Online Academic Exams: Does Multiplicity of Exam Versions Mitigate Cheating?

BSE Working Paper 1430

February 2024 (Revised November 2024)

Flip Klijn, Mehdi Mdaghri Alaoui, Marc Vorsatz

bse.eu/research

Online Academic Exams: Does Multiplicity of Exam Versions Mitigate Cheating?*

Flip Klijn† Mehdi Mdaghri Alaoui‡ Marc Vorsatz§

October 26, 2024

Abstract

We study academic integrity in a final exam of a game theory course with 463 undergraduate students at a major Spanish university. The exam is an unproctored online multiple-choice exam without backtracking. A key characteristic is that for each (type of) problem, groups of students receive different versions. Moreover, each problem version is assigned to one subgroup during one stage of the exam and to another subgroup during an immediately consecutive later stage. Thus, we can exploit grade points and timestamps to study students' academic integrity. We observe a significant decrease in completion time at each later stage; however, surprisingly, there is no corresponding impact on average grade points. The precise number of different versions does not seem to have an effect on either variable. Our findings thus suggest that employing a limited number of distinct problem versions (as few as two) can diminish cheating effectiveness in online exams.

Keywords: field experiment; academic integrity; online exam; multiple versions; completion time.

JEL–Numbers: A22, C93, D9, I21, I23.

*We thank the Editor, two anonymous reviewers, Lata Gangadharan, and Erik Wengström for useful comments and suggestions.

†Corresponding author. Institute for Economic Analysis (CSIC) and Barcelona School of Economics, Campus UAB, 08193 Bellaterra (Barcelona), Spain; e-mail: flip.klijn@iae.csic.es. He gratefully acknowledges financial support from AGAUR–Generalitat de Catalunya (2021-SGR-00416) and the Spanish Agencia Estatal de Investigación (MCIN/ AEI /10.13039/501100011033) through grant PID2020-114251GB-I00, PID2023-147136NB-I00, and the Severo Ochoa Programme for Centres of Excellence in R&D (Barcelona School of Economics CEX2019-000915-S).

‡Department of Economics and Business, Universitat Pompeu Fabra, C/ Ramon Trias Fargas 25, 08005 Barcelona, Spain; e-mail: mehdi.mdaghri@upf.edu.

§Departamento de Análisis Económico, Universidad Nacional de Educación a Distancia (UNED), Paseo Senda del Rey 11, 28040 Madrid, Spain; e-mail: mvorsatz@cee.uned.es. He gratefully acknowledges financial support from the Spanish Ministry of Science and Innovation through grant PID2021-122919-NB-I00.

1 Introduction

Academic integrity promotes learning, ensures that all students are evaluated based on their own efforts and abilities, and is essential for the reputation of educational institutions and the degrees they confer. Therefore, faced with the task of preparing a final exam of an academic course, instructors usually create a number of different (but similar) versions of exam questions with the aim to reduce potential cheating. This seems an obvious and effective measure in the case of classroom exams, where neighboring students are to receive different exam versions.

Randomized field experiment

In this paper, we investigate whether having a number of different exam versions may also mitigate potential cheating at online exams without proctoring. Our study was conducted during the Covid-19 pandemic, a time when Spanish public universities were unprepared for online exams and lacked standardized proctoring systems. Instructors had to independently devise their own measures to prevent online cheating, leading to a variety of informal and ad hoc strategies. Even today, many Spanish public universities still do not employ proctoring systems for online assessments. Possible challenges with online proctoring systems in the Spanish context include: (a) limited availability in co-official languages, (b) potentially high costs for universities or students, and, most importantly, (c) concerns about privacy.[1] Our study does not aim to invalidate online proctoring systems or compare the efficiency of different cheating deterrence methods. Instead, it addresses what can be done when there is no proctoring system in place.

Recent studies of online exams during the Covid-19 pandemic showed that high levels of cheating are very common (see, e.g., Basken, 2020, Janke et al., 2021, and Newton and Essex, 2023). Due to the increasing importance of online education and examination (see, e.g., Imran et al., 2023 and Ratten, 2023), our research question holds significant practical relevance. We aim to study whether a larger number of different versions leads to a more prominent mitigation of cheating. Our randomized field experiment is based on the final exam of an introductory course on game theory that took place at Universitat Pompeu Fabra, Spain in the second trimester of academic year 2020-2021. The four main characteristics of this online exam are as follows. First, students took the exam simultaneously. Second, the exam consisted of multiple-choice questions grouped into problems. Each problem appeared randomly at an earlier "round" (stage of the exam) for half of the students and at a later round for the other students. Third, backtracking was not possible, i.e., once a student moved to the next problem, there was no possibility to go back to the previous problem to change his/her answers. Fourth, for each of the problems there were 2, 4, or 6 different versions and each student was randomly assigned one of the versions.

Main results

For any given problem, we compare the performance of the students that face the problem in an earlier round with the performance of the students that face the same problem in a later round.[2]

[1] See also the article `https://theconversation.com/universities-shouldnt-use-software-to-monitor-online-exams-heres-why-188327` in The Conversation (August 12, 2022) on ethical concerns regarding proctoring software; and the article `https://www.nytimes.com/2022/08/25/us/remote-testing-student-home-scan-privacy.html` in the New York Times (August 25, 2022) on a U.S. federal judge's ruling about potential intrusion of proctoring software and the Fourth Amendment.

[2] Recall that both groups receive the same set of versions of the problem.

We measure performance through correctness and completion time. The *average correctness*[3] of a given problem in a particular round is defined as the average points per question by the students that face the problem in that round. The *average completion time* of a problem in a particular round is the average time taken for the problem by the students that face the problem in that round.

As expected, we find that the average completion time decreases significantly for each problem, i.e., on average, students in the later rounds finish each problem faster. Taking the average over all problems, the average completion time of the later rounds is 19.5% shorter than the earlier rounds. However, surprisingly, we find that the average correctness of each but one problem decreases, i.e., on average, students in the later rounds obtain lower scores (but not significantly so).

Related literature

Our study has similarities with other recent field studies.[4] Bilen and Matros (2021) investigated cheating behavior during an online examination of a large public university amid a Covid-19 lockdown. Students received the same set of questions in a random order, without multiple-choice options and without backtracking. Bilen and Matros (2021) examined correctness and completion times, focusing in particular on two students who displayed unusual time allocations and exceptional performance relative to midterm results.

Klijn et al. (2022) presented a randomized field experiment to study academic integrity at an online exam of a large public university in Spain during a strict Covid-19 lockdown. Students received the same set of questions in a random order, with multiple-choice options but without backtracking. Most importantly, in contrast with Bilen and Matros (2021), the assigned orders were crafted with the objective of systematically examining later-round effects. Klijn et al. (2022) found a significant later-round advantage in terms of correctness (7.7%) as well as completion time (18.1%). Since the exam questions in Klijn et al. (2022) are common (i.e., one exam version), the results of our current study suggests that having as few as two exam versions can already reduce the efficacy of cheating in terms of grade points.[5] Klijn et al. (2022) used a reminder of the university's code of ethics as a honesty nudge, sending it to a subgroup of students halfway through the exam; however, the nudge did not affect cheating levels.[6]

Vazquez et al. (2021) conducted a randomized field experiment at a large public university in Illinois to explore the impact of proctoring on exam grades across two classes (face-to-face and online) of an introductory microeconomics course. Face-to-face exams had live proctors or were unproctored and online exams had web-based proctors or were unproctored. Students whose exams were not proctored scored, on average, over 11% higher than those whose exams were proctored.

[3]For the precise definitions we refer to Section 2.

[4]We refer to Holden et al. (2021) and Noorbehbahani et al. (2022) for comprehensive reviews on cheating in a variety of online examinations and settings that are less related to our focus and approach.

[5]While the types of exams studied in Klijn et al. (2022) and the current paper are similar, there are limitations to comparing the findings due to different populations of students, distinct (Covid-19) circumstances, and slightly different evaluation schemes.

[6]Recently, Le Maux and Necker (2023) implemented a wheel of fortune game using Amazon Mechanical Turk (MTurk). In each of ten rounds, each participant spinned a wheel. The task was to guess the outcome of the wheel spin and then to report whether the guess was correct. Le Maux and Necker (2023) examined the impact of honesty nudges. They found that reminding individuals about the right thing to do increased honesty. Moreover, including information that it is possible to assess an individual's dishonesty strengthened the effect.

A notable difference between Vazquez et al. (2021) and our study lies in the treatment variables: while we vary the order of problems, Vazquez et al. (2021) compared students who are either proctored or unproctored, with the treatment groups examined at different time points.

Using data from an economics program at a Dutch university, Arnold (2022) studied the potential impact of transitioning from face-to-face proctoring to online proctoring services on cheating behaviors. He found that variables measuring students' human capital remained strongly related to course grades. Moreover, the exam data from online proctoring did also not exhibit suspicious grade patterns. The findings in Arnold (2022) suggest that examination with online proctoring services is not necessarily more conducive to cheating than face-to-face proctoring. In our study, we test whether it is possible to mitigate potential cheating without recurring to online proctoring services.

Elzinga and Harper (2023) compared student performance across two semesters of the same large introductory economics course at the University of Virginia, one taught in-person in 2019 and the other taught online in 2020. Between the 2019 final exam and the 2020 final exam, 61 out of 75 questions were identical. At the aggregate level, Elzinga and Harper (2023) did not find differences in student performance between online and in-person instruction. In our study, we do not compare two different student cohorts nor do we compare different exam or teaching modalities.

