and find that approximately 60% of the business students and 64% of the non-business students admitted to cheating because they had the desire to get ahead. King et al. (2009) sample 121 undergraduate business students and find that 73.6% of the students think it is easier to cheat in an online course than in a traditional course. Josien et al. (2015) survey 256 students and 52 instructors using l6 different scenarios and find that while the instructors think take-home exams meant to be done individually, the students believe it is a collaborate work with their friends. Further, some students believe that attempting to cheat without succeeding is not academic dishonesty while faculty members do. Guyette et al. (2008) use a survey to measure 22 business faculty members' views about the questionable online exam behaviors and find that even though most faculty hold strict views on what constitutes cheating, there are different interpretations by administrators and younger faculty. Ruey-Shin et al. (2010) survey 25 faculty members and 257 students about the credibility of online testing, and 80 percent of the teachers and 98 percent of the students doubt the integrity of the online exam scores. Therefore Ruey-Shin et al. suggest a face recognition technology to identify students taking online exams. Granitz & Loewy (2007) apply a content analysis on the written records of students formally charged with plagiarizing to determine what reasoning the students used to justify their cheating and find that 41.8% of the students used deontology, 19.9% used situational ethics, and 18.4% used Machiavellianism. Josien & Britton (2013) survey 256 students and 95 of them report that they had cheated previously in college. Further, they find that freshmen and sophomores seem to cheat less than their junior and senior counterpart. Swartz & Cole (2013) investigate 162 student responses using a 12-question survey and find that 64.6% of the students think that academic integrity online or in the classroom was the same. Harmon & Lambrinos (2008) study the results from two economics courses that have identical structures except for the final exam in one of them was not proctored. Their findings support the idea that cheating took place in the non-proctored exam and, therefore they suggest online exams to be administered in a proctored environment to prevent academic dishonesty. Sheets & Waddill (2009) use survey data from 177 students, and approximately half of them reported cheating in the past. Moreover, they show that cheating was more prevalent among younger male students with relatively lower grade point averages.

Unfortunately, using identical exam questions in an online teaching environment without any monitoring may provide additional opportunities to cheating students who might share exam answers among each other through emails, text messages, phone calls, or sitting next to each other. For example, the study by Harmon and Lambrinos (2008) states that the exam scores from a face-to-face teaching environment may significantly differ from the purely online teaching environment (not proctored) because information transfer among students is less plausible with proctored exams. Relatedly, Jones (2009) suggests using large pools of exam questions to construct exams that none of the students will be asked identical test questions.

There are many suggestions on how to reduce online cheating in the literature and one of them is about constructing randomized exam questions. This paper tests the effectiveness of randomized exams in online teaching. In addition, the paper explores a relatively under-researched area; the difficulties and challenges faced by an average instructor in constructing randomized exams as stated in Rowe (2004) that "most instructors will not have the patience to provide an adequately large pool" for randomized exams. In summary, the motivation of our paper is twofold: to provide direct evidence for the effectiveness of randomized exams in online teaching, and highlight the difficulties that instructors face in creating randomized exams.

RESEARCH METHOD AND ANALYSIS

Quantitative research based on historical data is employed in this study. Historical data are actual student grades obtained from a professor teaching an elective undergraduate online business course; International Finance, at the School of Business and Economics of Sonoma State University. The online course was taught at one of the semesters between 2014 and 2018 period. There were 23 senior students enrolled in the course and all quizzes and exams were given online using the Moodle Learning Management System (LMS). During the semester, the students had taken three online synchronous exams. Each exam had 40 questions (multiple-choice and true-false). Exam 1 was not randomized, and the students received the same exam questions. After Exam 1, the professor realized that the average score was unusually high; 87.83/100, and therefore she decided to randomize the remaining two exams (Exams 2 and 3) by providing different sets of questions to each student. She randomized Exam 2 by 80.0% [=32/40] and Exam 3 by 92.5% [=37/40]. As a result, the average scores of the students at Exams 2 and 3 went down significantly (71.74/100 and 69.57/100 respectively).