Hill and LoPalo (2024) studied the performance of students at midterm exams in two large, introductory courses at a US state university. Students were divided into two groups with the first group assigned to take exam 1 online and unproctored and exam 2 in-person and proctored, and the opposite for the second group. Hill and LoPalo (2024) found that students, on average, performed considerably better on online tests. This effect was concentrated at the lower end of the distribution of scores. Hill and LoPalo (2024) also evaluated several potential mechanisms for the difference between the two exam modalities. They found evidence consistent with students cheating by using unauthorized resources but little evidence of differences in effort before or during the exam or in testing anxiety across modalities. Finally, their analysis suggests a possible implementation of online testing for instructors that want to avoid the potential privacy concerns of webcam proctoring: presenting students with "new" questions significantly attenuates cheating. In our current study we find a complementary solution: a limited number of distinct exam versions can also diminish cheating effectiveness in online exams.

Organization

In Section 2, we describe the field experiment and our hypotheses. In Section 3, we present our results. The Online Appendix contains details about the course, screenshots of the final exam, a summary of the nomenclature, subject pool information, additional analyses, and a sample exam.

2 Randomized field experiment

Our randomized field experiment is the final exam of an introductory course on game theory that took place at Universitat Pompeu Fabra in the second trimester of academic year 2020-2021. The final exam was part of a continuous evaluation scheme which is discussed in more detail in Online Appendix A. The final exam was programmed and executed in Moodle. All 463 students started

the exam around the same time. For convenience and future reference, the concepts defined in the following sections are collected in Online Appendix B.

Design of final exam

The final exam consisted of *20 multiple-choice questions which were distributed over 6 problems*.[7] For each question we fixed a number of possible answers (of which only one was correct). For each question and for each student, the order of the possible answers was chosen randomly. Selecting the correct answer gave 5 points, an incorrect answer 1 negative point, and not answering 0 points. Students did not receive any feedback whatsoever on their answers until two weeks after the exam.

The first screen[8] provided the exam instructions, which included information on the number of questions, the number of problems, the number of points for a correct/incorrect/blank answer, and a reminder of the maximal duration (120 minutes). Moreover, it was emphasized in boldface that *moving back to a previous problem would not be possible*. Along with the instructions, part of the university's code of ethics was displayed, including the following excerpt:

> *"Truthfulness in academic assessments. ... Copying and plagiarism are forms of misconduct to which the corresponding prescribed punishments must be applied, not only to demonstrate the university community's rejection thereof but also to prevent the reputation of the University and its graduates being harmed. ..."*

After subscribing to the code of ethics by clicking on the "continue" button, the student faced the first problem. All questions that pertained to the same problem appeared on the same screen. Each subsequent problem (and all its questions) appeared on a new screen, but only after answering the previous problem or leaving it completely or partially unanswered purposely. For each of the 6 problems there were a number of different versions[9] obtained by applying permutations, scaling, etc. of numerical values, which should not vary in their level of difficulty.

Table 1 indicates for each problem the number of different versions and the number of questions (which is the same for each version of the problem). Furthermore, Table 1 indicates which pair of problems appeared in which two rounds, i.e., stages of the exams. For instance, for each student, problem I appeared either in round 1 or 2. *Throughout the paper we use the following nomenclature*. Rounds 1, 3, and 5 (rounds 2, 4, and 6) will be called *earlier rounds* (*later rounds*). When discussing problem I or II, the earlier group of students (or *earlier students*) refers to the group of students that worked on the problem in round 1, while the later group of students (or *later students*) refers to the group of students that worked on the problem in round 2. Similar terminology is used for the other two pairs of problems.

[7] See Online Appendix J for a sample exam and Online Appendix I for an estimate of the relationship between the final exam grade and the continuous assessment during the course.

[8] See Online Appendix A for screenshots.

[9] So, all students that are given the same version of a given problem face the exact same questions that belong to the problem.

rounds	problem	# versions	# questions	questions
1 and 2	I	2	3	[1,2,3]
	II	6	3	[4,5,6]
3 and 4	III	2	2	[7,8]
	IV	4	5	[9,10,11,12,13]
5 and 6	V	4	3	[14,15,16]
	VI	6	4	[17,18,19,20]

Table 1: Number of versions and questions for each of the 6 problems (labeled I, II, III, IV, V, VI). In the first two rounds, each student is faced with problems I and II. For each student, the particular order I,II or II,I was determined randomly. The other two pairs of rounds had similar random assignments.

Table 2 exhibits the $2 \times 2 \times 2 = 8$ (feasible) orders of the problems. For each of the 8 orders A,B,...,H, we indicate the resulting order of the questions and the number of students that were randomly assigned to it. For instance, the seventh question for students with order G was a "question 9" (the student is randomly assigned one of the four versions of problem IV (questions 9–13), see Table 1). Students were not informed of the existence of different orders and different versions of problems. We refer to Online Appendix C for an analysis of our subject pool.

order	round						# students
	1	2	3	4	5	6	463
A	**I**	**II**	**III**	**IV**	**V**	**VI**	
	[1,2,3]	[4,5,6]	[7,8]	[9,10,11,12,13]	[14,15,16]	[17,18,19,20]	56
B	**II**	**I**	**III**	**IV**	**V**	**VI**	
	[4,5,6]	[1,2,3]	[7,8]	[9,10,11,12,13]	[14,15,16]	[17,18,19,20]	60
C	**I**	**II**	**IV**	**III**	**V**	**VI**	
	[1,2,3]	[4,5,6]	[9,10,11,12,13]	[7,8]	[14,15,16]	[17,18,19,20]	56
D	**II**	**I**	**IV**	**III**	**V**	**VI**	
	[4,5,6]	[1,2,3]	[9,10,11,12,13]	[7,8]	[14,15,16]	[17,18,19,20]	58
E	**I**	**II**	**III**	**IV**	**VI**	**V**	
	[1,2,3]	[4,5,6]	[7,8]	[9,10,11,12,13]	[17,18,19,20]	[14,15,16]	61
F	**II**	**I**	**III**	**IV**	**VI**	**V**	
	[4,5,6]	[1,2,3]	[7,8]	[9,10,11,12,13]	[17,18,19,20]	[14,15,16]	59
G	**II**	**I**	**IV**	**III**	**VI**	**V**	
	[4,5,6]	[1,2,3]	[9,10,11,12,13]	[7,8]	[17,18,19,20]	[14,15,16]	56
H	**I**	**II**	**IV**	**III**	**VI**	**V**	
	[1,2,3]	[4,5,6]	[9,10,11,12,13]	[7,8]	[17,18,19,20]	[14,15,16]	57

Table 2: Orders (A,B,...,H) of the 6 problems (I, II, III, IV, V, VI) and the distribution of the 20 questions (labeled $1, \ldots, 20$) over the problems. Questions are in brackets [ ] to indicate that they appeared simultaneously (and in this order) on the same screen. The numbers on the right hand side indicate the number of students that were assigned to each order.

Hypotheses

We expect that (correct or incorrect) solutions/answers to any given question in the earlier round accumulate and start to circulate so that students that are confronted with the same question in the later round are more likely to make a "more informed" decision, inducing more correct and/or quicker answers.[10] Formally, the *average correctness* of a given problem in a particular round is defined as the average points per question by the students that were faced with the problem in that round. The *average completion time* of a problem in a particular round is the average time taken to answer the problem by the students that were faced with the problem in that round.[11] In a setting with the exact same questions for all students, Klijn et al. (2022) found clear evidence for a later-round advantage, i.e., later students perform better than earlier students. The same hypothesis in the current setting is stronger because of the existence of multiple versions.

Hypothesis 1 (**Order effect: later-round advantage**). *For each pair of problems (I,II), (III,IV), and (V,VI), the later round presents higher average correctness and shorter average completion time.*

Assuming that our analysis does not reject Hypothesis 1, we expect that increasing the number of different versions also increases the mitigation of cheating.

Hypothesis 2 (**Mitigation through multiplicity of versions**). *For each pair of problems (I,II), (III,IV), and (V,VI), the later-round advantage is less pronounced for problems with a larger number of different versions.*

3 Results

Considering the earlier round of each problem, (almost all) different versions have similar average correctness and similar average completion time.[12] For this reason we pool the data from different versions for the same problem and the same round.

Analysis of the evidence for Hypothesis 1

First, we analyze whether students who face a problem later on in the exam have an advantage in the form of a higher average correctness and a shorter average completion time. The box plots in Figure 1 visualize the data. In each box plot, red circles ○ indicate outliers and the white diamond ◇ corresponds to the group average. The thicker dark blue line within a box plot is the median of the empirical distribution.[13]

[10] It cannot be excluded that students perform better in later rounds due to an increased familiarity with the exam. It is also possible that students reduce the time spent on questions because they get more time-constrained as the exam progresses. Since there is no control treatment in which it is impossible for students to exchange information, it is difficult to single out these effects. Nevertheless, our main conclusion that multiple exam versions are rather effective at reducing the possibility of cheating is not affected by these confounding factors.

[11] A problem is considered completed by a student when he/she moves to the next problem/screen. Note that the answers to any questions of the same problem can be changed until the student moves to the next screen. The completion time of a problem, measured in minutes where the minimum is 1 minute, also includes the time used to read and think about any of its constituent questions and the possible decision of leaving some of the questions unanswered. Due to Moodle limitations, we could not track the time a student spent on individual questions.

[12] We refer to Online Appendix D for details.

[13] The scatter plots of the subject data in Figure 4, Online Appendix E, complement the box plots.

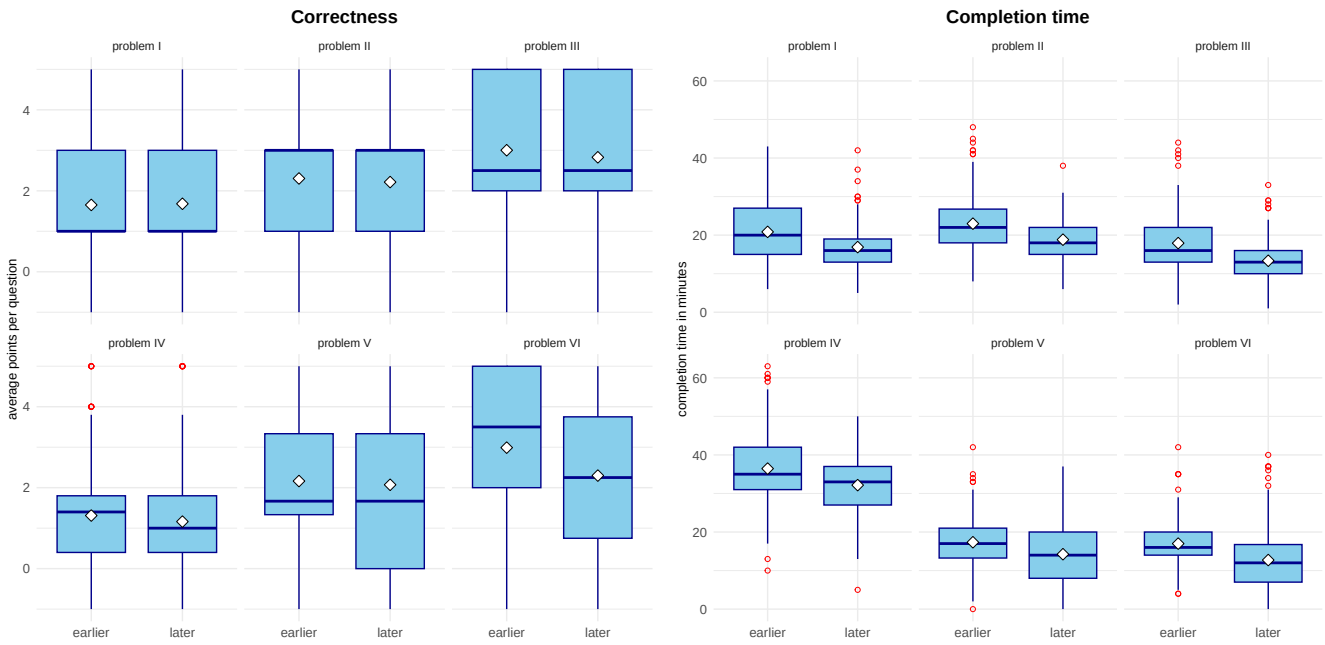


Figure 1: Box plots of correctness and completion time.

Figure 1 supports Hypothesis 1, on a possible later-round advantage, only partially. Specifically, later students have on average a shorter completion time than earlier students, which is in line with Hypothesis 1. However, we also observe that for each problem, the later students obtain on average fewer points per question than earlier students, which is clearly not in line with Hypothesis 1.

To formally analyze Hypothesis 1, we statistically test for order effects between earlier and later students in Table 3, for both correctness and completion times. Surprisingly, for almost all problems, the average points per question is not higher among the later students.[14] The only exception is problem I, but the 1.75% increase is not statistically significant. However, for each problem, the average completion time is significantly lower among the later students. Thus, we can reject Hypothesis 1 concerning correctness, but not concerning completion time. We argue that the decrease in average completion time can be attributed to cheating rather than other factors, such as increased familiarity with the exam format or perceived time constraints (see Footnote 10). If these additional factors indeed have an effect, it can be expected that this impact would manifest more prominently in the later stages of the exam. However, we observe a systematic decrease of the average completion time in all three pairs of rounds (1,2), (3,4), and (5,6).[15] Moreover, students' grades are determined by the points obtained for the questions, i.e., not by the completion times. However, we find an almost systematic (but non-significant) decrease of the average points per question for all problems in each of the later rounds (rounds 2, 4, and 6).

[14] Appendix F gives insights into changes of the answers from earlier to later rounds by providing the distributions of correct, incorrect, and blank answers.

[15] Recall that these are the three relevant pairs of rounds over the course of the experiment. In this context, Table 8 in Online Appendix G provides, for each version, an overview of the average time elapsed since the start of the exam.

problem	avg. points per question: earlier	later	% *in*crease	p	avg. completion time (min): earlier	later	% *de*crease	p
I	1.65	1.68	1.75	0.993	20.8	16.9	18.9	**0.000**
II	2.30	2.22	-3.88	0.581	23.0	18.8	18.1	**0.000**
III	3.00	2.83	-5.82	0.485	17.9	13.4	25.5	**0.000**
IV	1.31	1.16	-11.40	0.412	36.5	32.2	11.8	**0.000**
V	2.16	2.07	-4.26	0.824	17.3	14.2	17.9	**0.000**
VI	2.99	2.30	-23.09	**0.000**	17.0	12.7	25.1	**0.000**

Table 3: Impact of order of problems. Recall that the correct answer to any question yields 5 points, an incorrect answer 1 negative point, and not answering 0 points. The % increase/decrease is computed for "later" relative to "earlier." We employ Mann-Whitney U tests at the student level.

A possible concern is that the above analysis of rounds 3 and 4 may be impeded by different starting times of round 3 for students with different orders in rounds 1 and 2. According to Table 3, the students who had problem I in round 1 and problem II in round 2 required on average 39.7 minutes to complete the first two problems. Similarly, the students who had problem II in round 1 and problem I in round 2 required a very similar time: on average 39.9 minutes. In fact, we cannot reject the hypothesis that the two groups started round 3 at the same time ($p = 0.975$, Mann-Whitney U test).

Analogously, we verify that students with different orders of the first four problems started round 5 at the same time. Specifically, for the four different histories (I,II,III,IV), (I,II,IV,III), (II,I,III,IV), and (II,I,IV,III), the average time required to complete the first 4 problems is 90.6, 88.7, 91.0, and 91.0 minutes, respectively. Since the two-sided p-value of the Kruskal-Wallis test for joint equality is equal to 0.235, we cannot reject the hypothesis that the four groups started round 5 at the same time.

Analysis of the evidence for Hypothesis 2

We now consider Hypothesis 2 regarding the impact of the number of versions (see Table 1). Given that Hypothesis 1 has been rejected with respect to correctness, the corresponding part of Hypothesis 2 becomes void. *Therefore, the analysis below will focus exclusively on completion times.* For rounds 1+2 we observe in Table 3 that problem I (with 2 versions) has only a slightly stronger reduction of average completion time relative to problem II (with 6 versions). However, in rounds 3+4, problem III (with 2 versions) has a stronger reduction of average completion time relative to problem IV (with 4 versions). Finally, in rounds 5+6, we obtain the opposite direction: problem V (with 4 versions) has a weaker reduction of average completion time relative to problem VI (with 6 versions).[16]

As mentioned above, for each of the three pairs of rounds 1+2, 3+4, and 5+6, the odd-numbered problem (I, III, V, respectively) has a smaller number of versions than the even-numbered problem (II, IV, VI, respectively). Thus, we estimate, for each pair of problems I+II, III+IV, and V+VI, the following econometric model via OLS in order to analyze whether a problem with fewer versions

[16]However, this finding in rounds 5+6 may be affected by last minute decisions.

has a stronger reduction of completion time:[17]

$$\texttt{time}_{sj} = \beta_0 + \beta_1 \texttt{points}_{sj} + \beta_2 \texttt{odd}_j + \beta_3 \texttt{later}_{sj} + \beta_4 \texttt{odd}_j \cdot \texttt{later}_{sj} + \varepsilon_{ij},$$

where $\texttt{time}_{sj}$ is the completion time of student s for problem j, $\texttt{points}_{sj}$ are the total points student s scores on problem j (this variable serves as a control), $\texttt{odd}_j$ is a dummy variable that takes value 1 if j is an odd number and 0 if j is an even number, and $\texttt{later}_{sj}$ is a dummy variable that takes value 1 if student s faces problem j in the later round and 0 if the student solves problem j in the earlier round. Table 4 presents the estimation results.

variable	problems I+II	problems III+IV	problems V+VI
`intercept`	22.773***	36.270***	11.767***
	(0.523)	(0.553)	(0.590)
`points`	0.030	0.027	0.437***
	(0.044)	(0.043)	(0.034)
`odd`	-2.077***	-18.512***	2.740***
	(0.611)	(0.665)	(0.631)
`later`	-4.145***	-4.270***	-3.058***
	(0.605)	(0.665)	(0.610)
`odd · later`	0.193	-0.290	0.076
	(0.856)	(0.940)	(0.856)
observations	926	926	926

Table 4: OLS regression on completion time. Standard deviations are in parenthesis. *** means that the variable is significant at 0.001.

The OLS regression on completion time in Table 4 shows that later-round students need on average less time (the estimated `later` is significantly smaller than 0 in all three estimations), which is consistent with the non-parametric tests from Table 3. The interaction variable $\texttt{odd}_j \cdot \texttt{later}_{sj}$ captures the diff-in-diff effect described in Hypothesis 2. Since the estimated β_4 is never significantly different from 0, we cannot reject the null hypothesis that the reduction of the completion time is independent of the number of versions of a problem.

Finally, in addition to the above OLS regression on completion time, we carry out a permutation test to evaluate Hypothesis 2. In an ideal scenario, we would conduct a non-parametric test based on observations of each individual's possible later-round effect. However, this is not possible because each student faces each problem only once. Therefore, we proceed by creating pairs of students and making comparisons within each pair as follows. Consider the pair of rounds $1 + 2$ (the procedure for the pairs of rounds 3+4 and 5+6 is similar). Let S_I be the set of students who faced problem I in round 1 and S_{II} the set of students who faced problem II in round 1. From Table 2 it follows that $|S_I| = 230 \leq 233 = |S_{II}|$.