Table 1 below summarizes the details of the quantitative research. Exam 1 is non-randomized and therefore needs only 40 identical questions. However, Exams 2 and 3 are randomized, and therefore each student receives a different set of questions. The degree of randomization at Exam 2 and 3 are 80.0% and 92.5% respectively indicating that there are several identical exam questions. In order to achieve 80.0% [=32/40] randomization in Exam 2, the professor uses a total of 744 questions [= (32 x 23) + 8] where 8 out of 40 questions were identical for every student in the class. Similarly, Exam 3 has 92.5% [=37/40] randomization with a total of 854 questions [= (37 x 23) + 3] where 3 out of 40 questions were identical for every student in the class.

Table 1: Summary of Exams and Randomization of Exam Questions

Exams with 40 Questions	Number of Students Taking Exam	Number of Chapters Covered	Non-Randomized (NR) or Randomized**	Percent of Randomization	Number of Questions Required for Exam*
Exam 1 (E1)	23	6	NR	0.0% [=0/40]	40
Exam 2 (E2)	23	3	R	80.0% [=32/40]	744 [=(32x23)+8]
Exam 3 (E3)	23	3	R	92.5% [=37/40]	854 [=(37x23)+3]

*Full randomization of an exam with 40 questions and 23 students requires 920 questions.
**The Moodle Learning Management System (LMS) is used and Moodle has randomization options in the preparation of quizzes using a text bank.

Table 2 further presents the specifics of student scores from Exam 1(NR), Exam 2(R), Exam 3(R) as well as average percent declines (increases) from Exam 1(NR) to Exam 2(R) and Exam 3(R). Group S [=SUSPICIOUS] refers to the student cases where the exam scores significantly go down after applying randomization in Exam 2 and 3. Students are placed in Group S if the average percent exam score decline (X) from Exam 1 to Exams 2 and 3 is less than -20% (X < -20%). There are 11 students assigned to Group S, and the remaining 12 students are put in Group N [=NORMAL or non-suspicious]. If the cut-off point is at -30% decline (X < -30%), the size of Groups S and N are 8 and 15 respectively. The average score of all students for Exam 1(NR), Exam 2(R), and Exam 3(R) are 87.8/100, 71.7/100, and 69.6/100 respectively, and the average percent exam score decline from Exam 1(NR) to Exam 2(R) and Exam 3(R) is -17.8%.

Table 2: Specifics of the Online Upper-Division (Elective) Course

NAMES	Exam 1(NR)	Exam 2(R)	Exam 3(R)	X = Average Percent Exam Score Decline/Increase from Exam 1	GROUPS*
Student 1	87.5	75	50.0	-28.60%	S
Student 2	82.5	100	97.5	19.70%	N
Student 3	97.5	70	65.0	-30.80%	S
Student 4	97.5	95	82.5	-9.00%	N
Student 5	95.0	90	87.5	-6.60%	N
Student 6	97.5	80	75.0	-20.50%	S
Student 7	97.5	55	60.0	-41.00%	S
Student 8	45.0	40	32.5	-19.40%	N
Student 9	62.5	70	70.0	12.00%	N
Student 10	85.0	80	72.5	-10.30%	N
Student 11	97.5	65	47.5	-42.30%	S
Student 12	92.5	90	85.0	-5.40%	N
Student 13	87.5	70	97.5	-4.30%	N
Student 14	100.0	57.5	62.5	-40.00%	S
Student 15	97.5	52.5	67.5	-38.50%	S
Student 16	97.5	37.5	47.5	-56.40%	S
Student 17	92.5	50	62.5	-39.20%	S
Student 18	77.5	90.0	90.0	16.10%	N
Student 19	97.5	67.5	72.5	-28.20%	S
Student 20	97.5	70.0	50.0	-38.50%	S
Student 21	87.5	87.5	85.0	-1.40%	N
Student 22	70.0	80.0	70.0	7.10%	N
Student 23	77.5	77.5	70.0	-4.80%	N
AVERAGES	**87.8**	**71.7**	**69.6**	**-17.8%**	