We randomly match each student from S_I with some student from S_{II}, so that each student from S_{II} is matched with at most one student from S_I. Thus, we obtain an injection (one-to-one

[17]Online Appendix H presents an alternative, joint data OLS regression. The results from this joint regression perfectly align with those obtained from the three separate regressions presented here.

function) $S_I \to S_{II}$. For our analysis, we randomly create in total $T = 1000$ injections, denoted by $\mu^t : S_I \to S_{II}$, $t \in \{1, \ldots, T\}$.

Let $t \in \{1, \ldots, T\}$. For each pair of matched students (s, s') at μ^t, let Δ_s^t denote the difference between the later-round effect for problem I and the later-round effect for problem II. Formally,

$$\begin{aligned}\Delta_s^t \;\equiv\; & \frac{(\text{completion time of } s \text{ for problem I } - \text{completion time of } s' \text{ for problem I })}{\text{average completion time for problem I in earlier group}} \\ - \; & \frac{(\text{completion time of } s' \text{ for problem II } - \text{completion time of } s \text{ for problem II })}{\text{average completion time for problem II in earlier group}}.\end{aligned}$$

The denominators acknowledge the fact that inherent differences between the two problems may result in distinct time demands for their resolution. We opt to utilize the average completion times from the earlier groups, as they are arguably less susceptible to potential cheating influences. With the measurements of the differences in hand, we apply a Wilcoxon signed-rank test to analyze whether the median of the vector $\Delta^t = (\Delta_s^t)_{s \in S_I}$ is significantly different from 0. Let p^t be the one-sided p-value of the Wilcoxon signed-rank test for μ^t. We find that the median of $p = (p^1, \ldots, p^T)$ is equal to 0.589 for rounds 1+2, 0.999 for rounds 3+4, and 0.174 for rounds 5+6. Thus, we again cannot reject the null hypothesis that the reduction of the completion time (from earlier to later round) is independent of the number of versions of a problem.

4 Concluding remarks

Our randomized field experiment was designed with the objective of systematically examining cheating during an online exam through potential information flow from earlier to later rounds. We found that students completed problems more quickly in later rounds compared to earlier rounds. Cheating is a potential explanation for this finding, but other factors such as time pressure[18] or learning cannot be dismissed. We also believe the likely source of information flow is students sharing their solutions via messenger apps. This is based on the fact that many students know each other and communicate through social networks and messenger apps in their daily lives. Hired imposters and note-taking sites, two other alternative sources of cheating, are less likely to be driving factors of our findings. In fact, if students exclusively hired imposters, we would not have observed a significant difference between the earlier and the later rounds. Hired imposters presumably have a higher ability to solve exam questions, which would result in a constant effect throughout the entire exam, independent of the particular round in which a problem is encountered. Furthermore, it is unlikely that the correct answers were available on note-taking sites, as the exam questions had been newly created from scratch and not shared with anyone or uploaded before the exam.

Students encountering different versions of the exam did not exhibit higher grades in later rounds, which contrasts with Klijn et al. (2022) who found a significant later-round effect on average correctness when the set of exam questions is common to all students. A possible explanation is that with multiple exam versions, the information circulating in chat groups becomes more

[18]It is important to recall that the later rounds are rounds 2, 4, and 6. However, round 2 takes place near the start of the exam, when time pressure is usually not a concern.

disorganized and less reliable. Thus, our study suggests that the introduction of even a small number of different exam versions can significantly reduce the effectiveness of cheating, indicating the preventive measure's success. In conclusion, employing a limited number of distinct exam versions can diminish cheating effectiveness in online exams, akin to implementing multiple versions in in-class examinations. This holds practical significance, particularly in scenarios where the usage of online proctoring services is not possible.

Our field experiment was conducted in 2021, prior to the first version of ChatGPT which was launched by OpenAI in November 2022. Educational institutions can leverage generative AI to create a more secure and fair assessment environment. For instance, generative AI can automatically create a diverse array of questions –either with or without randomized elements– based on a pool of topics. Thus, each student could receive a customized exam that tests the same concepts but in different ways, significantly reducing the likelihood of cheating *among students*. However, students may cheat by using easily accessible chatbots to submit AI-generated responses as their own. While advanced proctoring tools can monitor students during exams to deter the use of these AI resources, such solutions, as we mentioned earlier, may not be possible in all settings. Additional strategies include implementing more frequent, shorter assessments instead of high-stakes exams and placing greater emphasis on alternative evaluation methods like short essay questions and oral exams. The former approach can help reduce student stress and minimize reliance on chatbots, while the latter provides a safeguard if chatbots significantly compromise the integrity of multiple-choice exams. Finally, AI can provide valuable insights into exam-taking behaviors, enabling educators to detect sudden surges in high scores or unusual patterns in completion times.

References

Arnold, I. J. (2022): "Online proctored assessment during COVID-19: Has cheating increased?" *Journal of Economic Education*, 53(4): 277–295.

Basken, P. (2020): "Universities say student cheating exploding in Covid era." *https://www.timeshighereducation.com/news/universities-say-student-cheating-exploding-covid-era*.

Bilen, E. and Matros, A. (2021): "Online cheating amid COVID-19." *Journal of Economic Behavior and Organization*, 182: 196–211.

Elzinga, K. G. and Harper, D. Q. (2023): "In-person versus online instruction: Evidence from principles of economics." *Southern Economic Journal*, 90: 3–30.

Hill, A. J. and LoPalo, M. (2024): "The effects of online vs in-class testing in moderate-stakes college environments." *Economics of Education Review*, 98: 102505.

Holden, O. L., Norris, M. E., and Kuhlmeier, V. A. (2021): "Academic integrity in online assessment: A research review." *Frontiers in Education*, 6: 639814.

Imran, R., Fatima, A., Salem, I. E., and Allil, K. (2023): "Teaching and learning delivery modes in higher education: Looking back to move forward post-COVID-19 era." *International Journal of Management Education*, 21(2): 100805.

Janke, S., Rudert, S. C., Petersen, A., Fritz, T. M., and Daumiller, M. (2021): "Cheating in the wake of COVID-19: How dangerous is ad-hoc online testing for academic integrity?" *Computers and Education Open*, 2: 100055.

Klijn, F., Mdaghri Alaoui, M., and Vorsatz, M. (2022): "Academic integrity in on-line exams: Evidence from a randomized field experiment." *Journal of Economic Psychology*, 93: 102555.

Le Maux, B. and Necker, S. (2023): "Honesty nudges: Effect varies with content but not with timing." *Journal of Economic Behavior and Organization*, 207: 433–456.

Newton, P. M. and Essex, K. (2023): "How common is cheating in online exams and did it increase during the COVID-19 pandemic? A systematic review." *Journal of Academic Ethics*, 22: 323–343.

Noorbehbahani, F., Mohammadi, A., and Aminazadeh, M. (2022): "A systematic review of research on cheating in online exams from 2010 to 2021." *Education and Information Technologies*, 27: 8413–8460.

Ratten, V. (2023): "The post COVID-19 pandemic era: Changes in teaching and learning methods for management educators." *International Journal of Management Education*, 21(2): 100777.

Vazquez, J. J., Chiang, E. P., and Sarmiento-Barbieri, I. (2021): "Can we stay one step ahead of cheaters? A field experiment in proctoring online open book exams." *Journal of Behavioral and Experimental Economics*, 90: 101653.

Online Appendix of "Cheating in an Online Academic Exam: Mitigation through Multiplicity of Exam Versions?"

Authors:

Flip Klijn

Corresponding author. Institute for Economic Analysis (CSIC) and Barcelona School of Economics, Campus UAB, 08193 Bellaterra (Barcelona), Spain; e-mail: flip.klijn@iae.csic.es.

Mehdi Mdaghri Alaoui

Department of Economics and Business, Universitat Pompeu Fabra, C/ Ramon Trias Fargas 25, 08005 Barcelona, Spain; e-mail: mehdi.mdaghri@upf.edu.

Marc Vorsatz

Departamento de Análisis Económico, Universidad Nacional de Educación a Distancia (UNED), Paseo Senda del Rey 11, 28040 Madrid, Spain; e-mail: mvorsatz@cee.uned.es.

Date: October 26, 2024

Contents:

A Course structure and screenshots of the exam

Course structure

The introductory course on game theory was distributed over 10 weeks in the second trimester of academic year 2020-2021. The (compulsory) course was taught in English to four different groups of students.[19] Students' evaluation was based on continuous assessment. A student's final grade was determined by three items: 2 (intermediate) tests, 7 seminars, and 1 final exam. Specifically,

- the two intermediate tests could give up to 12 and 13 points, respectively [25 points in total],
- each of the 7 seminars gives 1 point for attendance, and up to 8 additional points for the participation/work in the 7 seminars [15 points in total], and
- the final exam could give up to 60 points.[20]

The final grade (between 0 and 10) was obtained by dividing the achieved number of points by 10.

The two intermediate tests took place through the university's online platform Moodle and consisted of 12 and 13 multiple-choice questions, respectively. For all questions of the tests, there were 2, 3, or 4 different (but similar) versions which were randomly assigned to students. Access to each test was open for approximately 24 hours, but once a student started, he/she had 75 minutes to complete it without backtracking.

The 7 (in-class) seminars took place in weeks 3 to 9 of the course. In each seminar, exercises from take-home problem sets were discussed in groups of 25-30 students. The main objective of the seminars was to give students the opportunity to ask questions and to present solutions (for which they could earn up to 8 points in total).

Students were informed of all aforementioned details of the continuous assessment in the first week of the course. The final exam was programmed and executed in Moodle.

[19] For each of the four groups, the eight different orders were distributed almost equally.

[20] Since at the final exam students could get up to 100 points, for the calculation of the final grade the exam points were multiplied by 0.6.

Screenshots of the final exam

Below we provide screenshots of the final exam which was in English (except for some sentences and buttons that are part of the university's online platform). The first screen that the students saw contained the exam instructions (Figure 2) and the university's code of ethics (Figure 3). Note that "Aula Global" is the name of the university's Moodle platform.

Carefully read the instructions below before you proceed.

- This exam consists of 20 multiple-choice questions.
- VERY IMPORTANT: You have to answer all questions on each screen before you move to the next screen. Moving back to a previous screen is not possible.
- HANDWRITTEN SOLUTIONS: Your hand-written solution to the questions must follow the order of the questions. Number your answers accordingly: 1,...,20 (even the questions that you leave unanswered). Start writing on a new sheet each time you move to the next screen.
- GRADING: Each question has a unique correct answer. A correct answer gives 5 points and an incorrect answer gives 1 negative point. Leaving a question unanswered gives 0 points. In total you can obtain up to 100 points for the exam.
- VALIDATION: Your answers to the multiple-choice questions are graded as described above. However, for each answer there must be handwritten explanations or calculations that show how you derived your answer. If you answer a question in aula global correctly, but do not provide a sufficiently detailed handwritten solution, the number of points for that question will be set at 0. To validate your answers, you have to upload your scanned solutions in a *single PDF document* to Aula Global. From the moment that you complete the exam you will have (at most) 25 minutes to upload your PDF document. A specific and clearly indicated link in Aula Global (but outside the exam) will allow you to upload your PDF. If you upload it after the period of 25 minutes or if you send it by email, your answers will not be valid. We have set the maximum size of the PDF at a generous 10 Mb. After uploading your PDF file, verify that the file was uploaded correctly. We cannot assume any responsibilities concerning the timely upload of your PDF file.
- REVISION: The results will be available in Aula Global as soon as we have been able to make all necessary verifications. We will send an email once everything is ready. Please be patient. The date and time of revision will be announced by email and through aula global.

Figure 2: Exam instructions.

Extract from The University Code of Ethics

III. ETHICAL PRINCIPLES ON WHICH UNIVERSITY LIFE IS BASED

1. a. Academic integrity

Academic integrity means all the forms of behaviour linked to teaching, from the perspective of students and lecturers alike, on the basis of shared moral principles. It is a major factor in establishing the trust a community requires. If academic integrity is to be preserved in changing circumstances, constant discussion of what is and is not deemed honest is necessary, with a view to reaching a new consensus thereon. In democratic societies, decisions should be based on consensus wherever possible, although that does not absolve academic authorities of their responsibility for making them.

1. e. Truthfulness in academic assessments

Academic assessments have two functions. The first is to check that the objectives and competences envisaged in a subject's course plan are being met and acquired respectively. The second is to demonstrate that students have attained the minimum levels established for them be awarded the qualification corresponding to their study programme. With regard to the former function, academic assessments are vital for lecturers and students alike, as they show whether or not all parties are working hard enough to ensure that the envisaged progress is made. The latter function is an aspect of the University's social responsibility, as it is society that has charged universities with providing higher education and asks that graduates be capable of performing their professional activity properly. Assessments are instruments for identifying what students have and have not learned, and must be unequivocally adapted to the educational goals and competences pre-established in each subject's course plan. Students should not confuse passing with learning. Both lecturers and students must focus on guaranteeing effective learning and not merely passing assessments. Copying and plagiarism are forms of misconduct to which the corresponding prescribed punishments must be applied, not only to demonstrate the university community's rejection thereof but also to prevent the reputation of the University and its graduates being harmed.

By starting the exam, I acknowledge and accept the University Code of Ethics.

Qüestionari cronometrat

El qüestionari té un temps màxim de 2 hores. El temps començarà a comptar des del moment en què inicieu l'intent i s'ha d'enviar abans que el temps expiri. Confirmeu que voleu començar ara?

Inicia l'intent | Cancel·la

Figure 3: Displayed part of the university's code of ethics.

The last part in Figure 3 translates as follows: "Timed questionnaire. The questionnaire has a maximal duration of 2 hours. The time counter starts at the moment that you start your "attempt" and [the answers to the questionnaire] have to be submitted before the time expires. Confirm that you would like to start now. [Start attempt] [Cancel]"

B Nomenclature

Question: A multiple-choice task with a finite number of possible answers of which one and only one is correct. An exam consists of a total of 20 questions. Questions are labeled using the Latin numerals 1,2,. . .,20.

Problem: A set of questions that are related to each other and clearly and explicitly grouped together. The 20 questions are divided into 6 problems. Problems are labeled using the Roman numerals I,II,. . .,VI.

Problem version: A problem either has 2, 4, or 6 versions. The different versions were obtained by applying permutations, scaling, etc. of numerical values to all questions of the problem.

Round: A stage/phase of the exam. The exam consists of 6 consecutive rounds, labeled 1,2,. . .,6. In rounds 1 and 2, students solve problems I and II. Afterwards, in rounds 3 and 4, students solve problems 3 and 4. And finally, in rounds 5 and 6, students solve problems V and VI. The order in which students solve problems in each pair of rounds 1+2, 3+4, and 5+6 is random. For instance, a student either solves problem III in round 3 and problem IV in round 4 or she faces problem IV in round 3 and problem III in round 4.

Order: A feasible order or sequence of the 6 problems, specifying which problem a student faces in which round. There are $2 \times 2 \times 2 = 8$ feasible orders, i.e., orders where problems I and II are in rounds 1 and 2, problems III and IV are in rounds 3 and 4, problems V and VI are in rounds 5 and 6. An example of a feasible order is (II,I,III,IV,VI,V). An example of an infeasible order is (II,IV,I,III,V,VI): problem I is always presented in round 1 or 2 (never in round 3) and problem IV is always presented in round 3 or 4 (never in round 2).

Earlier rounds: Rounds 1, 3, and 5.

Later rounds: Rounds 2, 4, and 6.

Earlier students: Students who work on a problem in an earlier round.

Later students: Students who work on a problem in a later round.

Average correctness of a given problem in a particular round: the average points per question by the students that were faced with the problem in that round.

Average completion time of a problem in a particular round: the average time taken to answer the problem by the students that were faced with the problem in that round.

C Subject pool

Table 5 below provides more information about our subject pool. The two-sided p-values of the Kruskal-Wallis tests show that for all four variables (columns in the table), we cannot reject the null hypothesis that the subject pools of the eight orders are equal.

Exam Order	Gender	Intermediate	Attendance	Participation
A	0.411	17.886	6.497	4.714
B	0.583	18.747	6.750	3.967
C	0.509	18.893	6.732	5.054
D	0.569	19.404	6.828	5.207
E	0.459	18.970	6.574	4.311
F	0.448	18.830	6.797	4.424
G	0.554	19.026	6.733	4.732
H	0.448	18.878	6.702	4.842
Kruskal-Wallis p	0.733	0.085	0.379	0.517

Table 5: Subject pool information (average). The personal characteristics are defined as follows: Gender (1 for female), Intermediate (total number of points obtained in the 2 intermediate tests [0-25]), Attendance (total number of seminars attended [0-7]), Participation (total number of points obtained for participating in seminars [0-8]).

D Differences between exam versions

For each of the 6 problems there were a number of different versions obtained by applying permutations, scaling, etc. of numerical values, which should not vary in their level of difficulty. Table 6 shows in this respect that the 6 versions in problem VI were not of equal difficulty in the earlier round. Nevertheless, our main insight that Hypothesis 1 only holds for completion time and not for average points per question does not depend on the exam versions: there is no problem version with a significant average point increment and there is only one problem version, V.d, for which the completion time does not decrease significantly.

problem.version	average points per question			completion time		
	earlier	later	MWU p	earlier	later	MWU p
I.a	1.79	1.82	*0.869*	20.7	16.3	*0.000*
I.b	1.47	1.54	*0.703*	21.0	17.6	*0.000*
Kruskal-Wallis p	*0.142*	*0.375*		*0.846*	*0.144*	
II.a	2.02	2.29	*0.624*	21.4	17.9	*0.017*
II.b	2.72	2.25	*0.183*	21.5	18.6	*0.024*
II.c	2.18	1.86	*0.144*	23.8	20.1	*0.018*
II.d	2.67	2.45	*0.608*	22.3	18.1	*0.000*
II.e	2.10	2.26	*0.541*	23.6	18.5	*0.002*
II.f	2.12	2.09	*0.880*	25.2	20.2	*0.004*
Kruskal-Wallis p	*0.187*	*0.404*		*0.162*	*0.148*	
III.a	3.24	2.85	*0.199*	17.5	12.5	*0.000*
III.b	2.74	2.81	*0.746*	18.4	14.0	*0.000*
Kruskal-Wallis p	*0.130*	*0.890*		*0.132*	*0.015*	
IV.a	1.10	1.19	*0.749*	36.9	34.0	*0.032*
IV.b	1.77	1.19	*0.122*	35.2	31.8	*0.027*
IV.c	1.34	1.04	*0.210*	35.8	30.9	*0.002*
IV.d	1.12	1.23	*0.544*	37.5	32.0	*0.001*
Kruskal-Wallis p	*0.114*	*0.814*		*0.594*	*0.132*	
V.a	2.09	2.11	*0.800*	18.7	13.0	*0.000*
V.b	2.08	2.11	*0.680*	17.0	14.8	*0.050*
V.c	2.54	2.20	*0.374*	17.2	14.6	*0.038*
V.d	1.97	1.89	*0.773*	16.5	14.5	*0.118*
Kruskal-Wallis p	*0.196*	*0.709*		*0.266*	*0.504*	
VI.a	3.11	1.89	*0.004*	17.2	11.5	*0.000*
VI.b	3.14	2.46	*0.111*	17.1	12.7	*0.000*
VI.c	2.81	2.41	*0.429*	17.5	12.0	*0.000*
VI.d	3.05	2.14	*0.045*	16.2	12.7	*0.001*
VI.e	2.28	2.29	*0.872*	16.5	12.8	*0.000*
VI.f	3.77	2.55	*0.003*	17.6	14.0	*0.001*
Kruskal-Wallis p	*0.010*	*0.659*		*0.918*	*0.928*	

Table 6: Disaggregated analysis. Mann-Whitney U tests (MWU) are used to validate Hypothesis 1. The equality of the exam versions is analyzed with the help of Kruskal-Wallis tests. We report two-sided p-values for Kruskal-Wallis tests and one-sided p-values for the MWU tests.