*Order of the students is shuffled, and students are identified as "Student" to protect their privacy. Group S represents suspicious cases for the students whose average percent grade from Exam 1 (non-randomized means identical exam for all students) to Exam 2 and Exam 3 (randomized means a different set of question to each student) the absolute value of declines less than -20% (X < - 20%). There are 11 students with "S" and 12 students with "N" which represent normal (non-suspicious) cases. If the cut-off point is changed to -30% (X < -30%), the size of the "S" group is 8 students and "N" is, therefore, is 14.

Table 3 shows the exam averages separately for Groups S and N. For Group S (left half of Table 3), the average Exam 1 (NR: with identical questions) score is unusually high at

96.4/100, and Exams 2 and 3 (R: with different set of questions) scores are relatively low at 61.8/100 and 60.0/100 respectively. Further, the average exam score decline from E1(NR) to E2(R) and E3(R) is noteworthy at -36.7%.

Table 3: The Average Scores and Declines or Increases in Group S an N

GROUP	E1(NR)	E2(R)	E3(R)	AVE %	GROUP	E1(NR)	E2(R)	E3(R)	AVE %
S	87.5	75.0	50.0	-28.6%	N	77.5	90.0	90.0	16.1%
S	92.5	50.0	62.5	-39.2%	N	87.5	87.5	85.0	-1.4%
S	97.5	67.5	72.5	-28.2%	N	70.0	80.0	70.0	7.1%
S	97.5	70.0	50.0	-38.5%	N	82.5	100.0	97.5	19.7%
S	97.5	70.0	65.0	-30.8%	N	97.5	95.0	82.5	-9.0%
S	97.5	80.0	75.0	-20.5%	N	95.0	90.0	87.5	-6.6%
S	97.5	55.0	60.0	-41.0%	N	45.0	40.0	32.5	-19.4%
S	97.5	65.0	47.5	-42.3%	N	62.5	70.0	70.0	12.0%
S	100.0	57.5	62.5	-40.0%	N	85.0	80.0	72.5	-10.3%
S	97.5	52.5	67.5	-38.5%	N	92.5	90.0	85.0	-5.4%
S	97.5	37.5	47.5	-56.4%	N	87.5	70.0	97.5	-4.3%
					N	77.5	77.5	70.0	-4.8%
AVE	**96.4**	**61.8**	**60.0**	**-36.7%**	**AVE**	**80.0**	**80.8**	**78.3**	**-0.5%**

On the other hand, for Group N, the score averages of Exam 1(NR), Exam 2(R), and Exam 3(R) do not change significantly at 80.0/100, 80.8/100, and 78.3/100 respectively. In addition, the average exam score decline from Exam 1 to Exams 2 and 3 is only -0.5%.

To understand how significantly the exam score averages differ from each other, a hypothesis test is provided below by using the ANOVA analysis:

$$H_0 = \mu_1 = \mu_2 = \mu_3 = 0$$
$$H_1 = \mu_1 \neq \mu_2 \neq \mu_3 \neq 0$$

H_0 (null hypothesis) claims that all exam score averages, μ_1, μ_2, and μ_3, for the population are equal and H_1 claims otherwise that the exam score averages are unequal.

Table 4 provides the ANOVA table results for Group S where F-statistics value (51.93) is significantly greater than the F-critical value (3.32), and therefore we reject the null hypothesis and conclude that the averages of the exams for Group S are not equal.

Table 4: Anova Single Factor for Group S

Groups	Count	Sum	Average	Variance
E1(NR)	11	1060	96.363636	11.704545
E2(R)	11	680	61.818181	156.36363
E3(R)	11	660	60	98.75

Source of	SS	df	MS	F	P-value	F crit
Between Groups	9236.363	2	4618.1818	**51.92504**	**1.80942E-10**	3.315829
Within Groups	2668.181	30	88.939393			
Total	11904.54	32				

Similarly, Table 5 provides the ANOVA table results for Group N, and the F-statistics value (.07) is less than F-critical value (3.28) and therefore, the null hypothesis, or the claim that the exam averages are equal, cannot be rejected for Group N.