E Scatter plots of individual data

The scatter plots in Figure 4 visualize the individual data. In each panel/problem, each point represents one student's average points per question and completion time. A caveat is that students in the same group ("earlier" or "later") with both the same number of correct answers and the same completion time (in minutes) are represented by overlapping circles or overlapping crosses. Earlier (later) students are represented by purple crosses + (orange circles ○).

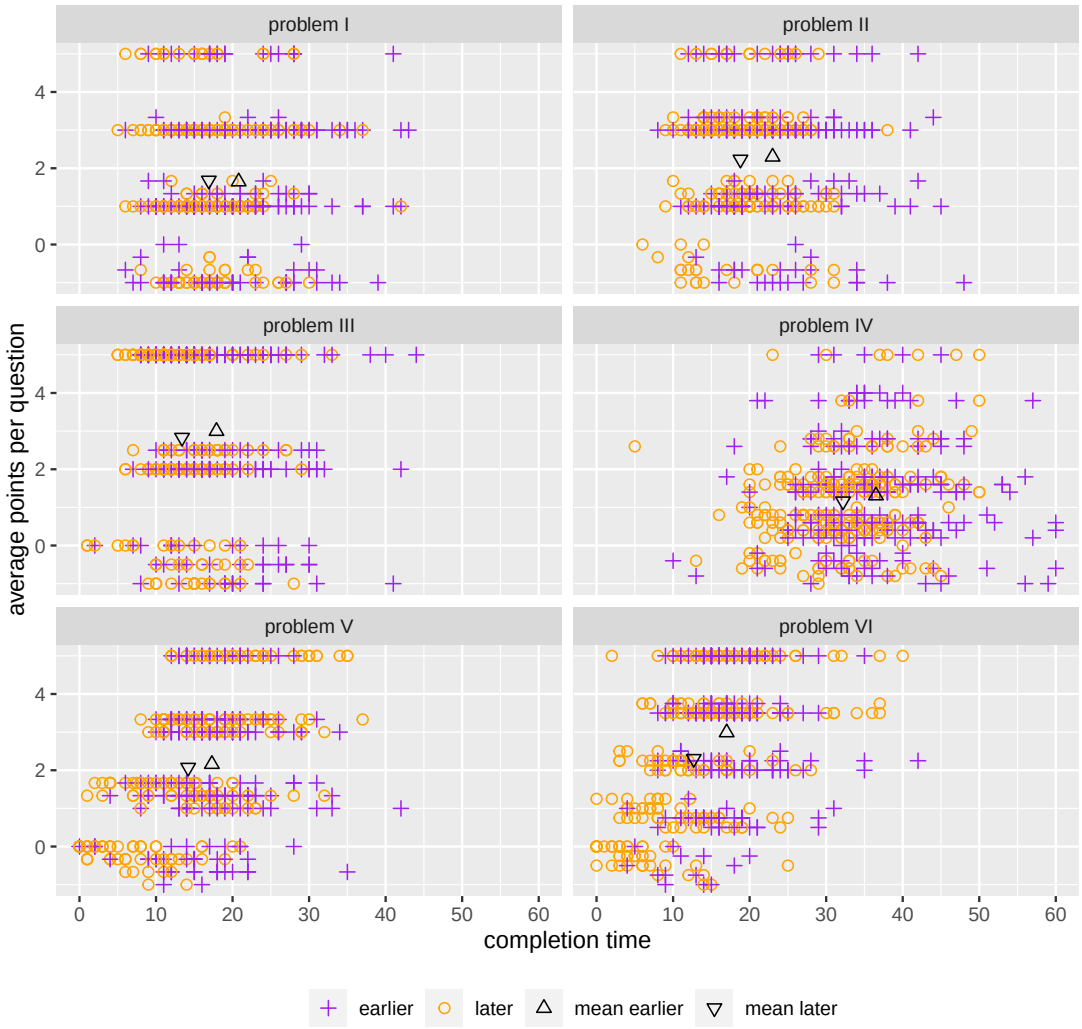


Figure 4: Panels describe correctness and completion time at the individual level for all six problems. The triangles △ and ▽ represent the averages of the earlier and later students, respectively.

We observe that for each problem, the pair (average completion time, average correctness) for the later students is positioned to the left and slightly below the corresponding pair for the earlier students, as indicated by the respective triangles ▽ and △.

F Distribution of answers

Table 7 below complements Table 3 by giving insights into changes of the answers from earlier to later rounds by providing the distributions of correct, incorrect, and blank answers.

With the exception of problem I, in all problems, the later round yields a smaller proportion of correct answers and a larger proportion of blank answers (relative to the earlier round). For problems I, II, and III, the later round also yields a larger proportion of incorrect answers (relative to the earlier round). However, for problems IV, V, and VI, the later round yields a *smaller* proportion of incorrect answers (relative to the earlier round). Therefore, the worse performance in the later round of problems IV, V, and VI is due to a smaller proportion of correct answers and a larger proportion of blank answers. A possible explanation of the large proportion of blank answers in the later round of problems V and VI is that these were the problems at the end of the exam and students may have run out of time.

problem	correct	incorrect	unanswered
problem I earlier	0.431	0.504	0.066
problem I later	0.440	0.521	0.038
problem II earlier	0.534	0.366	0.100
problem II later	0.517	0.368	0.115
problem III earlier	0.646	0.226	0.129
problem III later	0.613	0.237	0.150
problem IV earlier	0.345	0.414	0.241
problem IV later	0.312	0.399	0.289
problem V earlier	0.483	0.243	0.275
problem V later	0.445	0.153	0.402
problem VI earlier	0.649	0.247	0.104
problem VI later	0.507	0.234	0.259

Table 7: Distribution of answers.

G Cumulative completion times

order	round					
	1	2	3	4	5	6
A	**I**	**II**	**III**	**IV**	**V**	**VI**
	20.7	39.4	57.4	90.6	108.2	121.4
B	**II**	**I**	**III**	**IV**	**V**	**VI**
	24.7	42.5	60.3	91.5	108.3	120.5
C	**I**	**II**	**IV**	**III**	**V**	**VI**
	20.1	39.0	75.3	89.1	106.1	119.2
D	**II**	**I**	**IV**	**III**	**V**	**VI**
	22.9	40.3	77.9	90.7	108.7	121.2
E	**I**	**II**	**III**	**IV**	**VI**	**V**
	22.5	41.2	58.6	90.7	106.7	120.5
F	**II**	**I**	**III**	**IV**	**VI**	**V**
	22.2	39.7	58.3	90.5	107.4	120.7
G	**II**	**I**	**IV**	**III**	**VI**	**V**
	22.1	36.7	72.2	86.0	104.3	119.7
H	**I**	**II**	**IV**	**III**	**VI**	**V**
	19.9	38.9	75.2	88.3	105.1	119.6

Table 8: Average cumulative completion times since the start of the exam for each of the orders (A,B,...,H) of the 6 problems (I, II, III, IV, V, VI).

H Joint data OLS regression model

We consider the following joint data OLS regression model in order to evaluate Hypothesis 2 with respect to completion time in an alternative way:

$$\texttt{time}_{sj} = \beta_0 + \beta_1 \texttt{points}_{sj} + \beta_2 \texttt{versions}_j + \beta_3 \texttt{later}_{sj} + \beta_4 \texttt{versions}_j \cdot \texttt{later}_{sj} + \varepsilon_{ij},$$

where $\texttt{time}_{sj}$ is the completion time of student s for problem j, $\texttt{points}_{sj}$ are the total points student s scores on problem j (this variable serves as a control), $\texttt{versions}_j \in \{2, 4, 6\}$ corresponds to the number of versions of the problem j, and $\texttt{later}_{sj}$ is a dummy variable that takes value 1 if student s faces problem j in the later round and 0 if the student solves problem j in the earlier round. According to Hypothesis 2, the later-round advantage is less pronounced for problems with a larger number of versions, which means that the interaction variable $\texttt{versions}_j \cdot \texttt{later}_{sj}$ is expected to have a positive impact on completion time. Table 9 presents the estimation results.

variable	estimate	std. error	p-value
`intercept`	21.142	0.688	0.000
`points`	0.077	0.032	0.016
`versions`	0.080	0.160	0.615
`later`	-3.954	0.965	0.000
`versions` · `later`	0.025	0.224	0.912

Table 9: Joint data OLS estimation on completion time.