Table 5: Anova Single Factor for Group N

Groups	Count	Sum	Average	Variance		
E1(NR)	12	960	80	225		
E2(R)	12	970	80.8333	252.6515		
E3(R)	12	940	78.3333	309.4696		
Source of Variation	**SS**	**df**	**MS**	**F**	**P-value**	**F crit**
Between Groups	38.8888	2	19.4444444	0.0741097	0.928723	3.28491
Within Groups	8658.33	33	262.373737			
Total	8697.22	35				

These ANOVA results from Tables 4 and 5 are consistent with the results shown in Table 3 where Group S has significant average score changes from Exam 1 (NR) at 96.4/100 to Exam 2 at 61.8/100 and Exam 3 (R) at 60.0/100. The findings raise questions about GROUP S students' performance by knowing that the only change from Exam 1 to Exams 2 and 3 was a switch to randomized questions. On the other hand, the average score change for Group N, from Exam 1 at 80.0/100 to Exam 2 at 80.8/100 and Exam 3 at 78.3/100, is relatively insignificant implying that randomization in Exams 2 and 3 did not affect Group N's performance.

Further, Table 6 summarizes three F-tests for Groups S and N. The first F-test has a null hypothesis with a claim that Exam 1(NR) and Exam 2(R) score variances are the same and therefore the variances (standard deviations) came from the same populations. The second F-test has a null hypothesis with a claim that Exam 1(NR) and Exam 3(R) score variances are the same and therefore the variances (standard deviations) came from the same populations. Moreover, the third F-test has a null hypothesis with a claim that Exam 2(R) and Exam 3(R) score variances are the same and therefore the variances (standard deviations) came from the same populations.

Table 6: F-tests for the Exam Scores of Exam 1(NR), Exam 2(R), and Exam 3(R)

HYPOTHESES FOR GROUPS S AND N	EXAMS COMPARED	NULL AND ALTERNATIVE HYPOTHESES
$H_0 = \sigma_1^2 = \sigma_2^2 = 0$ $H_1 = \sigma_1^2 \neq \sigma_1^2 \neq 0$	Exam 1 (NR) versus Exam 2 (R).	H_0 = Exam 1 and 2 were sampled from populations with identical variances. H_1 = Exam score averages (population) are unequal.
$H_0 = \sigma_1^2 = \sigma_3^2 = 0$ $H_1 = \sigma_1^2 \neq \sigma_3^2 \neq 0$	Exam 1 (NR) versus Exam 3 (R).	H_0 = Exam score averages (population) are equal. H_1 = Exam score averages (population) are unequal.
$H_0 = \sigma_2^2 = \sigma_3^2 = 0$ $H_1 = \sigma_2^2 \neq \sigma_3^2 \neq 0$	Exam 2 (NR) versus Exam 3 (R).	H_0 = Exam score averages (population) are equal. H_1 = Exam score averages (population) are unequal.

Table 7 provides the details of the F-test statistics calculated at α=.05 critical value. The results for Group S are presented at the upper half of Table 7. It is evident that in the case of Group S, the null hypothesis, the variances of Exams 1(NR) and Exam 2(R) are equal, is not supported because F-value, 13.36, is significantly greater than F-critical value of 2.98 [or P(F<=f) one-tail value, 0.000163, is below α=.05]. Similarly, the claims of null hypotheses, that assumes Exam 1(NR) and Exam 3(R) score variances are equal, is not supported because F-value, 8.44, is greater than F-critical value of 2.98 [or P(F<=f) one-tail value, 0.001165, is below α=.05]. Therefore, it can be concluded that the variance populations for Exam 1(NR) and Exam 3(R) are not identical. This finding is consistent with the earlier results of the ANOVA at Table 4 indicating that the variances (or standard deviations) of Exam 1(NR) are coming from a different population.

In contrast, for Group S, the null hypothesis, claiming that the variances of Exam 2(R) and Exam 3(R) scores are equal, cannot be rejected because F-value, 1.58, is less than F-critical value, 2.98 [or P(F<=f) one-tail value, 0.240167, is greater than α=.05].

The bottom half of Table 7 presents F-values for Group N, indicating that the variances of Exam 1(NR) in comparison to either Exam 2(R) or Exam 3(R) are equal. Similar, the variances of Exam 2(R) and Exam 3(R) are equal. Again, this is a consistent result with the earlier results of ANOVA from Table 5 indicating that the variances (or standard deviations) of Exams 1(NR), Exam 2(R), and Exam(R) are coming from the same population for Group N.

Table 7: F-tests and Results

H_0: Null Hypothesis H_1: Alternative Hypothesis	$H_0 = \sigma_1^2 = \sigma_2^2 = 0$ $H_1 = \sigma_1^2 \neq \sigma_1^2 \neq 0$		$H_0 = \sigma_1^2 = \sigma_3^2 = 0$ $H_1 = \sigma_1^2 \neq \sigma_3^2 \neq 0$		$H_0 = \sigma_2^2 = \sigma_3^2 = 0$ $H_1 = \sigma_2^2 \neq \sigma_3^2 \neq 0$	
GROUP S	**E2(R)**	**E1(NR)**	**E3(R)**	**E1(NR)**	**E2(R)**	**E3(R)**
Mean	61.82		60		61.82	60
Variance	156.36		98.75		156.36	98.75
Observations	11	11	11	11	11	11
df	10	10	10	10	10	10
F	**13.36**		**8.44**		1.58	
P(F<=f) one-tail	**0.000163**		**0.001165**		0.240167	
F Critical one-tail	**2.98**		**2.98**		2.98	
GROUP N	**E2(R)**	**E1(NR)**	**E3(R)**	**E1(NR)**	**E3(R)**	**E2(R)**
Mean	80.83	80	78.33	80	78.33	
Variance	252.65	225	309.47	225	309.47	
Observations	12	12	12	12	12	12
df	11	11	11	11	11	11
F	1.12		1.38		1.22	
P(F<=f) one-tail	0.425497		0.303039		0.371243	
F Critical one-tail	2.82		2.82		2.82	

RANDOMIZATION OF ONLINE EXAMS AND THE CHALLENGES

Randomization of an exam is demanding for the instructors because it requires a deep test bank with many questions readily available. However, the textbook publishers provide test banks that often include an insufficient number of questions to construct fully randomized exams. Most test banks from the publishers have 50 to 100 questions per chapter for introductory business classes. This number might be much lower for the test banks of elective business courses, perhaps ranging from 30 to 60 questions per chapter.

Table 8 presents a hypothetical matrix to calculate how many questions needed to construct a fully randomized exam depending on (a) Number of Students Taking the Exam (Row 2), and (b) Number of Questions in the Exam (Column 2). The gray shaded matrix cells on Table 8 shows the required number of questions needed in constructing a fully randomized exam and obtained by multiplying (Number of Students Taking the Exam) x (Number of Questions in the Exam).

Table 8: Feasibility of Randomized Exams*

		Number of Students Taking the Exam									
		10	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
Number of Questions in the Exam	**10**	100	200	300	400	500	600	700	800	900	1,000
	20	200	400	600	800	1,000	1,200	1,400	1,600	1,800	2,000
	30	300	600	900	1,200	1,500	1,800	2,100	2,400	2,700	3,000
	40	400	800	1,200	1,600	2,000	2,400	2,800	3,200	3,600	4,000
	50	500	1,000	1,500	2,000	2,500	3,000	3,500	4,000	4,500	5,000
	60	600	1,200	1,800	2,400	3,000	3,600	4,200	4,800	5,400	6,000
	70	700	1,400	2,100	2,800	3,500	4,200	4,900	5,600	6,300	7,000
	80	800	1,600	2,400	3,200	4,000	4,800	5,600	6,400	7,200	8,000
	90	900	1,800	2,700	3,600	4,500	5,400	6,300	7,200	8,100	9,000
	100	1,000	2,000	3,000	4,000	5,000	6,000	7,000	8,000	9,000	10,000