We observe that the results from this joint data estimation procedure perfectly align with the results obtained from the three separate regressions in the main text. First, students who score more points on a problem take slightly (but significantly) more time to complete the problem. Second, there is a highly significant later-round advantage of about 4 minutes per problem. And finally, we cannot reject the null hypothesis that the reduction of the completion time is independent of the number of versions of a problem.

I Informativeness of exam grades

We use a random effects model to estimate the relationship between the final exam grade and the continuous assessment during the course (i.e., attendance/performance at the 7 seminars and the 2 intermediate tests).[21] Note that while attendance/performance at the seminars is basically cheating-proof, this is not necessarily the case for the (online) intermediate tests. In the estimation, we control for gender. Table 10 summarizes the estimation results.

Intercept	-8.116
	(4.183)
Intermediate tests	1.333^{***}
	(0.138)
Seminar attendance	0.742
	(0.668)
Seminar participation	0.757^{***}
	(0.171)
Gender	-1.970^{*}
	(0.937)
Observations	463

Table 10: Random effects estimation of the dependency of the final exam grade. We refer to Online Appendix C on subject pool information for a formal definition of the explanatory variables. The final exam grade is scaled proportionally from [0,100] (the maximum number of points in the final exam is 100) to the interval [0,60] (the weight of the final exam in the overall grade is 60%). In parenthesis, we present standard deviations. $^{*}\, p < .05$; $^{**}\, p < .01$; $^{***}\, p < .001$.

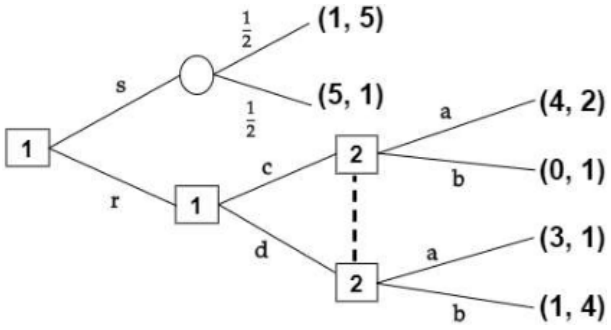
We observe that there is a positive, highly significant relationship between the final exam grade and both the grade in the intermediate tests and seminar participation (all expressed in "points" as explained in Online Appendix A). Hence, final exam grades can be considered informative.

[21] It would be interesting to match the exam grade with performance measures from other courses. Unfortunately, this data is not available.

J Sample exam

Below we reproduce a sample exam of 20 questions together with the possible answers. Note: (i) for each problem, only one version is chosen and reproduced, (ii) problems are in order A, (iii) the correct answer to all questions is a) but in each student's exam all answers were shuffled randomly, and (iv) most of the format has been removed.

Problem I. Consider the following two-player game in extensive form with risk-neutral players. In each payoff vector, the first number indicates the payoff of player 1 and the second number indicates the payoff of player 2.



Question 1. How many subgames are there? And how many strategies does player 1 have?

a) ⊗ There are exactly 3 subgames. Player 1 has exactly 4 strategies.
b) ◯ There are exactly 2 subgames. Player 1 has exactly 4 strategies.
c) ◯ There are exactly 2 subgames. Player 1 has exactly 2 strategies.
d) ◯ There are exactly 2 subgames. Player 1 has exactly 3 strategies.
e) ◯ There are exactly 3 subgames. Player 1 has exactly 2 strategies.
f) ◯ There are exactly 3 subgames. Player 1 has exactly 3 strategies.
g) ◯ There is exactly 1 subgame. Player 1 has exactly 2 strategies.
h) ◯ There is exactly 1 subgame. Player 1 has exactly 3 strategies.
i) ◯ There is exactly 1 subgame. Player 1 has exactly 4 strategies.

Question 2. How many Nash equilibria in pure strategies does the game have?

a) ⊗ The game has exactly 3 Nash equilibria in pure strategies.
b) ◯ The game has no Nash equilibria in pure strategies.
c) ◯ The game has exactly 1 Nash equilibrium in pure strategies.
d) ◯ The game has exactly 2 Nash equilibria in pure strategies.
e) ◯ The game has exactly 4 Nash equilibria in pure strategies.

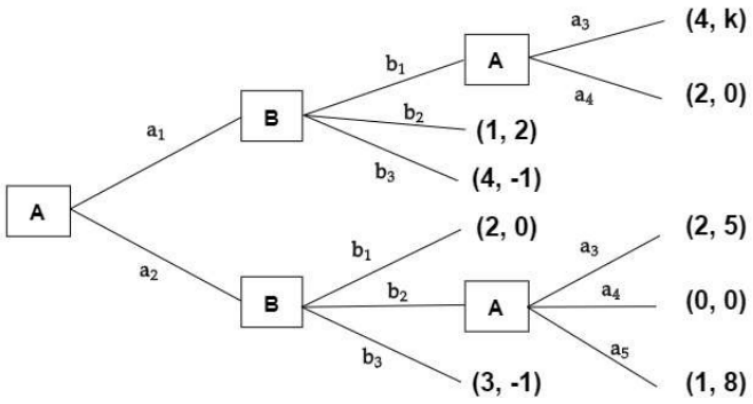
Question 3. What is the number of subgame perfect Nash equilibria in pure strategies?

a) ⊗ There are exactly 2 subgame perfect Nash equilibria in pure strategies.
b) ◯ There are no subgame perfect Nash equilibria in pure strategies.
c) ◯ There is exactly 1 subgame perfect Nash equilibrium in pure strategies.

d) ◯ There are exactly 3 subgame perfect Nash equilibria in pure strategies.
e) ◯ There are exactly 4 subgame perfect Nash equilibria in pure strategies.

. . . . . .

Problem II. Consider the following two-player game in extensive form with players A and B. In each payoff vector, the first number indicates the payoff of player A and the second number indicates the payoff of player B.



The constant $k \neq 2$ is a *real* number.

Question 4. How many strategies does each player have?

a) ⊗ Player A has exactly 12 strategies. Player B has exactly 9 strategies.
b) ◯ Player A has exactly 5 strategies. Player B has exactly 3 strategies.
c) ◯ Player A has exactly 5 strategies. Player B has exactly 6 strategies.
d) ◯ Player A has exactly 5 strategies. Player B has exactly 9 strategies.
e) ◯ Player A has exactly 6 strategies. Player B has exactly 3 strategies.
f) ◯ Player A has exactly 6 strategies. Player B has exactly 6 strategies.
g) ◯ Player A has exactly 6 strategies. Player B has exactly 9 strategies.
h) ◯ Player A has exactly 12 strategies. Player B has exactly 3 strategies.
i) ◯ Player A has exactly 12 strategies. Player B has exactly 6 strategies.
j) ◯ None of the other answers.

Question 5. Is (b_2, b_1) a best response to (a_1, a_4, a_3)?

a) ⊗ Yes, for all $k \neq 2$.
b) ◯ Yes, but only for all $k < 2$.
c) ◯ Yes, but only for all $k > 2$.
d) ◯ We cannot determine this, because there is perfect information.
e) ◯ We cannot determine this, because there is imperfect information.
f) ◯ We can determine this, but none of the other answers is correct.

Question 6. Assuming common knowledge of rationality, when does player A get a strictly higher payoff than player B?

a) ⊗ For all $k \in (2, 4)$ only.
b) ◯ For all $k < 2$ only.
c) ◯ For all $k > 2$ only.
d) ◯ For all $k \neq 2$.
e) ◯ For all $k > 4$ only.
f) ◯ None of the other answers is correct.

. . . . . .

Problem III. We consider the following situation with players 1 and 2. Each player i chooses an effort level $x_i \in [0, 10]$, i.e., a *real* number in the interval [0,10]. The payoff function of player 1 is $u_1(x_1, x_2) = 20x_1 - x_1^2 - x_1x_2$, while the payoff function of player 2 is $u_2(x_1, x_2) = 20x_2 - x_2^2 - x_1x_2$.

Question 7. Consider the normal-form game where the two players choose their effort levels simultaneously. Which of the following statements about the Nash equilibria in pure strategies is true?

a) ⊗ There is a unique Nash equilibrium (x_1, x_2) in pure strategies and $x_2 = 20/3$.
b) ◯ There is a unique Nash equilibrium (x_1, x_2) in pure strategies and $x_2 = 10/3$.
c) ◯ There is a unique Nash equilibrium (x_1, x_2) in pure strategies and $x_2 = 10$.
d) ◯ There is a unique Nash equilibrium (x_1, x_2) in pure strategies and $x_2 = 8$.
e) ◯ There is a unique Nash equilibrium (x_1, x_2) in pure strategies and $x_2 = 5$.
f) ◯ There is a unique Nash equilibrium (x_1, x_2) in pure strategies and $x_2 = 4$.
g) ◯ There is a unique Nash equilibrium (x_1, x_2) in pure strategies and $x_2 = 2.5$.
h) ◯ There is a unique Nash equilibrium (x_1, x_2) in pure strategies and $x_2 = 0$.
i) ◯ There are exactly two Nash equilibria in pure strategies.
j) ◯ There is an infinite number of Nash equilibria in pure strategies.

Question 8. Consider the sequential game with perfect information where first player 1 has to choose his effort level and then player 2, after having observed player 1's choice, makes her choice. What effort level x_2 will player 2 choose?

a) ⊗ $x_2 = 5$.
b) ◯ $x_2 = 20/3$.
c) ◯ $x_2 = 10/3$.
d) ◯ $x_2 = 5/3$.
e) ◯ $x_2 = 10$.
f) ◯ $x_2 = 2.5$.
g) ◯ $x_2 = 0$.
h) ◯ None of the other answers.