It should be stressed that constructing a fully randomized exam is too difficult and, at times, it is not even feasible if the online class has too many students and/or too many questions in an exam. Therefore, the main obstacle to constructing fully randomized exams lies in the fact that there is a limited number of questions available in today's test banks provided by textbook publishers. Alternatively, instructors can write their own exam questions. However, writing 500, 1,000, or more questions per exam is beyond the capacity of any instructor. As a result, instructors teaching online courses may give up from fully randomized exams when there are limited number of questions and ask proportionally more of identical questions in an exam which may in return reduce the effectiveness of online exams against suspicious student activities.

CONCLUSION AND SUGGESTIONS

A suggested method to reduce suspicious student activities is to provide randomized exams to students. The findings of this paper show that randomizing exams is relatively effective and reduces the student payoffs from suspicious activities. However, constructing fully randomized exams is challenging for instructors because there is a limited number of questions in a typical test bank provided by college textbook publishers. The lack of readily available questions practically limits the applications of fully randomized exam questions in online teaching. Therefore, the college textbook publishers should be aware of the increasing number of online course offerings that require a deep test bank with numerous questions to construct randomized exams.

REFERENCES

Granitz, N., & Loewy, D. (2007). Applying Ethical Theories: Interpreting and Responding to Student Plagiarism. *Journal of Business Ethics*, 72 (3), 293-306.

Guyette, R. W. Jr., & King, C. G., & Piotrowski, C. (2008). Business Faculty Views of Online Course Cheating: Evidence for a Cohort Effect, *Organization Development Journal*, 26(4), 25-31.

Harmon, O. R., & Lambrinos, J. (2008). Are Online Exams an Invitation to Cheat? *Journal of Economic Education*, 39(2), 116-121.

King, C. G, Guyette, R. W. Jr., & Piotrowski, C. (2009). Online Exams and Cheating: An Empirical Analysis of Business Students' Views. *The Journal of Educators Online*, 6(1): 1-11.

Josien, L., & Broderick, B. (2013). Cheating in Higher Education: The Case of Multi-Methods Cheaters. *Academy of Educational Leadership Journal*, 17(3), 93-105.

Josien, L., & Seeley, E., & Csipak, J., & Rampal, R. (2015). Cheating: Student and Faculty's Perception on Potential Cheating Activity. *Journal of Legal, Ethical and Regulatory Issues*, 18(2), 21-37.

Jones, I. M. (2009). Cyber-Plagiarism: Different Methods-Same Song. *Journal of Legal, Ethical and Regulatory Issues*, 12(1), 89-100.

Rowe, N. (2004). Cheating in Online Student Assessment: Beyond Plagiarism. *Online Journal of Distance Learning Administration*, VII(11), 1-8.

Ruey-Shin, C., & Kevin K. H., & Tsai, C. H. (2010). The implementation of face recognition technology and its effect on e-quiz credibility. *Asian Journal on Quality*, 11(2), 125-136.

Sheets, B., & Waddill, P. (2009). E-Cheating Among College Business Students: A Survey. *Information Technology, Learning, and Performance Journal*, 25(2), 4-19.

Simkin, M. G., & Mcleod, A. (2010). Why Do College Students Cheat? *Journal of Business Ethics*, 94(3), 441-453.

Swartz, L. B., & Cole, Michele T. (2013). Students' Perception of Academic Integrity in Online Business Education Courses. *Journal of Business and Educational Leadership*, 4(1), 102-112.