. . . . . .

Problem IV.

Question 9. Consider the following functions:

- $v(x) = (32 + 16\sqrt{x})^2$
- $w(x) = 16 + 8x$
- $k(x) = 2 - \sqrt{x}$
- $r(x) = 4 - 2\sqrt{x}$
- $s(x) = -2 + \sqrt{6x}$
- $t(x) = -4 - 2\sqrt{x}$

Which/how many of the functions v, w, k, r, s, t represent(s) the same preferences over lotteries as the von Neumann-Morgenstern utility function $u(x) = 32 + 16\sqrt{x}$?

a) ⊗ only s
b) ◯ only v
c) ◯ only w
d) ◯ only k
e) ◯ only r
f) ◯ only t
g) ◯ none of the functions
h) ◯ exactly two functions
i) ◯ exactly three functions

The next three questions (10, 11, and 12) concern an agent called Arnau. Arnau's wealth consists of $\omega = 10$ euros and two lottery tickets. One lottery ticket (T_1) gives a prize of 100 euros with probability 0.5 and a prize of 4 euros with probability 0.5. The other lottery ticket (T_2) gives a prize of 11 euros with probability 0.5 and a prize of 34 euros with probability 0.5. The two lottery tickets are independent: T_1 is a ticket in El Gordo, while T_2 is a ticket in La Grossa. Arnau's preferences can be represented by the von Neumann-Morgenstern utility function $u(x) = -1+5\sqrt{x}$ where x expresses euros.

Question 10. What is the expected value of Arnau's wealth?

a) ⊗ 84.5
b) ◯ 42.25
c) ◯ 94.5
d) ◯ 74.5
e) ◯ 134
f) ◯ 15
g) ◯ none of the other values

Question 11. Suppose Laia is interested in buying the two tickets from Arnau. What is the minimum amount of money she would have to offer him so that Arnau is willing to give her his tickets?

a) ⊗ The minimum amount is between 66 and 66.5 euros.
b) ◯ The minimum amount is between 66.5 and 67 euros.
c) ◯ The minimum amount is between 67 and 67.5 euros.
d) ◯ The minimum amount is between 67.5 and 68 euros.
e) ◯ The minimum amount is below 66 euros.
f) ◯ The minimum amount is above 68 euros.

Question 12. (requires substantial calculations) Suppose Arnau's parents tell him that they will compensate him with some money if Arnau's promises to give his prize from lottery T_2 to his little brother Blai. What is the minimum compensation that Arnau will demand from his parents?

a) ⊗ The minimum compensation is between 21.5 and 22 euros.
b) ◯ The minimum compensation is between 21 and 21.5 euros.
c) ◯ The minimum compensation is between 22.3 and 22.7 euros.
d) ◯ The minimum compensation is between 22 and 22.3 euros.
e) ◯ The minimum compensation is between 22.7 and 23 euros.
f) ◯ The minimum compensation is below 21 euros.
g) ◯ The minimum compensation is above 23 euros.

Question 13. Cecilia's preferences on the interval $(0, \infty)$ can be represented by a von Neumann-Morgenstern utility function u. The only other information that Cecilia is willing to reveal to us is that she is risk-averse.

Cecilia's friend Daniel has initial wealth $\omega > 0$. Daniel says that his preferences on the interval $(0, \infty)$ can be represented by the von Neumann-Morgenstern utility function v given by

$$v(x) = 2u(x) + 30 - \gamma x,$$

where x expresses euros and $\gamma < 0$ is a negative constant. What, if anything, can we say about Daniel's attitude towards risk on the interval $(0, \infty)$?

a) ⊗ Daniel is risk-averse.
b) ◯ Daniel is risk-loving.
c) ◯ Daniel is risk-neutral.
d) ◯ Since $\gamma < 0$ is negative, the function v is not a von Neumann-Morgenstern utility function.
e) ◯ We cannot determine Daniel's attitude towards risk, because we need further information about ω.
f) ◯ We cannot determine Daniel's attitude towards risk, because we need further information about γ.
g) ◯ We cannot determine Daniel's attitude towards risk, because we need further information about u.

. . . . . .

Problem V.

Question 14. Consider the following two-player game in normal form where players 1 and 2 make their decisions simultaneously. In each payoff vector, the first number indicates the payoff of player 1, while the second number indicates the payoff of player 2.

$1\backslash 2$	L	C	R
U	$2,3$	$1,5$	$2,-2$
M	$a,1$	$3,2$	$2,4$
D	$1,6$	$3,1$	$0,-1$

Here the constant a is a *real* number.

Considering pure strategies only, which of the following is true regarding the maxmin payoff m_1 of player 1?

a) ⊗ For each real number a, $1 \leq m_1 \leq 2$.
b) ◯ For each real number a, $m_1 \leq 0$.
c) ◯ For each real number a, $0 \leq m_1 \leq 1$.
d) ◯ For each real number a, $2 \leq m_1 \leq 3$.
e) ◯ For each real number a, $m_1 \geq 3$.
f) ◯ None of the other answers.

Question 15. Consider the following two-player zero-sum game in normal form where players 1 and 2 make their decisions simultaneously. In each payoff vector, the first number indicates the payoff of player 1, while the second number indicates the payoff of player 2.

$1\backslash 2$	L	R
U	$-1,1$	$a,-a$
D	$4,-4$	$-3,3$

Here the constant a is a *real* number.

We are told that player 1's only mixed security strategy is $(\frac{7}{10}, \frac{3}{10})$. What is a?

a) ⊗ $a = 2$
b) ◯ $a = -3$
c) ◯ $a = -2$
d) ◯ $a = -1$
e) ◯ $a = 0$
f) ◯ $a = 1$
g) ◯ $a = 3$
h) ◯ None of the other answers.

Question 16. Consider the following two-player games in normal form where players 1 and 2 make their decisions simultaneously. In each payoff vector, the first number indicates the payoff of player 1, while the second number indicates the payoff of player 2.
Which of the four games G_a, G_b, G_c, G_d are strictly competitive?

$G_a =$

$1\backslash 2$	L	R
T	2, -1	3, -2
B	4, 0	9, -4

$G_b =$

$1\backslash 2$	L	R
T	1, 0	-1, 2
B	0, -2	3, -4

$G_c =$

$1\backslash 2$	L	R
T	1, 9	5, 5
B	8, 3	3, 7

$G_d =$

$1\backslash 2$	L	R
T	7, -1	6, 0
B	-2, 8	1, 5

a) ⊗ G_c and G_d only.
b) ◯ G_a and G_b only.
c) ◯ G_a and G_c only.
d) ◯ G_a and G_d only.
e) ◯ G_b and G_c only.
f) ◯ G_b and G_d only.

. . . . . .

Problem VI.

Question 17. Consider the following normal-form game. The first/second/third number in the payoff vectors represents the payoff of player A, B, and C, respectively.

Player A chooses table a_1 or a_2. Player B chooses row b_1, b_2, or b_3. Player C chooses column c_1 or c_2.

Table a_1	c_1	c_2
b_1	2,3,3	1,3,2
b_2	1,2,1	3,3,3
b_3	1,2,5	3,2,2

Table a_2	c_1	c_2
b_1	1,3,3	2,4,2
b_2	2,3,2	0,3,3
b_3	1,2,2	1,4,5

How many Nash equilibria in pure strategies are there exactly?

a) ⊗ 2
b) ◯ 0
c) ◯ 1
d) ◯ 3
e) ◯ 4
f) ◯ 5
g) ◯ 6

Question 18. Consider the following two-player game in normal form where players 1 and 2 make their decisions simultaneously. In each payoff vector, the first number indicates the payoff of player 1, while the second number indicates the payoff of player 2.

1\2	b_1	b_2	b_3
a_1	3,0	1,5	3,2
a_2	0,3	4,1	0,2
a_3	5,1	0,2	1,1

Which are the rationalizable strategies of player 1?

a) ⊗ a_2 and a_3 only
b) ◯ a_1 and a_2 only
c) ◯ a_1 and a_3 only
d) ◯ a_1 only
e) ◯ a_2 only
f) ◯ a_3 only
g) ◯ a_1, a_2, and a_3

Question 19. Consider the following two-player game in normal form where players 1 and 2 make their decisions simultaneously. In each payoff vector, the first number indicates the payoff of player 1, while the second number indicates the payoff of player 2.

1\2	b_1	b_2	b_3
a_1	2,6	3,9	8,5
a_2	5,6	5,3	3,2
a_3	3,7	6,5	5,3

How many of the 9 strategy-profiles lead to a Pareto-efficient outcome?

a) ⊗ Exactly 3 strategy-profiles.
b) ◯ Exactly 1 strategy-profile.
c) ◯ Exactly 2 strategy-profiles.
d) ◯ Exactly 4 strategy-profiles.
e) ◯ Exactly 5 strategy-profiles.
f) ◯ Exactly 6 strategy-profiles.

Question 20. Consider the following two-player game in normal form where players 1 and 2 make their decisions simultaneously. In each payoff vector, the first number indicates the payoff of player 1, while the second number indicates the payoff of player 2.

1\2	b_1	b_2	b_3
a_1	3,2	3,1	2,4
a_2	2,6	5,4	1,2
a_3	2,3	1,2	4,-3

There is a unique Nash equilibrium in mixed strategies where each player plays precisely two of her pure strategies with strictly positive probability. At this Nash equilibrium, with which probability does player 1 play strategy a_3?

a) ⊗ 1/4
b) ◯ 3/4
c) ◯ 1/3
d) ◯ 2/3
e) ◯ 1/2
f) ◯ 2/5
g) ◯ 3/5